Foreign Policy at BROOKINGS

UNCOMFORTABLE GROUND TRUTHS: PREDICTIVE ANALYTICS AND NATIONAL SECURITY

HEATHER ROFF

NOVEMBER 2020

Reducing uncertainty in all aspects of life is undoubtedly an action that all individuals, societies, and governments seek to achieve. With fair warning of potential future states, we can craft strategies, change courses of action, produce policy interventions, allocate funding or assistance, or pursue any range of actions to either avoid or bring about a given future. In the national security policy space, such forewarning takes on even greater importance owing to the high stakes and lives on the line. It is no surprise, then, that forecasting is a longstanding tradition, both within the intelligence community and the Department of Defense. More recently, the Department of State also started to seek its own oracle through the establishment of the Center for Analytics, the "first enterprise-level data and analytics hub" that will utilize big data and subsequent data analytic tools to "evaluate and refine foreign policy."[1] The belief is that if we have the data, not only can we understand what occurs in the world but we might also be able to predict what will occur in the world, and if we know what will occur, we can intervene in that causal chain.

However, the volume of data available to make such predictions is staggering. Moreover, because foreign policy and national security are complex spaces, understanding which data matter, when, and whether there is any causal structure to various events in these domains is no small feat.[2] Artificial intelligence (AI) promises to provide the necessary power to ingest, analyze, and predict a variety of events.[3] Governments, likewise, are taking notice. For instance, China developed a "geopolitical environment and prediction platform," basically a giant predictive model of geopolitics and a recommender system for various courses of action to support Chinese foreign policies.[4]

The United States, likewise, developed various AI forecasting systems at differing degrees of granularity, from using AI to provide commanders with mission planning and courses of action (COAs) during conflict, to forecasting the onset of civil unrest around the globe.[5] [6] One such system, EMBERS (Early Model Based Event Recognition using Surrogates), was designed for "anticipatory intelligence" to support decision-making in the national security space.[7] EMBERS has gone through various iterations over the years and expanded from being used to look at only Latin American countries to use in countries in the Middle East and North Africa.

This paper focuses not on those predictive analytics systems that attempt to predict naturally occurring phenomenon, such as when components on an air platform might fail from use, corrosion, or heat, or on planning problems where some form of optimization is sought. Rather, it draws attention to a potentially troublesome area where AI systems attempt to predict social phenomenon and behavior, particularly in the national security space.[8] This is where caution must be advised for the policy crowd. In this paper I do not discuss the laws of thermodynamics nor how to optimize for speed. Instead I discuss human behavior in complex, dynamic, and highly uncertain systems. Using AI

to predict ever more complex social phenomena, and then using those predictions as grounds for recommendations to senior leaders in national security, will become increasingly risky if we do not take stock of how these systems are built and the "knowledge" that they produce. Senior leaders and decision-makers may place too much reliance on these estimations without understanding their limitations.

In this paper I argue that greater efforts need to be made by the policy community to look under the hood of these AI prediction systems and gain some estimation of what "good" systems look like. For if those making recommendations to senior leaders do not understand how the systems work, then their estimations may be wildly incorrect and have serious national security implications. As one expert noted, "We do our models but then throw them over the wall," and "We don't know what happens to them, and we don't know specific cases when they cause change in policy."[9] There is a need for a more intensive endeavor to create a tech-policy nexus where the "wall" begins to crumble.

"GOOD SCIENCE"

One way to begin to break down the wall between technology and policy is to understand "good science" and where AI, in particular, may have some challenges within this paradigm. To start, philosophers of science have long argued about "inductive risk" and the "inductive gap" between theories and evidence. Any scientific theory, including social scientific theories, requires empirical observations to buttress the claims of the theory and other scientific statements. However, a gap exists between the generalized theory and the set of observations used to test and support it. This gap exists because "empirical observations never guarantee the truth of generalized conclusions" such as scientific knowledge.[10] That gap can be wider or narrower, depending upon the data and evidence, but there is always a gap. Making any leap across that gap involves risk that the prediction, estimation, or conclusion is wrong. This is called inductive risk. That risk assumes lesser or greater importance depending upon the consequences of being wrong.

