

On Political Theory and Large Language Models

Political Theory

1–33

© The Author(s) 2023



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/00905917231200826

journals.sagepub.com/home/ptxEmma Rodman¹ 

Abstract

Political theory as a discipline has long been skeptical of computational methods. In this paper, I argue that it is time for theory to make a perspectival shift on these methods. Specifically, we should consider integrating recently developed generative large language models like GPT-4 as tools to support our creative work as theorists. Ultimately, I suggest that political theorists should embrace this technology as a method of supporting our capacity for creativity—but that we should do so in a way that is mindful of the content and value of theorizing, the technical constraints of the models, and the ethical questions that the technology raises.

Keywords

methodology in political theory, creativity, judgment, computational linguistics, text-as-data

Introduction

In 1968, long before the development of modern large language models, the National Research Council’s Behavioral Sciences Committee imagined an “ideal” computer system that would provide researchers with the “computer analogue of the intelligent, all-informed colleague.” The system would be the perfect collaborator: it would have “read widely, have total recall, synthesize

¹Department of Political Science, University of Massachusetts Lowell, Lowell, MA, USA

Corresponding Author:

Emma Rodman, Department of Political Science, University of Massachusetts Lowell, Dugan Hall, Suite 201, 883 Broadway St., Lowell, MA 01854, USA.

Email: emma_rodman@uml.edu

new ideas, always be accessible,” and suggest “facts or literature of interest.” It could also respond to requests for “data and documentation” and react intelligently to a researcher’s work (“analyze its logic, trace implications, suggest tests”).¹ Such a system, the committee imagined, would dramatically improve the process of research in the social sciences.

A year after the committee’s report, as triumphant technologists were putting a man on the moon, Sheldon Wolin (1969) published a defense of political theory in the *American Political Science Review*. Given the rising commitment to what he termed “methodism” and behavioralism, as well as the breathless championing of science and technology, he worried that political theory was being devalued and misunderstood. The National Research Council’s description of the ideal computer system was one example Wolin highlighted in his essay. To him, it “disclosed the fantasies of the behavioral scientist about theories” and theory building. Their view “trivializes what is involved in a theory’s formulation and thereby obfuscates the importance of the choice among rival ways of constituting the world” (1075). The computer “colleague”—with all of its impressive capacities for recall, memory, synthesis, and logical analysis—misses something essential about what it means to do theory and, relatedly, why theory matters. The fantasy of automating knowledge elides the inextricably human activities of theorizing, which Wolin described as political judgment, contextual understanding, creativity, and vision.

This essay begins from Wolin’s worry. In 1969, the fantasy computer system merely reflected the “contemporary mood” rather than a practical possibility; today, such a system has become a tangible reality of scholarly research (1075). Publicly available large language models can now do many, if not all, of the tasks that the National Research Council dreamed about a half-century ago. Contemporary computational text analysis is being widely used and valorized in empirical studies in a way that seems to validate Wolin’s concerns about hegemonic “methodism.” We appear to have reached a moment where the meaning and value of theory again require defense, this time against the widespread fetishization of big data and machine learning; what has been termed neo-positivism confronts us as “a naïve empiricism for the digital age” (Skees 2022, 147).²

-
1. “The Profession: NRC Behavioral Sciences Committee Report,” 1968, *Political Science: Newsletter of the American Political Science Association* 1(1): 25, col. 1.
 2. See also Fuchs (2017); Gardiner (2022). The threat posed by these technologies has been described as not only epistemological but also ontological and political (Amoore 2020; Kehlenbach 2022).

This paper, however, approaches these technologies with a different and, I think, more foundational question: was Wolin right? Is the computer system he described inescapably tied to empiricist logic and thus always the enemy of theory? Or, as I argue in this paper, is our answer contingent—both on *how* we use such tools and on *what* they are used to study? If that is the case, it is time for political theory to make a perspectival shift on such technology. We can and should use large language models to support our work *on our own terms*—and only to the extent that we can do so without compromising either theory’s content or its value. In that sense, this paper points us toward a methodology for using large language models for political theory, as political theorists. It is neither inevitable nor desirable for these tools to be used solely for empirical studies. In fact, I suggest that—both technically and ethically—large language models are best suited for creative and humanistic research rather than empirical research.

I begin the paper by describing what I take to be the content and value of political theory, particularly as contrasted with the content and value of empirical social science. I argue that political theory, like other humanistic fields, is distinguished from other subfields of political science by its emphasis on judgment and creativity, not on the production of empirical truths. In section two, I build out an account of what such creativity consists in. I then circle back to Wolin’s concerns by thinking alongside studies of machine creativity and creativity in artificial intelligence. In sections three and four, I ask: can we meaningfully reimagine computer systems, like the one Wolin describes, not as tools for empirical methods but as tools for political theory *qua* theory? What, if anything, do large language models have to offer the creative and imaginative theoretical process? I suggest that we can, in fact, develop new methodological approaches to these technologies, approaches oriented to theoretical rather than empirical enterprises. I offer the outline of one such approach, which builds on outputs from a recently developed large language model (GPT-4). Finally, I conclude by raising my own set of concerns and questions about the scholarly use of large language models, with a particular focus on empirical studies, pedagogy, and global justice.

What Is Political Theory?

What is distinctive about political theory? Wolin’s bifurcation between theory and empiricism is widely echoed; a broadly unifying theme across our diverse subfield is that, whatever else we are, we are not social

scientists.³ Social science's *vita methodica*, with its commitment to “objectivity, detachment, fidelity to fact, and deference to intersubjective verification,” presupposes a world that, through the application of scientific methods, can be made to yield truth statements that are rigorous, precise, and quantifiable (Wolin 1969, 1063). This empiricism assumes that the only valuable objects of study are those that are tractable to its chosen techniques and tools of investigation. It also precludes broadly normative questions about how the political world could or should be.⁴ Political theory, on the other hand, is the pursuit of what Wolin terms political knowledge or wisdom: an investigation of political questions whose answers are never final or true in some transcendent or foundational sense. Theorists approach these questions by way of judgment, reflection, argument, values, context, “an indwelling or rumination,” and an “astonishment at the variety and subtle interconnection of things” (1071, 1073).

But more is at stake here than just a choice of what and how we study. The most central issue for Wolin—and, I think, the least controversial of his worries—is his concern about the preservation of the distinctive human capacities we use when theorizing. Political theory is a humanism in the sense that “the most significant task of political theory is the protection of the political—the human—itself” (McWilliams 2015, 196). Empiricism, by contrast, is “a form of discipline designed to compensate for”—that is, to devalue and erode—what it sees as certain “unfortunate proclivities of the mind” (Wolin 1969, 1067). When Descartes writes that he is “amazed when I consider how

-
3. Although this view is not without detractors (e.g., Blau 2015; Dienstag 2016), it is generally held. Berlin (2014), for instance, offers a distinction between empirical and formal questions on the one hand and philosophical ones on the other, while the *Oxford Handbook of Political Theory* highlights the “irreducibly normative component” that unifies political theory (Dryzek, Honig, and Phillips 2008). Another way to put it is that political theorists “do not aim at empirical explanation or prediction and instead deal with the realms of ideas, concepts, texts, values, and norms” (March 2009, 534). It should be noted that opponents of theory also rely on a distinction between social science and theory—for example, that humanistic political theory is incompatible with the social scientific method of political science (Rehfeld 2010).
 4. Of course, there are normative theorists who intentionally work closely with, or are highly responsive to, empirical data and case studies (for instance, some theorists working on questions of deliberative democracy, e.g., Bagg 2018).

