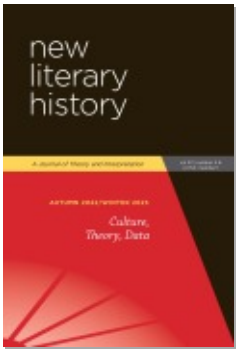
Culture, Theory, Data: An Introduction

Ted Underwood, Laura McGrath, Richard Jean So, Chad Wellmon



New Literary History, Volume 53, Number 4, Autumn 2022 / Volume 54, Number 1, Winter 2023, pp. 519-530 (Article)

Published by Johns Hopkins University Press

Culture, Theory, Data: An Introduction

Ted Underwood, Laura McGrath, Richard Jean So, Chad Wellmon

Culture. Theory. Data. In that group of three terms, the obvious interloper is *data*—a word not widely used in the humanities until this century. Not that humanists ever lacked data. Our sources and archives are data in the oldest sense of the word. Daniel Rosenberg's careful history of *data* concludes that it has long been "a rhetorical concept," meaning simply "that which is given prior to argument."[1] Of course, evidence is never given in an absolute sense. As Lisa Gitelman and Virginia Jackson have reminded us, every number in a table is collected and constructed by human hands.[2] But these constructions are converted into data when someone provisionally accepts them as given for the sake of a particular argument. Historical sources play a similar rhetorical role, and Rosenberg shows that the seventeenth century discussed them in a similar way, writing about "historical data" and even "scriptural data."[3]

By the late eighteenth century, however, *data* was specializing to contexts like "medicine, finance, natural history, and geography."[4] By the twentieth century, it had become strongly associated with the sciences and with systematic, numeric recordkeeping. Economists certainly had data. But historians believed they had data only when they were working with a document like a ledger or a census. The word felt alien enough that it was often invoked to mark a divide between the humanities and social sciences. In 1979, for instance, Tony Judt pushed back against social-scientific work in history by remarking that such projects "resort to quantified and quantifiable data to compensate for the lack of an argument and the glaring absence of conceptual insight."[5] If "conceptual insight" is akin to theory, Judt saw *data* as a poor substitute.

By the 1970s, *data* was becoming a metonym not only for the evidentiary procedures of science but for computers in particular. Lawrence Stone, criticizing quantitative work in the humanities around the same time as Judt, envisioned a scene where "squads of diligent assistants assemble data, encode it, programme it, and pass it through the maw

New Literary History, 2022 & 2023, 53 & 54: 519–530

of the computer."[6] Computers may seem less alien today than they did in 1979 (at any rate, they less commonly have a "maw"), but they are still strongly associated with the word *data*. So including *data* in the title of this special issue is probably enough to suggest that its essays will discuss the role that computational methods and digital media play in contemporary research on culture. Which they do.

On the other hand, as computers become increasingly embedded in daily life, *data* has expanded to cover a much wider range of things than it did forty years ago. The word still tends to imply that the information it describes is numeric or can be turned into numbers. But these days, what can't be turned into numbers? Even our recipes and home movies are transmitted digitally. Moreover, contemporary statistical models are no longer restricted to overtly quantitative problems with a small number of variables.[7] Machine learning can create variables as needed to model images, social networks, and unstructured text. Most of us regularly use models created this way to answer questions and find documents. Studying culture with data is no longer a specialized practice requiring "squads of diligent assistants" who "encode" variables for analysis. It is something most scholars already do, consciously or not.

For that reason, the title of this issue is not something like "Culture, Theory, Digital Humanities." Instead of advancing a subfield, we hope to explore theoretical questions that may be relevant to anyone studying culture, whether they use computational models explicitly in their research or "only" to translate unfamiliar languages, navigate social networks, and find sources. In shaping the event that brought these authors together (a symposium in Charlottesville, Virginia in May 2022), we invited people who work in media studies, information science, sociology, and anthropology, as well as historians and literary scholars. Literary history was often the center of our conversation, as you might expect from the title of this journal. But we came to that topic from many different angles.

The assumption guiding this symposium was simply that computation has become important enough in a wide range of disciplines to prompt a broad conversation about its implications for cultural theory. As recently as 2010, that would have been a rash assumption. It was still far from clear that computation could produce substantive insights in the humanities at all. Literary studies in particular was not "a 'counting' discipline," as James English concisely and correctly put it.[8] A writer using numbers to illuminate literary history in 2010 was well advised to skip theory and focus on providing some concrete results.