Good science, many would argue, needs to minimize this gap as much as possible. However, what makes science "good?" Many cite the need for some formalized set of rules, such as the scientific method; others may claim that replication or reproducibility is needed, while others may claim that reliability and predictability are also core elements. There is also the time-honored notion that science deals with facts, not values. In some fields of study in the social sciences many argue that although explanatory power and scope are important, one should also balance these against parsimony. One's theory must "travel" and not be bound to a single point in time or place, but also not be overdetermined by including every possible variable. Debates in the philosophy of science, as well as within the scientific community, have shifted over time, but one of the core elements is that good science ties to good theories.

The problem we face with the rise of AI, especially related to systems for *predicting* social phenomena, is that these systems are not designed to jump the inductive gap. In fact, they live in the gap. They are typically atheoretical, or the theories are enmeshed and then parsed down to such an extent with proxy variables that they suffer from a weight of evidence problem.[11] While they may have massive amounts and varieties of data on which to draw (furthering a weight-of-evidence problem), they will never have *all* the data. To more clearly see why this is a problem, we return to the example of EMBERS.

EMBERS

EMBERS at its core is an events-based predictive meta-model developed by ten institutions and over seventy academics to forecast "socially

significant population-level events, such as civil unrest incidents, disease outbreaks, and election outcomes" from purely open-source data.[12] It was developed for the Intelligence Advanced Research Projects Activity (IARPA) Open Source Indicators (OSI) program. EMBERS "deployed" for several years, but only as a research activity.

As an events-based predictive AI model, its predictions either do or do not come to pass, and so there are ways of testing the accuracy of EMBERS predictions. In many cases, EMBERS accurately predicted a high percentage of civil unrest events.[13] This analysis will now interrogate some of the subcomponents, especially those dealing with sentiment analysis.

EMBERS comprises four subcomponents: data ingest, message enrichment, analytic modeling, and prediction fusion. The data input in EMBERS originates from a dozen different sources, including Twitter, newspapers, and government reports. Of most interest here are the text-based data. To be sure, there are correlated events, such as flu outbreaks and cold weather, or discussions about elections before a planned election. Yet, we should be mindful of the underlying science and the machine learning used to predict human behavior based on natural language.

Much of the natural language processing takes place mostly in the "enrichment" process. The first part of this process "uses linguistic information to both expand on the content of the textual content of the message and extract information contained in the text into a structured format."[14] In short, this process looks to various words, tags them as various parts of speech (noun, verb, object, date, place, and so forth), and shortens words to their roots (a process called lemmatization). The outputs from this process then are provided as inputs to further semantic and sentiment analysis. The sentiment analysis is but one area on which we should focus.

Sentiment analysis within EMBERS relies on the Affective Norms for English Words (ANEW) lexicon, which was created by Margaret Bradley and Peter Lang at the University of Florida in 1999.[15] EMBERS applies the ANEW lexicon, translates it into Spanish and Portuguese, and then arrives at a three-dimensional "sentiment score." Thus, when a word on the lexicon matches a data point ingested into EMBERS, the enrichment process provides that score as evidence of the level of pleasure, dominance, or arousal of the producer of that text. However, how much weight should be placed on the ANEW lexicon to determine sentiment scores?

We should look to the research design, methodology, and findings that Bradley and Lang provide. First, this lexicon was originally designed for English.[16] Although we can translate, those translations may in fact not carry the same meaning, weight, or affect in different populations or dialects.[17] Second, the experiments were conducted on introductory psychology students at the University of Florida as part of a course requirement. The population used to generalize to several continents is a group of students eighteen to twenty years old, from a particular part of the United States, with all of their demographic, cultural, and linguistic regularities. Third, the instructions for the experiment required the students to "bubble in" one of potentially nine points on a scale, and these were ranked ordinally. Each bubble has a corresponding figure along the nine-point range (for example, a smile ranging to a frown). Scores are summed on each dimension, and the mean score for each word is thus used as "the sentiment" score for that word. Students were shown between 100 and 150 words.