weak my mind is and how prone to error” or when Richard Hooker urges restraint on “the mind of man that it may not wax over-wise,” Wolin sees empiricism’s skepticism of capacities like reflectiveness, judgment, creativity, and vision.⁵ This is why, in part, he worried about the required empirical methods training of political science graduate students. Wolin is concerned about the effects of this disciplining of the mind: “The impoverishment of education by the demands of methodism,” he writes, “poses a threat not only to so-called normative or traditional political theory, but to the scientific imagination as well” (1073). To theorize is to be mentally open; to draw on “diverse, even ill-assorted baggage”; and to be trained within a “meditative culture which nourishes all creativity” (1071, 1073). In defining itself against empiricism, theory seeks to preserve these human capacities of judgment and creativity from colonization and destruction.⁶

Rather than taming the wayward proclivities of the mind in order to better discover facts in the world, theorists recognize that facts are perpetually underdetermined. It is only through exercising judgment and creativity that we are able to give an account of them, and the sense we make is always contingent and contestable. Democratic theorists, to give one example, commonly rely upon the “all-affected principle”—that is, the idea that everyone affected by a governing structure has a right to participate in shaping it—as an answer to the “boundary problem” of how to legitimately constitute, in an original sense, membership in the polity.⁷ If we accept the all-affected principle, our next logical step might seem to be social science: to empirically discover who is affected. But the empiricism of that question, as Fraser (2009) has written, is deeply fraught:

The problem is that, given the so-called “butterfly effect,” one can adduce empirical evidence that just about everyone is affected by just about everything.

-
5. Quoted in Wolin 1969, 1066–67. Cartesian skepticism also provokes mental closure: having once identified a seemingly trustworthy path via procedures of scientific method, the empiricist continues along them with “rigidity and single-mindedness” (1068).
 6. Brown (2002) suggests that this defensiveness is always the framework within which we think when we attempt to define political theory: our definition seeks to protectively exclude “that which rivals or displaces the enterprise or that contains the specter of its colonization” (557).
 7. For a classic account of the boundary problem and the all-affected principle, see Whelan (1983).

What is needed, therefore, is a way of distinguishing those levels and kinds of effectivity that are deemed sufficient to confer moral standing from those that are not. Normal social science, however, cannot supply such criteria. On the contrary, to operationalize the all-affected principle requires complex political judgments. (40)

Notice that Fraser's account implies a two-step process: we first imagine many possible ways and degrees to which possible members might be affected by the polity, and then we make complex political judgments about which of these should confer membership in the polity. In the first step, theorists draw on the creative faculties of the mind; in the second step, theorists choose, and they do this by drawing various modes of thinking together—"some evidentiary, some interpretive, some normative, some historical, some conceptual"—to form "a wide-ranging, open-ended" capacity for making such political judgments (41).

There is, of course, no settled method for precisely *how* to judge political questions (or, we might say, make political arguments).⁸ Berlin (2014) has argued, in fact, that this is a hallmark of making such judgments: "No obvious method of settling these questions lies to hand . . . there is no automatic technique, no universally recognized expertise . . . for accepting or rejecting earlier answers to these questions" (191). I take this, in fact, to be the epistemological basis for political theory's heterogeneity, at least to some degree. So while I agree that we lack a "settled consensus on the meaning and purpose of political theory," this is somewhat paradoxically a product of our shared general commitment to the distinctively human capacities of creativity and judgment (March 2009, 533).⁹ It is this epistemological humanism of

8. I draw on Kant (2001) and Arendt (1982) in characterizing judgment as a distinctively human practice of mind in this paper. *Pace* Arendt, judgment involves the subject's construction of meaning: either retrospectively in an assessment of what has occurred (we can see this particularly in the history of political thought) or in the context of the possibility of political action in the present moment (especially apparent in normative and critical theory). Reflective judgment—to borrow Kant's language—is what allows us to make meaning (Kant would say "construct universals") out of our experience of particularity. A full consideration of judgment as a faculty must be deferred to another paper, though I return briefly to the question of judgment later when I consider the idea of judging whether or not something is creative. On Kant and Arendt's ideas of judgment, see Zerilli (2016); Nedelsky and Beiner (2001).

9. I also agree that attempts to remedy this lack of disciplinary unity are suspect (Gunnell 1988; Kaufman-Osborn 2010).

political theory that defines it against the methodological empiricism of social science.

What Is Creativity?

In the last section, I glossed the theorist's creative capacity as generative—that is, as the moment at which a theorist's mind opens up the myriad of possibilities inherent in a given question or, meta-cognitively, in the formation of the questions themselves. This capacity is what psychologists refer to as divergent thinking: a form of imaginative and playful thinking that generates wide-ranging, numerous, and varied ideas in response to open-ended tasks or prompts (Runco 2014). To do this kind of thinking well implies producing a large number of unusual or unique ideas that extend across a wide range of varying categories. The kinds of questions used on psychological tests for divergent thinking analogize to various instances of creative thinking that theorists do (e.g., think of titles for a story, list the consequences of the world being suddenly covered in water, generate uses of a common object like a brick or coat hanger). Divergent thinking is the capacity to generate a large number of potentially creative ideas in response to such prompts. This cognitive capacity is mediated, of course, by an element of chance as well as one's accumulated prior knowledge and experiences; together, they determine just how creative one will be in any given situation. Someone's ability to imagine the varied consequences of a world underwater depends on their capacity for divergent thinking, but it may also depend on how recently they last saw the film *Waterworld*.

Divergent thinking, however, is only the beginning of creativity; it is necessary but not sufficient. We recognize intuitively that we might generate a large number of unique and unexpected ideas without any of them being truly creative. What, then, is creativity? To be creative, an idea must have three attributes: it must be novel, surprising, and valuable.¹⁰ A creative idea is something new, something that does not follow in an obvious way from things that are already known, and this novel and surprising idea must be understood by someone or some set of persons as having value. Divergent thinking is necessary in order to produce ideas that are potentially novel and surprising, but creativity requires a third piece as well: an act of judgment to determine whether the ideas are truly novel and surprising, and also whether they have value or utility. So, while creativity is not randomness, it is also not

10. This formulation is indebted to Boden (2004).

order. To be novel, surprising, and of value, a creative idea must unexpectedly exceed the rational confines of prior ideas while at the same time remaining intelligible and valuable within an established vernacular. That is to say, it is surprising: it pushes the limits of what is expected given what is already known.

A few clarifying notes and points of emphasis are in order. First, an idea might be novel to me or to my discipline, and thus creative, without necessarily being new to human knowledge generally. Surprise also exists on a continuum, from mildly unlikely combinations of existing ideas, to highly unexpected developments of existing conceptual frameworks, all the way to “the shock we experience when presented with a new idea that is seemingly not just improbable and/or unexpected, but downright impossible” (Boden 2014, 228). Importantly, the experience of surprise occurs not just when we encounter someone else’s creative idea but also when we encounter our own creative ideas. The classical trope was that creative inspiration was an inhabitation by the Muses, while today “an insight is said to emerge from the unconscious mind, showing up in consciousness as a kind of pleasant surprise (Eureka!)” (Paul and Kaufman 2014, 10). An idea pops into our head as though it were generated by someone else and then handed to us. We may even be startled or experience a jolt when a creative idea occurs to us. In these senses, creativity is experienced as outside our control or as coming from elsewhere, rather than as a product of conscious agency or intention.

Phenomenologically, creativity is experienced as dialogic. This is true in several ways beyond what I have just described (that is, the experience of creative ideas as visitations or as external to our conscious minds). In the first instance, the criteria of novelty and surprise presuppose existing understandings. An idea can only be novel or surprising in conversation with the ideas that already exist. In the same sense, an idea can only be judged useful with respect to existing frameworks of knowledge, understanding, and values. In both cases, however, neither existing knowledge nor novel ideas speak for themselves; it is a human interlocutor who judges which ideas are creative or, as Kant (2001) would put it, “exemplary” (43). Creativity is not a fact that an idea can be discovered to possess but an individual’s assessment of the meaning and value of the idea by way of “conceptual thinking, perception, memory, and reflective self-criticism” (Boden 2004, 2). In this sense, “creativity is not really a property of products or processes at all, but a category of judgment in the minds of observers” (Cropley 2011, 363). It follows that the value of the creative idea may, in the most interesting cases, ultimately restructure one’s prior standards of valuation and judgment. The reflexive and dialogic character of judgment, as Zerilli (2005) has argued, marks it as a distinctively

political (and human) capacity: the ability to make a free and undetermined reflective judgment while also holding open the space to begin again—that is, to revise both our prior judgments and the standards we used to arrive at them.