But we have seen many concrete results over the last decade. Computational models have been used to estimate the number of medieval books

missing from our libraries, to trace the changing senses of "freedom" and "justice" in abolitionist newspapers, and to contrast the visual styles of *Bewitched* and *I Dream of Jeannie*.[9] There is no longer much question that computation can produce cultural knowledge. The questions that remain are about the risks and implications of doing so. Whom are we empowering? What new disciplinary arrangements are we creating? What theories should guide our inquiries? For a conversation on that scale, we felt there should be social scientists in the room as well as humanists.

This is not the first time advances in information technology have drawn social scientists and humanists into conversation about new approaches to culture. As Bernard Geoghegan has pointed out, the new field of information theory exerted a powerful influence in the 1940s and 1950s.[10] The mathematical analogy between sound vibrations, written glyphs, and electrical signals helped prepare readers for the structuralist premise that all aspects of culture (from phonemes to food) could be understood as systems of communicative signs. Claude Lévi-Strauss ended *The Savage Mind* with a triumphant, three-page discussion of information theory, arguing that it validated the prehistoric tendency to see a "universe made up of meanings."[11] Fernand Braudel similarly cited information theory as support for his ambitious proposal to reunify history with other human sciences and organize them collectively around a new "qualitative mathematics."[12] Perhaps no ambitious proposal goes exactly as planned, but Geoghegan argues that aspirations borrowed from information theory and cybernetics long remained legible in the works of Roland Barthes, Michel Foucault, and Julia Kristeva.[13]

In recent years, digitization and machine learning have played a role analogous to the bridging role information theory played in the middle of the twentieth century. These technologies link different cultural domains—connecting sound to text, text to image, and image to spreadsheet. In doing so, they have also connected the humanities to the social and computational sciences, leading literary scholars to study mathematics and sociologists to read the rhetorical theory of Kenneth Burke.[14] It is possible that we are beginning to see the emergence of a new theoretical lingua franca, where words like "bias" and "model" draw meaning simultaneously from quantitative and qualitative disciplines.

But there are also reasons to be wary of these new connections. The transformation of culture into data is not an automatic or neutral process. If data is always constructed, we must ask who does the construction. What blind spots have been introduced along the way, and whose interests are being served? These questions will grow especially urgent if the relatively lightweight models produced by statistical machine learning continue to be displaced by deep learning technologies with a raven-

ous appetite for computation and data. At present, the most powerful language models are the private property of large corporations. If these models grow more capable while remaining proprietary, universities may confront serious competition. To return to our midcentury analogy: this wouldn't be the first time the discovery that culture is a system of signs has been followed by a more chastened discovery that power and knowledge are therefore mutually constitutive.[15]

Of course, the outcome of these contemporary struggles is impossible to predict. We have traced loose analogies to the aspirations of midcentury structuralism not because the twenty-first century is really doomed to follow the same path, but simply to convey the scale of the challenge and potential opportunity we see unfolding. The interpretive, methodological, and political questions that emerge when culture is represented as data come with real consequences and risks. In the pages that follow, twenty scholars explore those questions. They come from different disciplines, engage technology in different ways, and reach different conclusions. But we have found that their contributions can be organized around four shared rubrics.

What Kind of Meaning Do New Technologies Create?

Machine translation is a utility taken for granted on the internet, but as Hoyt Long points out, it "has rarely been taken seriously as an object of theorizing" in literary studies (722). Long's essay, "Learning to Live with Machine Translation," acknowledges and categorizes the failures of machine translation but goes on to ask "what good enough machine translation of literary texts might be good for" if we could view a model as a collaborator rather than a replacement for human reading (723). What could we learn about literary texts from an imperfect translation—or even, perhaps, from the failures of translation?

In a loosely similar way, N. Katherine Hayles takes a closer look at generative language models, such as GPT-3, that have often been dismissed as "simply lacking in meaning" (661). Hayles agrees that models don't create meaning in quite the same way as human utterances do—for one thing, a language model can't interact with the world in the same way we do, and its words have a certain "fragility of reference" (636). But she nevertheless concludes that the utterances of language models deserve interpretation and can have a kind of meaning grounded in the *Umwelt*, or world-horizon, of the model.