Immediately apparent in the ANEW lexicon is that the highest-ranking scores fell along general trends, in particular, trends along values associated with Western liberal democracy, capitalism, Christianity, heteronormativity, and status as a student. For example, religious terms used in the lexicon all refer to the Christian faith: Christmas, angel, heaven, hell, church, demon,

God, savior, devil, and so forth. There were no other terms in the lexicon to denote other faiths or belief systems. This may not constitute a "traveling problem" for a predominantly Christian Latin America, but this cannot be said for the Middle East and North Africa.[18] Moreover, the affect associated with these words tends to mirror the teachings of Christianity. Another region, faith, or population may not in fact feel the same way.

In addition, not only was there a strong heteronormative bias, but the structure of the available words on the questionnaire also demonstrates bias. For instance, there were twelve words relating to women (vagina, hooker, whore, wife, woman, girl, mother, rape, breast, abortion, lesbian, bride).[19] There were five for men (penis, man, brother, father, boy). The lack of any balance on this alone is alarming from an instrument design perspective. However, the scores associated with such words are also telling. In some cases, looking to the affective division between male and female respondents shows not merely their valuation of heteronormative roles but also underlying cultural connotations of devaluing gender stereotypes.[20]

Looking at just the higher-ranking scores, they show a particular cultural and even regional affect. For example, "diploma" and "graduate" garner some of the highest scores in the lexicon. This cultural, regional, and demographic skew is then taken by EMBERS and used to produce a sentiment analysis on a completely different region, population, demographic, and language.

Some might object that EMBERS's entire architecture and the amount of data collected and analyzed makes up for the difficulties with its sentiment analysis, and only the end product itself matters. This may be true, but we also must look to some of the limitations with the preprocessing of ingested text. These limitations involve the difficulties of accurately coding events-based data using natural-language processing.

Wei Wang and others used two projects—not identical to EMBERS—to assess the accuracy of events-based coding: the Integrated Crisis Early Warning System (ICEWS) and the Global Database of Events, Language, and Tone (GDELT). These are systems intended to forecast international crises for U.S. intelligence analysts. Even though Wang and others did not assess EMBERS in their study, they found "major discrepancies between automated systems that use primarily English-language corpus and a hand-coded system that uses news sources in both English and Spanish."[21] Automated coding demonstrates problems of event duplication—where consecutive reports of the same event is recorded as multiple events—as well as misclassification of event type. The noisiness of the data and the limitations of the textual extraction and classification leads to significant problems. In testing GDELT, for example, Wang and others found that the average accuracy for event category classification was 16.2 percent, but even the most accurate category, protest events, was 35.3 percent—worse than flipping a coin.[22] In short, the way in which we use AI for events-based coding is also subject to severe limitations because AI *cannot understand context* from the text it ingests.

Some might object and claim that EMBERS is sufficiently better than its predecessors, and it has one component neither ICEWS or GDELT can access: a ground-truth dataset from which to test its accuracy. This dataset, the Gold Standard Report (GSR), was produced by human analysts at MITRE and was a hand-coded events-based dataset of protest events for the ten Latin American countries observed. The dataset included the event description, location, and time stamp of the first mention by a major news source.[23] Access to such a ground truth can thus prove whether an event or action actually occurred, and it also acts as a check on the AI's ability to correctly process the text.

However, we must note that the GSR was produced for the first version of EMBERS, not its further expansion. Furthermore, the GSR was only good for a period of time set out by the original EMBERS project. Without ongoing human-coded, events-based data, the ability to test the validity of EMBERS's predictive capabilities dwindles. Therefore, in 2016, Parang Saraf and Naren Ramakrishnan proposed extending the GSR for additional periods by way of automation.[24] This "AutoGSR" acts as a persistent ground truth to check the validity of predictions. However, as Wang and others showed in their work on automated events-based coding, AI is not as reliable as a human coder. Although AutoGSR identified protest events for a three-month period in 2015 at a high accuracy and recall rate, the system is still susceptible to event duplication and false positives based on volume of reported stories. The ability of natural language processing (NLP) to have some form of global context awareness is still lacking. Moreover, and more important, AutoGSR cannot continue as a true ground truth for EMBERS. It is now a probabilistic system, just like ICEWS and GDELT, and although it may be more sophisticated in various ways, ultimately, it is no longer the gold standard.