Moreover, creativity can literally take the form of dialogue; that is, the judgment about whether an idea is creative can come from outside the originator of the idea. This happens all the time: when we judge a student's work, when we go to a colleague for advice on a paper we are writing, when we workshop a research project, when we engage in peer review, or when we read someone else's work to find the pieces which spark our own thinking and ideas. In the last case, we may see something in the work—judge it to be creative—in ways that completely exceed and even elide what the author themselves finds to be valuable about the work. That is, our assessment of creativity does not rely upon the author's intention. Like the ancient practice of *sortes Virgilianae* or medieval bibliomancy, the creative meaning or provocation we find in a text can even have an element of randomness (though of course in such cases we still judge the overall creative fecundity of a given text). In fact, we typically presuppose that there is value in taking a wider perspective or having some critical distance from an idea before we judge its value. At minimum, there is nothing about creative judgment that necessarily demands that the subject who judges must be the one who produced the idea.

This separation within creativity between divergent thinking and judgment opens space to begin to reconsider Wolin's objection to the ideal computer colleague. Imagine this water-cooler conversation: a human colleague asks us to consider three possible paper topics to pursue. They want our judgment of which one, in other words, would be most novel, surprising, and valuable to the field. Is there a meaningful difference if our interlocutor is a machine, offering the same three possible topics for our consideration? I do not mean at this point to ask the question of whether the “divergent thinking” process that a machine and human might engage in is the same—that is, whether a machine thinks, or whether the brain or the mind is, in fact, analogous to the neural networks and algorithmic structures of computers.¹¹ The question here is more straightforward: since creativity is dialogic and often separates the generation of ideas from the process of judgment, is there

11. I do not intend to align myself with thinkers who posit that human individuals—their thoughts, emotions, judgments, sentiments, and values, in sum, their consciousness—are nothing more than, and can be reduced to, their neural networks firing in “an exquisite neuro-computational dance” (Churchland 1995, 225).

something distinctive or problematic about bringing a machine in at the stage of divergent thinking?

The same caveats and cautions apply here, I think, as would apply to human collaboration. Not all suggestions—from either humans or models—are creative; in both cases, we continue to rely on our discerning faculty of judgment. We also want to retain our own capacity for divergent thinking, even if we rely on or draw inspiration from brainstorming or talking with others. But as contemporary theory has de-centered the myth of the thinker as a solitary and self-sufficient monad, it has cleared space to recognize the ways in which divergent thinking is dynamically iterative, responsive to provocation, intersubjective, and generated from randomness—in short, open to the possibility of bringing in computational systems without compromising the distinctively political and human enterprise of political theory. In the sections that remain, I explore what this might look like in practice. First, I discuss—with a critical eye—some ways that political theorists have already drawn on computational systems. Then I build out my own perspective: that we both can and should work to integrate large language models as tools to aid us in the divergent thinking stage of our work.

Texts as Data

Texts form the bedrock of political theory: we interpret, synthesize, contextualize, and draw normative inspiration from them. Despite prevailing disciplinary skepticism, some political theorists have even attempted to compute them—that is, to experiment with computational methods of natural language processing that treat texts as data. Primarily, these methods have been employed by theorists who work in the history of political thought or the *Begriffsgeschichte* traditions. These studies, while covering a range of topics, have shared a common goal: to distill facts from text data. This makes sense because this is how the promise of natural language processing has been framed.¹² These methods, ported into the humanities and social sciences from computational linguistics, have promised validity and efficiency in the process of taking in large text corpora, synthesizing the

12. An important exception comes from literary theory, where Moretti (2013) has advocated for machine-based “distant reading,” a method of engaging with texts that is fundamentally different than reading *simpliciter* and, importantly, normatively preferable if one is truly committed to expanding “the canon.”

regularities and patterns within them, and generating previously unknown facts about them.¹³

Early natural language processing studies in political theory addressed the most basic factual question a theorist might have about texts, which is the question of who wrote what. These studies, going as far back as the 1960s, included efforts to statistically infer authorship of disputed *Federalist Papers* (Mosteller and Wallace 1964) and to attribute anonymous writings to Thomas Hobbes (Reynolds and Saxonhouse 1995).¹⁴ In the *Federalist* study, Mosteller and Wallace considered a collection of papers whose authorship—Madison or Hamilton—was disputed. The problem is a classification problem: given a subset of *Federalist* texts where the authors are known, can the author's pattern of language somehow be determined and then applied in order to classify the disputed papers via inference? While earlier historians had tried to trace similarities in a more *ad hoc* sense—looking for phrases in the disputed papers that were similar to Madison's notes from the Constitutional Convention, for instance—Mosteller and Wallace used “data internal to *The Federalist* but not depending on its intellectual content” (6). More precisely, they found that Hamilton and Madison used words at quite different rates in their writing: not weighty content words like “republican,” “federal,” or “judiciary” but primarily small function words like “by,” “upon,” and “to.” The *Federalist* classification problem became a problem of statistical inference. By modeling the occurrences of a list of these discriminating function words in known texts, Mosteller and Wallace were able to calculate the odds that an unattributed document was written by either Hamilton or Madison. This task is by no means mathematically straightforward, as the nearly three-hundred-page book on the method makes clear. But the overarching principle *is* straightforward: texts can be treated mathematically in order to infer patterns, trends, and facts lying below the surface of the language.

As text-as-data methods have developed greater power, political theorists have sought to computationally infer more complex patterns and facts from texts. Recently published work has sought to explore and quantify topics and themes in texts, as well as map shifts over time in the meanings of concepts (Jockers and Mimno 2013; Kozłowski, Taddy, and Evans 2019; Rodman

13. For general overviews of text-as-data methods, see Grimmer, Roberts, and Stewart (2022); Grimmer and Stewart (2013); Schwartz and Ungar (2015).

14. For a more recent attribution study along these lines, see Neocleous and Loizides (2020).

2020). In their sophisticated statistical analysis, for instance, Blaydes, Grimmer, and McQueen (2018) take an explicitly “empirical approach” to extracting and quantifying themes in a corpus of medieval advice texts for princes and sultans (1151). Each text is divided into short sections that the model mathematically simplifies and clusters with other sections based on linguistic similarities. Though the researchers specify the number of themes, they do not pre-specify the content of these themes; the model discovers the “topics,” which the researchers must afterward interpret by drawing on substantial disciplinary knowledge. In analyzing these topics, trends in thematic content over time and across cultures emerge: they find, for instance, a long and slow decline in the religious themes of the European advice texts, which they historically contextualize. They also use this fact to refigure Machiavelli's secular *Prince* as a culmination rather than a radical break in thinking. Again, the principle of the method is clear: to extract facts about the texts “generally unavailable even to the most discerning of readers” by analyzing them in a different way and “at a larger scale than close reading” (1165).

Despite the contributions of these studies, theory in general has remained uneasy with, and largely uninterested in, text-as-data analyses. The basic paradigm of text-as-data—analyzing texts computationally in order to infer implicit patterns and facts—is at best an awkward fit with a discipline that is not particularly fact-motivated. The problems that technologists have excitedly told humanists that text analysis can help them solve—clarifying matters of attribution and literary dating, filling gaps in damaged texts by means of text predictions, or quantifying networks of thinkers or the meanings of words—are not problems that lie at the heart of what most theorists take themselves to be doing. These types of problems are not wholly tangential to theory as a vocation, but they are distant from the practices of creativity and judgment that constitute the core of the discipline. To the extent that computational text analysis remains fact-motivated, it will remain at the fringes of political theory.