Although the internet is not governed by a single algorithm, we might think of a platform like *Medium* as a computational system that

generates meaning (in collaboration, to be sure, with human authors). In "Content's Forms," Tess McNulty inquires about the kinds of meaning conveyed by new genres of internet "content" and the theories we might need to understand them. Combining close and distant reading, she argues that the genres that thrive on *Medium* represent the human person "not as a fixed entity, but rather as a mysterious object of perpetual, experimental discovery" (842).

How Should Computation Acknowledge Subjectivity?

While computational methods have become increasingly common in literary criticism over the last decade, there is still little agreement on what they're good for or where their utility comes to an end. Michael Gavin tackles this question head-on by asking "Why Distant Reading Works." In particular, why should we assume that the aggregation of texts in a corpus tells us anything meaningful about the past? Gavin's answer is that we can't rely on corpora to tell us the objective truth of the past. Even if all documents were preserved (which of course they are not), the principle of "selective attention" guarantees that many aspects of the past would be left out of a textual record. But precisely because attention is selective, we can rely on language to reveal what authors believed would be relevant to their audience. In that particular sense, variations in a corpus are guaranteed to map variations in the "cognitive environment" that produced it (620).

If Gavin asks how texts represent the past, Katherine Bode asks whether scholars should understand their work as representing the literary past at all—or whether it would be more accurate to say that they constitute it and enact it. In discussing the performativity of literary studies, Bode is careful not to insist on a crisp separation from science. On the contrary, scientific disciplines have their own kind of performativity. Bode sees problems emerging when scholars imagine that computational methods compel them to approach writing as a collection of natural objects that pre-exist scholarly inquiry. This assumption can be used either to exclude computational methods from literary study or to argue that literary study must become a (naïvely representational) science. Bode recommends instead a flexible epistemology that holds its objects *"loosely, but not lightly"* (536).

Approaching computational epistemology from a sociological perspective, Laura K. Nelson is more willing than Gavin or Bode to endorse objectivity. But "embodied" objectivity, in her account, accepts "that all views are necessarily views from somewhere" and that "subjectivity [is]

necessary for producing objective knowledge" (855). Nelson acknowledges that computational methods have often been used to advance a simpler conception of objectivity, which she identifies with the universal laws of a "social physics" (857). But she argues this simplification is not a necessary consequence of computation. On the contrary, "computational methods . . . model the depth and complexity of the multiple perspectives captured in large data" and can make the researcher's own subjective choices more visible (861). She illustrates this point by measuring intersectional relationships between embodied perspectives in a collection of first-person narratives from the nineteenth-century US.

Where is the Cultural Criticism in Computational Approaches to Culture?

Digital and computational approaches to culture have been shadowed by questions about their capacity to support critique at least since 2012, when Alan Liu asked, "Where is the cultural criticism in the digital humanities?"[16] Some skepticism stems from the nefarious uses to which data has been put; data, in Jessica Marie Johnson and Sarah Bruno's words, has long been the "raw material of empire" (583). In "Que Recogan Este Memoria," Johnson and Bruno argue that an embrace of Black feminist digital praxis challenges the dehumanizing effects of colonialism, recenters the human, and can "potentially dismantle imbalances of epistemic power at the structural level" (589). While digital humanists have inherited the legacy of colonialism, this need not define the field moving forward. Drawing on two cases of data deformation in Puerto Rican history, Johnson and Bruno challenge researchers to sift through the data of the transatlantic slave trade in a manner that "may help set out new terms for the survival of Black life in the future" (588).

In "Medium Specific Sexuality," Joan Lubin traces a "computational history of sexuality" to the Kinsey Reports, published in 1948 and 1953. These landmark studies "introduced big data into the study of human sexuality" (758) by converting individual narrative into computational data and "marked a tipping point in the datafication of the discourse of sex" (759), changing both the study of human sexuality and its literary archive. Lubin asks, "What happens to the literary archive of sexology when sexology becomes more computational than narrative?" (759). Lubin shows how the relationship between sexology and the novel was transformed, "undermin[ing] high literary values . . . and accelerat[ing] the popularization of new mass cultural genres" (761). Computation shaped historic understanding of sexuality, changing the course of the novel.