IMPLICATIONS AND CONSIDERATIONS

Although this paper highlights some of the difficulties of the EMBERS project, it should not be interpreted as discounting its approach or its ability to generate interesting predictions. Instead, the aim is to take a very close look at the design of a system and perceive where problems may arise owing to the limitations of various parts of its architecture. The limitations of EMBERS notwithstanding, it is still an events-based prediction model, and although it is certainly dependent upon the behavior of people, it attempts to cross correlate and validate in various ways to reduce uncertainties in the prediction.

Nevertheless, there is an important lesson to be learned by assessing this sophisticated project: in multiple areas the AI either is limited in its abilities or is limited owing to the inadequacy of the model construct and/or data. These limitations become compounded for various reasons, such as appropriateness, time, volume, or sheer noise. Therefore we should begin to question the appropriateness and limitations of such tools, especially when they attempt to predict human behavior.

In the case of relying on textual evidence of human intent to act in a particular way, we need to establish good theories of human action. We can look to human psychology but also to sociology and political science for insights. Even these theories are not monolithic, however, and choosing one theory over another requires various justifications. Even in the case of EMBERS, decisions to use various theories of collective action, social identity, and group membership carry significant weight. Also, following Douglas, determining which evidence to weigh in a reliable and scientifically rigorous manner is no easy task.[25] Different theories of why people act in various ways require different forms of evidence, and in the realm of social phenomena, there is no way ethically or accurately to perform controlled human studies to claim definitively that all humans in situation X will do Y. The studies could never be complete because social, environmental, and cultural interactions are too complex. Thus, no one source of evidence is going to be definitive or provide a gold standard.[26] In short, we do not have access a priori to a ground truth about human actions in any specific situation; all we truly have is some level of correlation when many people chat about performing some action with other variables. Thus, predictive models of social phenomenona can at best rely on crudely coded past events, where those past events are shaped by a myriad of latent processes and variables unaccounted for by the relevant data.

Furthermore, we must begin to acknowledge that AI systems—especially those relying on machine learning—are not value-free and neutral. Rather, they are inherently biased because of their subjectivity. As Missy Cummings and Songpo Li argue, "Machine learning results can be greatly affected by the subjectivity of the machine learning practitioner, where the practitioner subjectively selects the machine learning algorithm and the algorithm parameters for a specific data set, and either this person or perhaps other people then interpret the results."[27] In addition, the machine learning practitioner can also inherit biases of previous researchers, as the case of ANEW shows: there are serious difficulties with the objectivity of the ANEW lexicon, as well as with the research design and instrument. Results from ANEW may be only reliable for a small and distinct population, and it would be inappropriate to generalize these findings beyond that population. Acknowledging this limitation is crucial where decisions concerning foreign policy or even foreign interventions are at stake, such as whether the United States provides aid, withdraws support, sends troops, or the like. We should be aware of the potential for inappropriate and biased data to lead to spurious results.

In this paper I used EMBERS—which sets a high standard—as a foil to highlight some of the underlying problems of using AI for predictive analytics for social phenomena. Yet despite EMBERS's creators' attempts at rigor, it too suffers from such problems. Thus, smaller studies or applications that do not have the ability or resources associated with such large projects should be equally scrutinized. For instance, in one rather infamous instance, researchers attempted to use facial image recognition photos to predict the likelihood of that person's being a criminal (which is a set of behaviors, not a feature of an individual). This rather confounding and ethically dubious endeavor highlights the conceptual errors of mixing classification and prediction.[28] We must not be lackadaisical and overly trusting of such systems. We must be increasingly careful in how we build and assess AI systems for prediction. These systems will suffer from bias, and they may never access a ground truth to assess.

CONCLUSION

The push for predictive analytics and forecasting tools will not slow down anytime soon. Policymakers, commanders, diplomats, and many more—dazzled by their promise—will increasingly rely on them. However, we should be careful not to make assumptions about the capabilities of prediction machines and should seek to understand how they work, when they do not, and why. In some instances, these could be purely technological limitations—the NLP does not pick up context well; in others, limitations could be due to human choices. The choices of assumptions in a model, the choices of variables or datasets, or the choices of one AI technique over another are all profoundly important when we consider that the goal of prediction in national security is not only to reduce uncertainty but to save lives, promote national interests, and defend against one's competitors and adversaries.