Nor is it merely that text-as-data methods focus on a set of questions that do not deeply interest theorists; the different questions of interest reflect a more fundamental disagreement about method. Theory has an implicit philosophy of language or, more precisely, of reading. Theorists read texts as thickly “culturally and socially situated” and as “reflect[ing] the ideas, values, and beliefs of both the authors and their audiences” in a complex and indeterminate sense; computational approaches struggle to integrate “such subtleties of meaning and interpretation” (Nguyen et al. 2020, 2). Nor is this merely a problem of improving in a technical sense how well

the computational model can integrate such complexities by, for instance, modeling polysemy, syntactic ambiguity, or idiom in a more sophisticated way. Texts are by nature complex and indeterminate objects on which theorists exercise their capacity for judgment, a capacity that defies computational modeling. We do not read to discern a fact of the matter but as a method of constructively uncovering the judgments of others, dialogically, as we form our own. Reading—that is, making ultimately contestable judgments about a text—is not separable from making our own political judgments; the processes cannot be disentangled. This is perhaps why the most theoretically interesting and generative elements of a text are often those that are least clear and that offer no agreement on ground truth.

Finally, there is a wholly practical objection that a theorist might raise. One of the major difficulties with training language models is that one has to train the model to perform the precise task one wants it to do. Not only does this require a lot of data—often including costly manual annotation of training data—but the model has no agility: if you train a model to parse a particular corpus of texts into sixty topics or to distinguish Hamilton's writing from Madison's, that is all the model is able to do. This is not insurmountable—you can train new models or permute a model to a new domain—but it contributes to a wider dilemma preventing the uptake of these methods in political theory: it requires a daunting level of technical skill to select, train, validate, and interpret these models. One of Sheldon Wolin's least contestable worries about quantitative methods training for theory graduate students is that it simply takes up an untenable amount of their time and energy.

Computational text-as-data analysis to this point, then, does not appear to have a great deal to offer theory. However, a major paradigm shift is currently taking place in computational linguistics. This shift, I argue in the next section, will bring these methods much more in sync with theory *qua* theory. Instead of seeing corpora of texts as data that can be studied to yield facts, large language models are now being developed that learn the overall structure of language with the intent to simulate human-like text generation. In essence, the logic of these models is inverted: instead of trying to see patterns in existing texts as a way of saying something about those texts and the world they represent, the model strives to model the full complexity of the language structure from existing texts so that it can dynamically generate more language. They are generative rather than reductive. The model does not introduce assumptions about the truth or meaning of texts, nor does it offer interpretable facts about the texts as an output. These massively large language models—currently, complex transformer models that are trained on

billions of words at a cost of millions of dollars' worth of electricity alone—have a straightforward goal: modeling the structure of natural language such that they can generate new written language in response to prompts that are themselves written in natural language.

Large Language Models as a Humanism

In this section, my goal is not to offer a fully-fledged methodology for how theorists should use large language models. Instead, I want to describe the features and general logic of these models, show what they are and are not good at doing, give some hopefully generative examples of the models at work, and then briefly describe one methodological direction that I believe theorists might fruitfully pursue. My intention in this section is to provide clarity around the structure and possibilities of large language models. I defer various technological, disciplinary, and ethical concerns until the last section.

Although a number of large language models have been developed recently, the most widely discussed, publicly available model is the fourth-generation generative pretrained transformer, shorthanded as GPT-4.¹⁵ This deep neural network was trained on a corpus of more than 500 billion tokens (roughly, words) from a crawl of the internet, including the whole text of Wikipedia, as well as massive collections of text from books. What is even more remarkable than the corpus size, however, is the complexity of the model itself. To get an intuition about the scope and complexity, we can imagine a neural network as a distillation of the training text into an incredibly complex mathematical function that encodes the relationships and structure of the data that trains it. GPT-3 had 175 billion terms in its function, a ratio of roughly three training words to every term, and the model size is only growing with each generation of the model. Given the inevitable repetition in language—that is, the training set assuredly has duplicative and similar language in it—it is easy to see that the model is encoding tremendous linguistic complexity and nuance in the neural network. In essence, the model learns rules and standards of written language inductively through exposure to a vast amount of it. That is, it learns language structure without ever being explicitly trained on rules of grammar, meaning, or style. The model also incidentally encodes a great deal of information that is contained

15. For more details on how to access the model as well as suggested technical literature, see appendix A.

within language; after all, a dictionary is a kind of compressed encyclopedia and language is a rich encoder of logic.

Given its knowledge of natural language, users can interact with GPT using natural language prompts rather than, as with most computer systems, a programming language. It is already clear that to get the most out of the model—that is, to write good prompts—is not always straightforward; it requires the kind of creative facility with language and concepts at which theorists excel. Through these natural language prompts, the model can be given a wide range of language tasks to perform: it can generate stories or poems in various styles, edit or simplify writing, suggest titles, brainstorm new ideas or offer criticisms of ideas, write passages, and generate outlines and content in response to prompts. This is only a tiny fraction of what is possible; if the user can explain the task in natural language and the output is natural language, odds are decent GPT can do it. The model can engage in what is known as “zero-shot learning”: it can deduce what you want it to do from the prompt, without prior instruction or training on that particular task. It can perform even more sophisticated tasks under a “one-shot” or “few-shot” paradigm, where you provide one or several completed examples of the type of answer you want it to give before asking it to do so.

It is important to clarify two points here: first, the model is reproducing patterns in existing natural language, not doing something we might call thinking. These types of models have been called “stochastic parrots,” which accurately captures what they are doing with language (Bender et al. 2021). Producing natural language is not the same type of task as, for instance, computing $2 + 2$, nor is it the same type of task as randomly generating a number between 1 and 1,000. GPT models language probabilistically in a way that threads determinism and randomness. It will not give you an answer to a prompt that is completely random and outside the scope of what you have requested, but if you input a given prompt twice it will not give the same response both times. To the extent that the model can produce language that looks like or mimics reasoning or thinking, it is drawing on a sophisticated model of the structure of language—not on a thinking or reasoning capacity that exists outside of its language modeling.

Second, the empirical knowledge that the model has encoded is incidental to the language modeling, not a primary goal of the model. What these models “know” about the world is limited to what is implicit in the structure of language. As a consequence, they are notoriously bad at facts—or good at imagination, depending on your perspective. The model is often willing to make up answers to absurd prompts like “How many pieces of sound are

there in a typical cumulonimbus cloud?” (apparently, 1,000 pieces).¹⁶ Along these lines, the model is very talented at inventing plausible sounding pieces of research that do not actually exist.¹⁷ This is not to say that large language models cannot be used empirically; there are a growing number of applications that embed the model into other programs in order to pair it with empirical data or web search tools, which I discuss in Appendix A.

Given everything we know about the model so far, it is perhaps unsurprising that GPT performs moderately well when asked to produce divergent outputs to a prompt. As I detail in Appendix B, the model outputs in response to such prompts are not dissimilar to human outputs, though humans outpace the model in terms of producing extreme originality. This suggests a framework for how we might utilize large language models: as tools to augment and spark our own divergent thinking in the production of ideas and written language.

It might help to work through a toy example of what I mean. Here is the abstract of a recent piece of political theory (Turner 2023) published in the *American Political Science Review*:

Walt Whitman’s *Democratic Vistas* (1871) has become a touchstone of democratic theory. Commentators of unusual ideological range uphold the book as politically exemplary. This article demonstrates that recent theoretical celebrations of *Democratic Vistas* are sanitized and incomplete. I expose the antidemocratic side of *Democratic Vistas* by analyzing (1) its philosophy of death and (2) its politics of race. Whitman framed his immortalist response to death within an imperialist historical teleology. That teleology entailed violations of Native sovereignty, the political inequality of Black Americans, and the projection of both Black and Native peoples’ evolutionary extinction. *Democratic Vistas* emerges from this analysis as both necropolitical and white supremacist. If, as Richard Rorty argues, *Vistas* models a salutary form of reformist “national pride,” then it also illustrates the dangerous susceptibility of such pride to moral innocence and self-deceit.