Although computational methods have become increasingly common, there remains some skepticism about what they can actually show us—what the use of computational modeling or machine learning might add to ongoing disciplinary conversations or how these findings might contribute to, inform, or enrich our understanding of culture, let alone change it. Two essays interrogate dominant cultural narratives about the so-called "racial awakening" in the aftermath of George Floyd's murder in June of 2020. Laura McGrath examines what this "awakening" did—and didn't do—in the literary world. Bringing computational methods to bear on literary sociology, McGrath examines the fields of creation, production, and reception through an analysis of book-deal announcements. Has the literary field become more equitable in the past twenty years, or in the past two years, as many claim? She shows how "race" circulates as a form of currency for book marketing and promotion, reinforcing dominant literary narratives, such as the metaslave novel and the multigenerational immigrant narrative, even while claims about the changes within the industry ring hollow.

While McGrath's account examines one industry's response to a cultural moment, challenging salutary claims of change, Long Le-Khac, Maria Antoniak, and Richard Jean So take up the same moment to offer new insights into how cultural change occurs in the first place. Analyzing a corpus of #BlackLivesMatter tweets from the days and weeks following Floyd's murder, Le-Khac, Antoniak, and So investigate the "intertwining of social movements with new communications technologies" to "study the relationship between race, discourse, and social change" (668). By showing how the discourse of #BLM was eventually absorbed by the racial discourse of "colorblindness," they add to Raymond Williams's influential, three-part model of cultural evolution (the dominant, the emergent, and the residual) the theory of *the insurgent*: "a long-building minority cultural strain that surges to contest the dominant culture in a moment of crisis" (670).

How Can We Bridge the Difference of Scale between Narrative and Social Data?

While debates about "distant reading" have made it easy to think about scale as a characteristic that distinguishes one critical method from another, several authors contributing to this special issue are also interested in scale as a feature of narrative itself. In "Literary Studies and Collective Life," Caroline Levine notes that the literary has often been defined in a way that privileges a relatively small scale of social

description. Novels tend to focus on "a small group of central characters," and literary critical methods similarly prize "the singular moment of alterity or disruption" (699). To draw out the full political potential of literature, Levine argues, we need to learn to move between forms and scales, finding the hinges that connect small, local actions to the data of collective life.

Alison Booth finds a similar kind of mediating connection in the genre of biographical nonfiction, which "has special affordances for combining the individual and the typical" (560). The scale of collective biography also lends itself to an approach Booth calls "mid-range reading," which expands beyond a single book without claiming to represent literature as a whole (564). What emerge instead are networks of a few dozen or a few hundred lives linked by provisional categories but also attentive to handcoded particularitities.

The claim that novels tend to focus on a relatively small scale of social description receives some empirical support from Andrew Piper and Sunyam Bagga's essay, "Toward a Data-Driven Theory of Narrativity." The probabilistic approach adopted by Piper and Bagga doesn't require them to select a definitive or essential characteristic of narrative. Instead, they consider narrativity as a question of degree and look for a short list of features that tend to characterize narrative, even if those features don't always occur together. The models that emerge from this process imply that readers "are most confident of narrativity when [they] encounter a highly focalized set of agents set at a distance to the teller who are situated within a particular time and place" (895).

Dennis Tenen is interested in stories that might seem to stretch this concrete and agent-centered theory of narrative: stories where "airports, hospitals, hotels, and corporations emerge as powerful characters in their own right" (904). To characterize a genre he calls the "organizational novel," Tenen focuses on the syntactic role played by distributed agents (905). While they may be central to the story, entities such as the Airport or Railroad tend to receive, rather than initiate, action. Along the way, Tenen defends exploratory data analysis from models of computational method that center hypothesis-testing.

In the case of a queer canon, Matt Warner argues, scale is not an affordance; in fact, "Queer people have longstanding reason to aspire to keep the data concerning themselves as small as possible" (940). The emphasis on large-scale data has resulted in an absence of queer studies in cultural analytics, because "big queer data might not be that big" (940). Taking a theoretical cue from the history of queer bibliographers, including many amateurs currently building lists and assembling canons online—"queer readers have always been invested in uncovering and

keeping track of stories about people like us"—Warner argues that a queer computational literary criticism must embrace data that is "smaller, quainter, and more marginal" (940). Warner uses these existing bibliographies to build a queer corpus, showing how even smaller data can yield considerable insights for computational analysis.

* * *

Our double special issue ends with two shorter response essays. Roopika Risam draws attention to the digital paratexts that could help document literary practice—especially postcolonial literary practice—if only they were being preserved. Clayton Childress asks why computation seems to be bringing humanists and cultural sociologists into conversation once again. He finds four good reasons to use computational methods that illuminate what these methods can and should do for us, along with one bad reason.