Human prediction may or may not be any better than an AI in some cases, and this is not to say that humans have some elevated status above their silicon counterparts.[29] When it comes to AI, we increasingly rely on systems we do not fully understand; techno-optimism may color our judgments about the predictions such systems make, and in even the best of systems we face serious epistemological challenges. Clearly, the nexus between the technology and policy communities needs to become more tightly linked. The designers and developers of AI prediction machines cannot solely focus on "just their part" and should "communicate uncertainty honestly to enable trustworthy assessment of what we do and do not know"; at the same time the policy community needs to become more comfortable learning about these systems and how they work.[30]

Despite the criticism of some of the elements of predictive systems, especially EMBERS, aired in this paper, there is great utility in pursuing these technologies. When predictive AI systems accurately present their assumptions, the sources of their data, and the choices of their architectures, features, and weights and attempt to ameliorate biases and communicate their limitations, this can greatly help human analysts, policymakers, and subject matter experts to better inform their decisions and judgments. However, we must continue to take notice that some systems will never be the Delphic Oracles we desire, and we may never be able to access a ground truth from which to test the accuracy of those predictions. Providing for causality in foreign affairs is (potentially) an insurmountable task, given that at any given time multiple factors may or may not be relevant depending upon the situation and the environment. Thus, we should accept the reality of remaining uncomfortable with the predictions and estimations of our AI world.

ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ANEW	Affective Norms for English Words
COA	Course of Action
EMBERS	Early Model Based Event Recognition using Surrogates
GDELT	Global Database of Events, Language, and Tone
GSR	Gold Standard Report
IARPA	Intelligence Advanced Research Projects Activity
ICEWS	Integrated Crisis Early Warning System
NLP	Natural Language Processing
OSI	Open Source Indicators

REFERENCES

1 U.S. Department of State, "Establishment of the Center for Analytics," press release, January 17, 2020 (www.state.gov/establishment-of-the-center-for-analytics/).

2 For example, see Defense Advanced Research Project Agency, "Big Mechanism" (www.darpa.mil/program/big-mechanism).

3 Abbreviations and acronyms used in this paper:

AI	Artificial Intelligence
ANEW	Affective Norms for English Words
COA	Course of Action
EMBERS	Early Model Based Event Recognition using Surrogates
GDELT	Global Database of Events, Language, and Tone
GSR	Gold Standard Report
IARPA	Intelligence Advanced Research Projects Activity
ICEWS	Integrated Crisis Early Warning System
NLP	National Language Processing
OSI	Open Source Indicators

4 Zhang Xun, Li Jiangtao, Zhang Xiaohu, Fu Jingying, and Wang Dongming, "Construction of a Geopolitical Environment Simulation and Prediction Platform Coupling Multi-Source Geopolitical Environmental Factors" *Science & Technology Review* 36, no, 3 (2018): 55–61; Abishur Prakash, "Algorithmic Foreign Policy" *Scientific American* 29 (August 2019; https://blogs.scientificamerican.com/observations/algorithmic-foreign-policy/).

5 The Defense Advanced Research Project Agency (DARPA) built the Integrated Conflict Early Warning System (ICEWS), and more recently the Intelligence Advanced Research Project Agency built the Early Model Based Event Recognition using Surrogates system (EMBERS). EMBERS originally was intended for use relating to Argentina, Brazil, Chile, Colombia, Ecuador, El Salvador, Mexico, Paraguay, and Venezuela. The Global Database of Events, Language, and Tone (GDELT) is a more transparent version of ICEWS.

6 One example (of much interest across all services) is the U.S. Navy's 2019 Automated Multi-System Course of Action Analysis using Artificial Intelligence program, which seeks to integrate the AEGIS weapon system with an AI mission planner that "enables faster than real-time COA generation and performance analysis in simulation, and support real-time and post mission analysis." See U.S. Navy, "Automated Multi-System Course of Action Analysis using Artificial Intelligence," Navy SBIR 2019.1—Topic N101-034, NAVSEA (www.navysbir.com/n19_1/N191-034.htm).