-
16. This is a widely used example of early GPT-3 “nonsense replies” to prompts.
17. In my early testing of GPT-3, I asked the model to suggest literature that I might want to read given the abstract of a paper I was writing. It suggested several pieces including an interesting sounding article by Linda Martín Alcoff called “Political Solidarities,” published in the *Journal of Political Philosophy* in 2019. After searching, it became clear that Alcoff had never written or published such a piece, and that the suggestion was a highly plausible-sounding confabulation (what OpenAI calls a “hallucination”) of the model.

What would be a good title for this article? I prompt the model as follows: “Suggest five creative and eye-catching titles for an academic political theory article in a highly ranked journal. The paper has the following abstract:” and then provide the text of the abstract.¹⁸ The model offers:

1. “Whitman's Democratic Vistas: A Façade of Democracy?”
2. “The Dark Truth Behind Whitman's Democratic Vistas”
3. “The Necropolitics of Democratic Vistas: Unmasking Whitman’s Deadly Ideology”
4. “Whitman's Democratic Vistas: Uncovering Its Antidemocratic Tendencies”
5. “Immortalism, Imperialism and the White Supremacy of Walt Whitman’s Democratic Vistas”

None of these titles are, to my mind, better than the title the author himself came up with—“Whitman’s Undemocratic Vistas: Mortal Anxiety, National Glory, White Supremacy”—which has a compactness of language in the “undemocratic vistas” and a suggestion of tension and ambiguity in the “anxiety.” But I have a fondness for the pun in the model's third suggested title, and, more to the point, the five titles are good enough to spark other thoughts or creative refinements. That is, they serve two purposes: first, they provoke our own divergent thinking, and second, they ask us to exercise our judgment as to which, if any, are novel, surprising, and valuable—that is, creative.

A methodology for using large language models in political theory could utilize the model's ability to produce divergent text in a number of ways. Titles and abstracts are areas where these tools can support the writing and thinking process. The model could be prompted with sections of early drafts or notes about a project to synthesize nascent ideas, suggest objections, develop lines of argument, raise related points, or highlight areas for clarification. Nor does this need to reproduce what a conversation with a human colleague would look like, as the model can output text in any format or style. In experimenting with some of these possibilities, for instance, I discovered that GPT is quite talented at writing short poems like haiku in reaction to prompt text, which I found particularly generative for my thinking. It also introduces moments of low-cost levity and enjoyment into my thinking

18. I thank Chip Turner for his indulgent permission to use his abstract for this example.

process, as when I asked the model to synthesize the introduction of this essay into a short poem in the style of Roald Dahl:

The dream of a perfect colleague,
Accessible and wide-eyed machine,
A computer filled with knowledge,
And quick to intervene.
To provide data, facts, and more,
Synthesize new ideas galore,
Analyze logic with ease,
And suggest further tests to please.
But is this the thing theory needs?
For Wolin warned us,
Judgment and creation are the heart of our acts,
Political theory needs more than brute force and facts.

The pleasure the model's poems give me is no doubt idiosyncratic; the ability of the model to generate valuable divergent content will depend on what the individual thinker finds valuable. Visual learners may benefit, for instance, from the growing integration of images as both inputs and outputs of large language models. Finding value entails a process of experimenting with the model. I think of this as roughly analogous to how we frame our questions about our research differently to different colleagues. We do not approach people as if they are omniscient, but neither do we approach them as if they know nothing; instead, we try to let them into our thinking as best we can by framing our questions to match what we understand about their own thought processes and experiences. We recognize that someone can be a valuable interlocutor even if they occasionally fail to understand what we mean or if they offer help that is too tangential to be helpful.

GPT is not a human colleague; I am not suggesting here that it deserves our grace or patience. It is a tool to augment our creativity as theorists; we are the ones who, as we work to figure out what best supports our creativity, deserve our own grace and patient acceptance. Working with large language models is a reflective and metacognitive process of thinking about your own creative processes, and how you might frame the support you want in order to ask for it. In this sense, I think theorists stand to benefit distinctively from

the process of engaging and experimenting with this tool. I offer a discussion of how to get started with this process of engagement and experimentation in appendix A, where I outline ways to access GPT as well as offer some specific advice on how to formulate prompts.

The more I reflect on the model, the more I am drawn to one possible line of methodological development. These large language models have a well-rounded corpus of texts in their training data, but they were obviously developed without an eye to specialization in the disciplinary language that political theorists use. Fine-tuning the model—that is, collecting a large corpus of theory texts and exposing the model to it such that it learns the specific structure of such language—is likely to make the divergent outputs of the model more precise with respect to the cross-cuttings meanings and subtle uses of discipline-specific language and concepts. For my own work, it would be valuable to me for the model to have a wider range of nuance, for instance, in the way it divergently relates the word “equality” or the word “alienation” to other words. I would value a model that offered me rich and allusive poetic meditations drawing on the linguistic relationships of concepts at the heart of our discipline.

Concluding Concerns

If I have attempted to sketch a clear and positive account of large language models in the prior section, it is not because there are no troubling implications and unanswered questions. To raise these concerns, however, requires a shared understanding of the structure of the models as well as a clear articulation of what we take to be the content and value of political theory and the nature of creativity. Without that clear-eyed grounding, we simply have our heads in the sand, lamenting the development of computational models as “too dreadful;” as Turing (1950) drily put it, that attitude is not “sufficiently substantial to require refutation. Consolation would be more appropriate” (444).

In trying to build a shared preliminary understanding, I have already touched on several concerns in passing: whether large language models are reasoning or thinking, for instance, or under what conditions of production creativity is still considered creative. GPT is a highly sophisticated model of language structure, but I have tried to demonstrate how it is not in any sense a generalized artificial intelligence. Similarly, I have argued that creativity and judgment are distinctive human capacities that cannot be reduced to computation; nevertheless, elements of creativity—namely, divergent

thinking—can be augmented by large language models without losing what makes these capacities distinctive.

Several interrelated concerns remain, however.¹⁹ To begin with the most pressing and perhaps least straightforward, many scholars (Bender et al. 2021; Henderson et al. 2020; Strubell, Ganesh, and McCallum 2019) have documented and described the environmental and financial costs of large models and the way that those costs produce cascading inequalities and unjust distributions of harms. The energy demands to train these models, as I have already noted, are extraordinary and therefore the models are deeply imbricated with questions of global environmental justice. Moreover, the benefits of these models generally accrue to the globally most privileged (Benjamin 2019). While most of our scholarly tools, practices, and institutions are linked to these kinds of harms and inequalities, large language models bring them into stark relief. In the best scenario, discussions of these issues with respect to language models might help us to get clarity on our broader disciplinary obligations, decision-making processes, values, and standards. I do not offer such clarity here, but I strongly concur with the disciplinary need for it.

There is another concern that is related to unjust distributions of harms from these models, and that is the widely acknowledged (Guo and Caliskan 2020) presence of biases and stereotypes: the models “encode stereotypical and derogatory associations along gender, race, ethnicity, and disability status” (Bender et al. 2021, 613). The language on the internet, which is the source of much of the training data, is not only shot through with these stereotypes, but it represents a hegemonic and privileged subset of human language. This is usually—but not universally—understood as a problem of practice, and suggestions have already been made for what mitigation might look like (Jo and Gebru 2020).²⁰ But it raises a much wider issue about the

19. The issues I raise in this section do, of course, relate to broader social, political, and existential concerns about generative artificial intelligence. Political theorists can and should be at the forefront of these conversations, and it is my hope that this paper inspires more theorists to join those already at work on these and related issues (e.g., Panagia (2021); Koopman 2022). The scope of this section, however, is confined to issues that directly confront a theorist choosing whether to use LLMs in their work (while acknowledging that the local, so to speak, is always imbricated with the global).