* * *

The essays summarized above certainly don't reach consensus on any of the questions they explore. But despite these differences of opinion, we have been struck by the coherence of the conversations between and among disciplines that we observed in Charlottesville. Historians, literary critics, and sociologists seem to be grappling with similar questions about the limits and affordances of computation. We are united by shared problems—and, to a surprising extent, by a shared lingua franca. Some fields may use large datasets. Others tend to close read a smaller sample of books. But many of us are comfortable talking about "samples" and "models" (of various sizes). Few of us are shocked to discover that knowledge is constructed by interested participants, and few of us feel that the word *data* belongs exclusively to the sciences.

We can't necessarily speak for all the contributors, but the four coeditors of this issue see this convergence of concerns and vocabularies as a hopeful sign. The last decade has been a dark one for disciplines that study culture. Literary scholars in particular have faced hard questions about our disciplinary prospects and about the future we promise graduate students. The intensity of internal debate about different modes of reading (close versus distant, surface versus deep) sometimes seems less proof of continued life than a symptom of our desperate search for it.

This may be a good time to build bridges and look outward. Literary theory has often mattered most when it contributed to a broader conversation about writing and culture where other disciplines and institutions also had roles to play. Perhaps we are once again entering one of those productively messy periods. Certainly stories and short dramatic scenes have never been more intricately interwoven with real life than they are on the internet. Literary scholars have long claimed that fiction gives writers and readers a way to refashion identities in response to social change. Now we can actually see that happening from week to week, as dramatic templates mutate on social media or in semifictional subreddits such as "Am I the Asshole?"

Models of culture are also becoming consequential in newly direct and visible ways. Language models like GPT-3 have attracted attention by mastering an immense repertoire of forms, voices, and genres that they remix as requested.[17] In truth, these models encode not just language but culture. The consequences of using a cultural model generatively—to produce yet more culture—are still far from clear. But it's a question that humanists and social scientists urgently need to answer. With apologies to Karl Marx: our models of culture have hitherto interpreted the world in various ways. Generative models, for good or ill, are liable to change it.

Can scholars really intervene in cultural practices mediated by such complex technologies? We are convinced that they must, and the essays we have collected here convince us that they can. Most generative models and internet platforms are owned by large corporations. Without scrutiny, technological mediation of culture will almost certainly heighten existing inequalities and biases.[18] So the need for critique is clear. Many of the essays in this special issue respond to that need, developing critical accounts of new data-gathering practices, internet genres, and technologically mediated social movements.

In some cases, it is possible for cultural theory not only to critique technology but to guide it directly. Generative models of culture are reproducing patterns that have long preoccupied humanists. If our cultural theories have any validity, they ought to be practically useful here. Admittedly, the theories that currently underpin language models are often radically simplified, ignoring social mediation to posit an observer who confronts a sea of undifferentiated text. But more complex theories produce better results. One valid criticism of GPT-3, for instance, was that the theory undergirding it lacked any model of the social context of language.[19] It is no accident that ChatGPT improved on GPT-3 by retraining the model with a clearly specified dialogic frame and with extensive feedback from a human audience.[20] Even a small amount of social framing made a difference.

It is admittedly alarming that we are now all part of an experiment where theories of language and culture get tested on live subjects. But the silver lining, for what it's worth, is that expertise in cultural theory has a newly critical role to play. Natural language processing is strong enough now at the paragraph level that the next thing it needs may really be a better account of the social contexts and purposes of human communicative practices. The theories of translation, narrative, and interpretation advanced in this special issue were not necessarily intended to contribute to that task. But, intentionally or not, cultural theories are likely to play an increasingly explicit role in guiding technology—in part because the instructions for language models are increasingly written in natural language rather than code.

In short, the convergence of culture and computing presents humanists and social scientists with an opportunity to make a difference in several new ways. We have the critical perspectives, theories, and metadata needed to inform an emerging project with potentially massive social consequences. Those consequences are not guaranteed to be good, and trying to provide course correction for them will sometimes force us out of our comfort zone. But that is often true when scholars engage big contemporary problems. In any case, reflection on the relation between data and culture is clearly urgent, and we hope the essays gathered here have laid a foundation for it.