7 Andy Doyle, Graham Katz, Kristen Summers, Chris Ackermann, Illya Zavorin, Zunsik Lim, Sathappan Muthiah, and others, "Forecasting Significant Societal Events Using the Embers Streaming Predictive Analytics System," *Big Data* 2, no. 4 (December 2014): 185–95; Andy Doyle, Graham Katz, Kristen Summers, Chris Ackermann, Illya Zavorin, Zunsik Lim, Sathappan Muthiah, and others, "The EMBERS Architecture for Streaming Predictive Analytics," paper presented at IEEE International Conference on Big Data, Washington, October 27–30; Dipak Gupta, Sathappan Muthiah, David Mares, and Naren Ramakrishnan, "Forecasting Civil Strife: An Emerging Methodology," paper presented at the

Third International Conference on Human and Social Analytics, location TK, Date TK, 2017; Parang Saraf and Naren Ramakrishnan, "EMBERS AutoGSR: Automated Coding of Civil Unrest Events," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data* (August 2016): 599–608.

8 Sydney Freedberg, "Fix It Before It Breaks: SOCOM, JAIC Pioneer Predictive Maintenance AI Breaking Defense," February 19, 2019 (https://breakingdefense.com/2019/02/fix-it-before-it-breaks-socom-jaic-pioneer-predictive-maintenance-ai/).

9 Philip A. Schrodt, quoted in Beth McMurtrie, "Can We Predict the Next War?" *Chronicle of Higher Education*, October 13, 2014.

10 Kevin C. Elliott and Ted Richards, *Exploring Inductive Risk: Case Studies of Values in Science*, edited by Kevin Christopher Elliott and Ted Richards (Oxford University Press, 2017), 2.

11 A weight-of-evidence problem is how to determine or "weigh complex sets of evidence from multiple disciplines." See Heather Douglas, "Engagement for Progress: Applied Philosophy of Science in Context," *Synthese* 177 (2010): 328; D. Weed, "Weight of Evidence: A Review of Concept and Methods," *Risk Analysis* 25 (2010): 1545–57.

12 Doyle and others, "Forecasting Significant Societal Events," 185.

13 Sathappan Muthiah, Patrick Butler, Rupinder Paul Khandpur, Parang Saraf, Nathan Self, Alla Rozovskaya, Liang Zhao, and others, "EMBERS at 4 Years: Experiences Operating an Open Source Indicators Forecasting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016 (https://dl.acm.org/doi/10.1145/2939672.2939709).

14 Sathappan Muthiah and others, ""EMBERS at 4 Years," 189.

15 Margaret Bradley and Peter Lang, Note Text TK.

16 In 2007, several scholars translated the ANEW lexicon into Spanish and conducted their own analysis. Like Bradley and Lang, they too sampled undergraduate students, but from several Spanish universities. Their sample, however, was again, limited to a particular demographic and particular dialect, and their sample was grossly over-represented by women (560 women and 160 men). The authors also found "remarkable" statistical differences between the translated versions of ANEW and the original version. In short, there is a lot of emotional difference between the Spanish and the English. See Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña, "The Spanish Adaptation of ANEW (Affective Norms for English Words)," *Behavior Research Methods* 39, no. 3 (2007): 600–605 A later (2012) study translated ANEW into European Portuguese, with greater attention on language representativeness for their respondents. However, this study also relied on undergraduate and graduate students. The study's authors found statistically significant differences in their population from the American and Spanish studies after translation, and in some cases they were unable to even translate some of the English words to retain the same meaning. See Ana Paula Soares, Montserrat Comesaña, Ana P. Pinheiro, Alberto Simões, and Carla Sofia Frade, "The Adaptation of Affective Norms for English Words (ANEW) for European Portuguese," *Behavior Research Methods* 44 (2012): 256–69. Even these two translations still limit their generalizability to Latin America.

17 One need only think of the differences across English-language countries—Canada, the United States, Great Britain, Australia, New Zealand—to consider not only that some words have different meanings, but also that the way people feel about those words are considerably different.