20. For a critical discussion of the problems of data and data formats, see Koopman (2022).

utility of large language models specifically for empirical research. I have explained the basic operations of the model, but I should emphasize here that GPT is a proprietary as well as technical black box: we can see and judge the outputs, but we do not know how, exactly, the model has arrived at any given output. In the case of divergent outputs, we retain full control over the output through our subjective judgment of its value; we are not in a position of assessing truth claims because we are not trying to produce facts from the data. Validating empirical outputs is much more fraught and unclear, a concern that has been raised repeatedly by thoughtful computational language methodologists in the last decade (Grimmer and Stewart 2013).

What would it mean to address this interrelated issue of algorithmic bias and empirical validation? As Amoores (2020) observes, dominant critical perspectives call for transparency, for “removing the ‘bias’ or ‘value judgments’ of the algorithm, and for regulating harmful and damaging mathematical models” (5). This is a model of ethics that centers openness and accountability and is grounded in the belief that models can be made ethical by being prevented from violating social norms like reproducing derogatory and biased language. Given the opacity, magnitude, and complexity of these new models, I am concerned about the technical feasibility of this approach. More broadly, I am troubled by this general ethical orientation. Drawing on William Connelly, Amoores notes that “actions one might consider harmful are not merely ‘actions by immoral agents who freely transgress the moral law’ but are ‘arbitrary cruelty installed in regular institutional arrangements taken to embody the Law, the Good, and the Normal’” (6). It seems to me that in thinking about the ethics of large language models, we should be wary of ethical paradigms that focus on tinkering with harms at the surface or that seek to displace us as agents who make judgments.

This wider ethical view also shapes my response to concerns about how large language models might (1) undercut the profession of writing political theory and, relatedly, (2) undermine the practices of writing by which we evaluate our students. These concerns are usually framed as total and direct: that is, what if the model can write an essay comparing Locke and Hobbes that is indistinguishable from an “A” student's effort, or what if the model can be trained to write political theory papers good enough to pass peer review. Whether or not this fear is well-grounded at the moment is almost irrelevant because the technology is developing with astounding rapidity. In both instances, we *will* eventually be forced to grapple with deeper questions about what writing is a cipher for and what writing means and does for us. After reflecting on it, I am not troubled by the possibility that a model might

write a professional political theory paper that I would read and judge to be novel, surprising, and valuable; I find value in many unexpected places, and I am grateful for it whenever I find it. To paraphrase Emerson, we lie in the lap of an immense intelligence; we remain at the center, judging what we find and bringing along the things that help us think about how to live together. I value my own writing as a means of thinking, and the practice has intrinsic value to me independent of the product.

The case of the student paper changes character if we think about it in these terms. In the most charitable characterization of pedagogical motives, I might argue that I want students to do their own writing because writing is instrumental to the process of thinking; I know they are thinking because I see it in their writing. This is a murky claim. GPT, as we have seen, demonstrates that language can be modeled computationally, without implying thinking or reasoning. Similarly, student facility with language often reflects other things besides thinking: a capacity for mimicry, working in a native language, tendencies toward behavioral compliance and rule following, vulnerability to negative sanctions, background educational privilege, a sense of the right to assert one's own position, time and space in which to write, neurotypicality, and physical health and social wellbeing. GPT requires us to confront a fact that was already true long before large language models: our pedagogy has been insufficiently attentive to the link between student thinking/learning and student assessment. As with the issues of global inequality and environmental justice, large language models highlight real and terribly urgent problems in our discipline that already exist rather than create dilemmas wholly new.

Whenever we borrow or appropriate methods from elsewhere for our own uses, they retain traces of their origins. I do not deny this, nor do I seek to claim that my re-imagination of large language models for political theory is definitively neutral—that is, that it carries no potential unforeseeable effects on how political theorists are trained, do their thinking, and assess one another's work.²¹

21. One question that looms concerns authorship—that is, of authorial attribution, citation, and intellectual property. Amoores's *Cloud Ethics* offers a provocative chapter on these themes for those who might like to think about these questions in more depth. I suggest, to begin, that authors disclose the ways they use LLMs, as I do in this paper's acknowledgments. Another question—again, I only gesture to this without speculation—regards the unforeseeable ripple effects that LLMs might have on the professional institutions and resources available to, but not coeval with, the practice of political theory. Will they be widened and improved? Or will our jobs, our departments, our journals, and our students no longer exist?

At the same time, I would stress that the technological developments are already happening and already having effects. My goal in this paper is to grapple with the technology that currently exists and to see how we might use it. By clearly articulating our theoretical enterprise and demanding to be met on our own terms, we are better positioned to both defend what we value and see the value in new possibilities. More, even: we can express what we value with a clarity that facilitates our engagement with the difficult problems of pedagogy and justice that were already looming over our field.

Appendix A

Getting Started with GPT

This overview is meant to smooth any initial barriers to entry for those who have not yet spent time experimenting with GPT in any form; this technology is evolving rapidly, and these details will certainly change.

Accessing the Model. There are two main ways to directly access the various GPT models. The first and most straightforward is a free and publicly available chatbot released in late 2022 by OpenAI, the developers of the model, at <https://openai.com/blog/chatgpt/>. ChatGPT is built on a GPT model that has been further fine-tuned to optimize it for producing texts structured like a conversation. ChatGPT *appears* to have memory because it uses the entire conversation up to that point, including the most recent question asked, as the prompt that it is answering. A consequence of this is that to start a chat on a new topic, you will want to begin a new chat rather than change topics within the existing chat window. OpenAI has placed *post hoc* social safety limits on ChatGPT, so it is a commonplace experience to ask the model to do something and to be told either it is not allowed to or capable of doing it. Even if it ultimately completes a prompt, the model reminds users frequently what it is and what it can do. For instance, I asked the model to write a speech in the style of Thucydides that defends the sovereignty of a hypothetical nation against the incursions of neighboring great powers. While it did produce the speech, it began with a preamble reminding me that it was a language model and not capable of reasoning or arguing. Because it frequently flags its capabilities and is able to conduct a chat dialogue in a way that users will find familiar, ChatGPT is an excellent interface to use to begin exploring the model.

The other way to access GPT is to create a developer account with OpenAI at <https://platform.openai.com/overview>. You can start by accessing the

models with the intuitive Playground page on the OpenAI website, which is a large text input-output box with a selection of controls and settings you can adjust on the righthand side. You can also access the model through the API (using programming language for model calls and prompts). The website offers a Quickstart Tutorial as well as a very illuminating Examples page. With either the API or the Playground, the user pays a very small amount of money to send their prompts to the model and receive output text. Why would a user want to pay money to access the model this way rather than use ChatGPT? You may not want the model to be optimized for conversation, you may want to adjust the hyperparameters of the model, or you may want to fine-tune a version of the model on a specific text collection. None of those things are currently possible within ChatGPT.

Writing Good Prompts. Writing good prompts is a skill, but political theorists are well-suited to the task since we are trained to produce and interpret language. Start with a clear and concise prompt to set the direction for the outputs, providing enough context to help the model generate relevant results. OpenAI has recently added “Custom Instructions” input boxes to the ChatGPT interface where you can specify context for your prompt and the desired form of the output. It can be helpful to tell the model what kind of role or character it is playing (i.e., “you are a peer reviewer for an academic political theory journal,” “you are Abraham Lincoln at Gettysburg”) or to describe the setting or style of the output (“for a brainstorming session,” “in the style of the attendees of the American Constitutional Convention in 1787”). Another way to give context is to provide an example of a completed prompt (i.e., “X is the title for a paper with Y abstract. Using Z abstract, write five possible titles for Q paper”). GPT limits the total combined length of the prompt and output, but these limits are constantly growing. It is now possible using GPT-4 to feed in most of a draft of an article as part of a prompt.