University of Illinois Urbana–Champaign

Temple University

McGill University

University of Virginia

NOTES

1 Daniel Rosenberg, "Data Before the Fact," in "*Raw Data" is an Oxymoron*, ed. Lisa Gitelman (Cambridge, MA: MIT Press, 2013), 36.

2 Gitelman and Virginia Jackson, "Introduction," *"Raw Data" is an Oxymoron*, 2–3.

3 Rosenberg, "Data Before the Fact," 19.

4 Rosenberg, "Data Before the Fact," 33.

5 Tony Judt, "A Clown in Regal Purple: Social History and the Historians," *History Workshop Journal* 7 (1979): 75; quoted in Steven Ruggles, "The Revival of Quantification: Reflections on Old New Histories," *Social Science History* 45, no.1 (2021): 13.

6 Lawrence Stone, "The Revival of Narrative: Reflections on a New Old History," *Past & Present* 85 (1979): 6. Quoted in Ruggles, "The Revival of Quantification," 12.

7 For the underlying cause of this shift, see Leo Breiman, "Statistical Modeling: The Two Cultures," *Statistical Science* 16, no. 3 (2001): 199–231.

8 James F. English, "Everywhere and Nowhere: The Sociology of Literature After 'the Sociology of Literature'," *New Literary History* 41, no. 2 (2010): xii.

9 Mike Kestemont, Folgert Karsdorp, Elisabeth de Bruijn, Matthew Driscoll, Katarzyna A. Kapitan, Pádraig Ó Macháin, Daniel Sawyer, Remco Sleiderink, and Anne Chao, "Forgotten Books: The Application of Unseen Species Models to the Survival of Culture," *Science* 375, no. 6582 (2022): 765–69; Sandeep Soni, Lauren F. Klein, Jacob Eisenstein, "Abolitionist Networks: Modeling Language Change in Nineteenth-Century Activist Newspapers," *Journal of Cultural Analytics* 6, no. 1 (2021); and Taylor Arnold, Lauren Tilton, and Annie Berke, "Visual Style in Two Network Era Sitcoms," *Journal of Cultural Analytics* 4, no. 2 (2019). https://doi.org/10.22148/

10 Bernard Dionysius Geoghegan, "From Information Theory to French Theory: Jakobson, Lévi-Strauss, and the Cybernetic Apparatus," *Critical Inquiry* 38, no. 1 (2011): 96–126.

11 Claude Lévi-Strauss, *The Savage Mind* (Chicago: Univ. of Chicago Press, 1966), 267.

12 Fernand Braudel, "History and the Social Sciences: *The Longue Durée*," trans. Immanuel Wallerstein, *Review (Fernand Braudel Center)* 32, no. 2 (2009): 171–203. Originally published as Braudel, "Histoire et Sciences Sociales: La Longue Durée," *Annales ESC*, 13, no. 1 (1958): 725–53.

13 Geoghegan, "From Information Theory to French Theory," 123–26.

14 John W. Mohr, Robin Wagner-Pacifici, and Ronald L. Breiger, "Toward a Computational Hermeneutics," *Big Data and Society* 2, no. 2 (2015).

15 "Between techniques of knowledge and strategies of power, there is no exteriority." Michel Foucault, *The History of Sexuality*, trans. Robert Hurley (New York: Pantheon, 1978), 1:98.

16 Alan Liu, "Where is the Cultural Criticism in the Digital Humanities?," in *Debates in the Digital Humanities*, ed. Matthew K. Gold (Minneapolis: Univ. of Minnesota Press, 2012), 490–509.

17 Tom B. Brown et al., "Language Models Are Few-Shot Learners," *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, https://arxiv.org/abs/2005.14165.

18 Jenny L. Davis, Apryl Williams, and Michael W. Yang, "Algorithmic Reparation," *Big Data & Society* 8, no. 2 (2021); Laura D. Tyson and John Zysman, "Automation, AI, & Work," *Daedalus* 151, no. 2 (2022): 256–71; and Erik Brynjolfsson, "The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence," *Daedalus* 151, no. 2 (2022): 272–87.

19 "Text generated by an LM is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind. It can't have been, because the training data never included sharing thoughts with a listener . . . ." Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Margaret Mitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 616.

20 Some key advances toward this goal were made in Nisan Stiennon et al., "Learning to Summarize from Human Feedback," *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, https://arxiv.org/abs/2009.01325.