18 EMBERS originally looked to Argentina, Brazil, Chile, Columbia, Ecuador, El Salvador, Mexico, Paraguay, and Venezuela. However, reports state that they expanded to twenty countries in Latin America and have begun to cover Iraq, Syria, Egypt, Bahrain, Jordan, Saudi Arabia and Libya. See Leah McGrath Goodman, "The EMBERS Project Can Predict the Future with Twitter," *Newsweek*, March 7, 2015 (www.newsweek.com/2015/03/20/embers-project-can-predict-future-twitter-312063.html).

19 Obviously, men can be raped, have breasts, and may engage in prostitution. However, the trend of this evidence and the cultural connotations throughout suggest they are directed at women.

20 For example, when asked about how much pleasure the word "lesbian" elicited, female students ranked the word at a 3.38. When male students were asked the same, they responded with a mean score of 6.00. Likewise, the word "whore" was scored by female students at 1.61, while their male counterparts ranked it at 3.92.

21 Wei Wang, Ryan Kennedy, David Lazer, and Naren Ramakrishnan, "Supplementary Materials for Growing Pains for Global Monitoring of Societal Events," *Science* 353, no. 6307 (September 2016): 1502.

22 Wang and others, "Supplementary Materials for Growing Pains."

23 Doyle and others, "Forecasting Significant Societal Events," 192; Wang and others, "Supplementary Materials for Growing Pains."

24 Parang Saraf and Naren Ramakrishnan, "EMBERS AutoGSR: Automated Coding of Civil Unrest Events," in *Proceedings of the 22nd ACM SIGKDD International Conference*, 599–608 (https://dl.acm.org/doi/10.1145/2939672.2939737).

25 Douglas, "Engagement for Progress."

26 Douglas, "Engagement for Progress."

27 Missy Cummings and Songpo Li, "Machine Learning Tools for Informing Transportation Technology and Policy," Duke University, Humans and Autonomy Laboratory, November 2019, 1 (http://hal.pratt.duke.edu/sites/hal.pratt.duke.edu/files/u39/HAL2019_2%5B1920%5D-min.pdf). Cummings and Li look at the subjectivity of machine learning practitioners in the realm of autonomous driving variables on the prediction of fatalities and injuries. However, their work applies here as well, for they are able to identify various opportunities for subjective decisions by those building machine learning models.

28 A paper titled "Automated Inference on Criminality using Face Images," by Xiaolin Wu and Xi Zhang, was posted in 2016 on an academic database called Arxiv. After an international outcry the authors took it down. However, the authors' written response to the criticisms is still incoherent. They write, "Our inquiry is to push the envelope and extend the research on automated face recognition from the biometric dimension (e.g., determining the race, gender, age, facial expression, etc.)

to the sociopsychological dimension." In short, they not only posit that some correlation of the shape of a face corresponds to whether that person breaks the law, but also that a machine learning agent can correlates the pixels in two-dimensional black and white photos with some "sociopsychological" dimension. See Xiaolin Wu and Xi Zhang, "Responses to Critiques on Machine Learning of Criminality Perceptions," Arxiv, May 26, 2017 (https://arxiv.org/pdf/1611.04135.pdf). More recently researchers from Harrisburg University in the United States also put forward a study stating that their AI "can likewise predict whether someone is a criminal, based solely on a picture of their face." The study was rejected for publication after a wide backlash by the AI community. See Sidney Fussell, "An Algorithm That 'Predicts' Criminality Based on a Face Sparks a Furor," *Wired*, June 24, 2020 (www.wired.com/story/algorithm-predicts-criminality-based-face-sparks-furor/).

29 Philip E. Tetlock and Dan Gardner, *Superforecasting: The Art and Science of Prediction* (New York: Broadway Books, 2016); D. Kahneman, P. Slovic, and A. Tversky, eds., *Judgement under Uncertainty: Heuristics and Biases* (Cambridge University Press, 1982); K. R. Hammond, *Human Judgement and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice* (Oxford University Press, 1996).

30 Charles F. Manski, "Communicating Uncertainty in Policy Analysis," *Proceedings of the National Academies of Science* 116, no. 16 (2019): 7635.

ABOUT THE AUTHOR

Dr. Heather Roff is a Senior Research Analyst in the National Security Analysis Department at the Johns Hopkins Applied Physics Laboratory and a nonresident fellow in the Foreign Policy program at the Brookings Institution. She may be reached at Heather.Roff@jhuapl.edu.