New users should try varying the language in the prompts so that they differ in small ways to see how the outputs change, and they should also respond directly to a model output with further directions. For instance, they can tell the model things like “try again,” “be more creative,” “rewrite what you just wrote in more sophisticated sounding language/as rap lyrics/as an outline,” “assess how interesting what you just wrote is,” “expand on that,” “keep going,” and so on. This can be counterintuitive at first, because we would not use these kinds of directives in a dialogue with human beings.

Fine-tuning, Hyperparameters, & GPT Pairings. As mentioned previously, creating a developer account allows you fine control over the settings of the model. You can control the randomness of the outputs (the “temperature” control lever), set the length of the outputs, choose the version of GPT you want to access (there are many submodels), establish penalties for language or topics that have already been used in the prompt or output in order to increase novelty later in the generated text, and have the model generate multiple outputs and then select the best one to show you. The ideal hyperparameter settings will depend on what you’re looking to prioritize in your model outputs. I find, for instance, that in generating divergent lists, the model tends to repeat itself and use similar structures in the language unless the frequency and presence penalties are set very high.

Fine-tuning a model on a corpus of texts might also make sense for certain language generation prompts. For instance, you might want to train the model on a collection of your own working papers that is substantially longer than the allowed prompt length. Another use case would be training to a specific task—like writing high-quality scholarly abstracts from introductions—where you want to provide the model with hundreds of successful examples. This is done using relatively simple computer programming via the command line, and the documentation section of the OpenAI website provides clear instructions for how to format and input the training data.

Finally, new pairings are emerging between GPT and various applications, data sources (including images), and search engines. For instance, a publicly available scholarly literature review tool, Elicit, uses GPT to provide on-the-fly written summaries of individual pieces of literature or of the literature on a given topic as a whole. You can type in a research question in natural language, and it will look for and summarize papers of interest. GPT plugins are a new line of development that are likely to pair large language models in the near future with a wide range of existing scholarly database, archival, and computational tools.

Annotated Technical Bibliography. Extremely readable and entertaining introduction to artificial neural networks:

Shane, Janelle. 2019. *You Look Like a Thing and I Love You*. New York: Little. Good technical introduction to artificial neural networks:

Buduma, Nikhil and Nicholas Locascio. 2017. *Fundamentals of Deep Learning: Designing Next-generation Machine Intelligence Algorithms*. Sebastopol, CA: O’Reilly.

General introduction to transformer-based text models, including social science applications:

Wankmüller, Sandra. 2022. "Introduction to Neural Transfer Learning With Transformers for Social Science Text Analysis." *Sociological Methods & Research*. Forthcoming in print; available online at <https://arxiv.org/abs/2102.02111>.

Original technical paper describing the model architecture of transformers:

Aswani et al. 2017. "Attention is All You Need." *Conference on Neural Information Processing Systems*. <https://arxiv.org/pdf/1706.03762.pdf>

Discussion of the training, data, performance information, and technical details of GPT-3:

Brown et al. 2020. "Language Models are Few Shot Learners." <https://arxiv.org/abs/2005.14165>

Updated description of the fourth-generation model's capabilities, including the ability of the model to take image inputs as well as texts:

OpenAI. 2023. "GPT-4 Technical Report." <https://cdn.openai.com/papers/gpt-4.pdf>

Appendix B

Divergent Thinking Tests

"Similarities" Test Results. Assessments of human divergent thinking look at the responses that someone gives to various standard prompts and judges those outputs based on their originality (Runco et al. 2016). Similarly, we can use these natural language prompts to ask GPT to produce outputs that we can also judge for originality. One popular prompt is a "similarities" test: "list as many similarities as you can between apples and pears. The more similarities you think of, the better." When I take this assessment, I am able to produce seventeen similarities over a roughly five-minute period. I draw on my own experiences with apples and pears, particularly recent experiences (i.e., with Harry and David holiday boxes). Some ideas flowed rapidly one after another; at other moments, I had to stare off into space for a minute to come

up with the next one. In about ten seconds, GPT is able to produce fifty-nine similarities before repeating itself, though some of the things on its list are closer in meaning than I would have allowed on my own list. The first seventeen items on the model's list are roughly comparable to my own outputs in terms of originality, and some of the later similarities the model produces are highly divergent from my own list (“they can both be grafted onto other trees to create new varieties”; “they can both be used as a natural remedy for minor ailments such as diarrhea or constipation”). The *process* of producing each list is clearly distinctive: I draw heavily on experience (“Both can fit in your hand”) while the model draws on language it has seen about apples and pears (many recipes, it appears) to produce new language about apples and pears. The outputs, however, are roughly equivalent.

Author	Model
Both are edible.	Both apples and pears are types of fruit.
Both fit in your hand.	They are both round or pear-shaped.
Both can be cooked.	They both have a skin that can be peeled or eaten.
Both can be eaten raw.	They both have a soft, pulpy inside with seeds in the middle.
Both can be made into cider.	They both grow on trees.
Both can be made into sauce.	They are both commonly used in cooking and baking.
Both grow on trees.	They can both be eaten fresh or used in preserves form (such as in jams or jellies).
Both are cultivated by humans for food.	They are both a good source of fiber.
Both contain small seeds inside.	They are both available in a range of colors, including green, red, yellow, and brown.
Both have thin skins.	They both have a sweet taste when ripe.
Both have stems.	They are both rich in antioxidants.
Both are sweet.	They can both be juiced.
Both can rot.	They can both be dried and eaten as a snack.
Both can be purchased at a store.	They can both be canned or frozen for later use.
Both splat if dropped from a great height.	They are both high in natural sugars, which give them their sweet taste.
Both sold by Harry and David.	They are both low in calories and fat.
Both come in many subvarieties.	They can both be used to make pies, tarts, and other baked goods.

“Uses” Test Results. Another common test of divergent thinking is the “uses” test. The verbatim prompt I used for this test was: “People typically use everyday items for specific purposes. Often there are alternative uses for the same objects. For example, a newspaper could be used as a hat or a blanket, and many other things. List as many alternative uses of a spoon as you can. The more uses you think of, the better.” In this case, I asked a political theory colleague (under the promise of anonymity) to take the test, and I also gave the prompt to GPT-3. My colleague worked for about five minutes to create their list, while I allowed the model to generate output until it repeated itself. I then evaluated both lists using a subjective assessment of originality. I present the lists of uses I judged to be highly original below. My colleague clearly surpassed the model in terms of the quantity of highly original uses, though the model has some; the cases of overlap are notable as well.

Colleague	Model
Cook drugs	Using the handle of the spoon to roll dough or pasta into long, thin strands
Make music (drum)	Using the handle of the spoon as a makeshift drumstick to play percussion instruments
Reflect light in someone's eye	Using the spoon as a makeshift chisel or pry bar to remove nails or other small objects
Play fetch with dog	Using the handle of the spoon to apply pressure to a button or switch in hard-to-reach places
Throw at a thief to distract them	Using the spoon as a makeshift magnifying glass by holding it at a distance from the eye and using the curved surface to focus sunlight
Steal from a neighbor to gaslight them	Using the spoon to apply makeup, such as eye shadow or concealer
Use as a pointer in a lecture	Using the spoon as a makeshift spoon rest while cooking on the stove.
Stick on your nose to comfort a crying child	Using the handle of the spoon to apply pressure to the back of a spoon or fork to remove a stuck object, such as food from between teeth

(continued)

(continued)

Colleague	Model
Throw really far as a track and field event	
Shovel out a tiny driveway	
Use as an alien ship in a play movie with model trains and figurines and stuff	
Dig carefully with [sic] excavating fossils	
Use as a secret symbol for a conspiracy	
Take on <i>Antiques Road Show</i>	
Use to type on a keyboard without touching keys	
Dip in ink and use as a pen	
Measure someone's height	
Drop down a well to measure its depth	

Acknowledgments

The paper was greatly improved by comments and suggestions from the editors and reviewers at *Political Theory*, as well as from audience members at the Western Political Science Association conference and the Politics and Computational Social Science conference. Tom Arnold-Forster, Anna Daily, Lisa Gilson, E. Stephen Kehlenbach, Greg Koutnik, Alison McQueen, Christopher Rytting, Noah Stengl, Liz Taylor, Seth Trenchard, Jack Turner, and John Wilkerson also gave helpful suggestions and encouragement. I am especially grateful for the close attention Phil Yaure gave to the manuscript in its early stages and for the rare confluence of theorists and methodologists who came together to workshop a later draft at the University of Wisconsin – Madison. AI Disclosure: GPT-3 was used to produce examples in this work, as noted, as well as to brainstorm ideas, poems, and motivational speeches throughout the writing process. No language presented as my own in the final version was generated by AI.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Emma Rodman  <https://orcid.org/0000-0002-9506-9531>

Bibliography

- Amoore, Louise. 2020. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Raleigh, NC: Duke University Press.
- Arendt, Hannah. 1982. *Lectures on Kant's Political Philosophy*. Ed. Ronald Beiner. Chicago, IL: University of Chicago Press.
- Bagg, Samuel. 2018. "Can Deliberation Neutralise Power?" *European Journal of Political Theory* 17 (3): 257–79.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. 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, Virtual Event, Canada, 610–23. New York, NY: Association for Computing Machinery.
- Benjamin, Ruha. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity Press.
- Berlin, Isaiah. 2014. "Does Political Theory Still Exist?" In *Concepts and Categories*. 2nd ed., edited by Henry Hardy. Princeton University Press.
- Blau, Adrian. 2015. "History of Political Thought as Detective-Work." *History of European Ideas* 41 (8): 1178–94.
- Blaydes, Lisa, Justin Grimmer, and Alison McQueen. 2018. "Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds." *Journal of Politics* 80: 1150–67.
- Boden, Margaret. 2004. *The Creative Mind: Myths and Mechanisms*. 2nd ed. London: Routledge.
- . 2014. "Creativity and Artificial Intelligence: A Contradiction in Terms?" In *The Philosophy of Creativity: New Essays*, edited by Elliot Samuel Paul and Scott Barry Kaufman, 224–44. New York, NY: Oxford University Press.
- Brown, Wendy. 2002. "At The Edge." *Political Theory* 30 (4): 556–76.
- Churchland, Paul. 1995. *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain*. Cambridge, MA: MIT Press.
- Cropley, Arthur. 2011. "Definitions of Creativity." In *The Encyclopedia of Creativity*, edited by Runco, Mark and Steven Pritzker, 358–68. Amsterdam: Elsevier Science & Technology.
- Dienstag, Joshua Foa. 2016. "On Political Theory, the Humanities, and the Social Sciences." *Perspectives on Politics* 14 (4): 1083–89.
- Dryzek, John, Bonnie Honig, and Anne Phillips. 2008. "Introduction." In *Oxford Handbook of Political Theory*, 3–42. New York: Oxford University Press.
- Fraser, Nancy. 2009. *Scales of Justice: Reimagining Political Space in a Globalizing World*. New York, NY: Columbia University Press.

- Fuchs, Christian. 2017. "From Digital Positivism and Administrative Big Data Analytics Towards Critical Digital and Social Media Research." *European Journal of Communication* 32 (1): 37–49.
- Gardiner, Michael. 2022. "Automatic for the People? Cybernetics and Left-Accelerationism." *Constellations* 29: 131–45.
- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton, NJ: Princeton University Press.
- Grimmer, Justin, and Brandon Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297.
- Gunnell, John G. 1988. "American Political Science, Liberalism, and the Invention of Political Theory." *The American Political Science Review* 82 (1): 71–87.
- Guo, Wei, and Aylin Caliskan. 2020. "Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases." arXiv preprint. arXiv:2006.03955.
- Henderson, Peter, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. "Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning." *Journal of Machine Learning Research* 21 (248): 1–43.
- Jo, Eun Seo, and Timnit Gebru. 2020. "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 306–16. New York, NY: Association for Computing Machinery.
- Jockers, Matthew, and David Mimno. 2013. "Significant Themes in 19th-century Literature." *Poetics* 41 (6): 750–69.
- Kant, Immanuel. 2001. *Critique of the Power of Judgment*. Eds. Eric Matthew and Paul Guyer. New York, NY: Cambridge University Press.
- Kaufman-Osborn, Timothy. 2010. "Political Theory as Profession and as Subfield?" *Political Research Quarterly* 63: 655–73.
- Kehlenbach, Stefan. 2022. "The Subatomic Person: A New Ontology of Big Data." *Theory & Event* 25 (4): 851–72.
- Koopman, Colin. 2022. "The Political Theory of Data: Institutions, Algorithms, & Formats in Racial Redlining." *Political Theory* 50 (2): 337–61.
- Kozłowski, Austin, Matt Taddy, and James Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings." *American Sociological Review* 84 (5): 905–49.
- March, Andrew. 2009. "What Is Comparative Political Theory?" *The Review of Politics* 71 (4): 531–65.
- McWilliams, Susan. 2015. "Teaching Political Theory as a Vocation." *The Good Society* 24 (2): 191–97.
- Moretti, Franco. 2013. *Distant Reading*. New York, NY: Verso Books.

- Mosteller, Frederick, and David Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. London: Addison-Wesley.
- Nedelsky, Jennifer, and Ronald Beiner (eds.). 2001. *Judgment, Imagination, and Politics: Themes from Kant and Arendt*. New York, NY: Rowman and Littlefield Publishers.
- Noecleous, Andreas, and Antis Loizides. 2020. "Machine Learning and Feature Selection for Authorship Attribution: The Case of Mill, Taylor Mill and Taylor." *IEEE Access* 9: 7143–51.
- Nguyen, Dong, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. 2020. "How We Do Things with Words: Analyzing Text as Social and Cultural Data." *Frontiers in Artificial Intelligence* 3 (62): 1–14.
- Panagia, Davide. 2021. "On the Possibilities of a Political Theory of Algorithms." *Political Theory*. 49 (1): 109–133.
- Paul, Elliot Samuel, and Scott Barry Kaufman. 2014. "Introduction." In *The Philosophy of Creativity: New Essays*, edited by Paul, Elliot Samuel and Scott Barry Kaufman. New York, NY: Oxford University Press.
- Rehfeld, Andrew. 2010. "Offensive Political Theory." *Perspectives on Politics* 8: 465–86.
- Reynolds, Noel, and Arlene Saxonhouse. 1995. *Three Discourses: A Critical Modern Edition of Newly Identified Work of the Young Hobbes*. Chicago, IL: University of Chicago Press.
- Rodman, Emma. 2020. "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors." *Political Analysis* 28 (1): 87–111.
- Runco, Mark. 2014. *Creativity: Theories and Themes: Research, Development, and Practice*. 2nd ed. Cambridge, MA: Academic Press.
- Runco, Mark A., Ahmed M. Abdulla, Sue Hyeon Paek, Fatima A. Al-Jasim, and Hanadi N. Alsuwaidi. 2016. "Which Test of Divergent Thinking is Best?" *Creativity: Theories, Research, Applications* 3 (1): 4–18.
- Schwartz, Andrew, and Lyle Ungar. 2015. "Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods." *The Annals of the American Academy of Political and Social Science* 659: 78–94.
- Skees, Murray. 2022. "A New Traditional Theory: Fetishizing Big Data Analytics." *Constellations* 29: 146–60.
- Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. "Energy and Policy Considerations for Deep Learning in NLP." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–50. Florence, Italy: Association for Computational Linguistics.
- Turing, Alan. 1950. "Computing Machinery and Intelligence." *Mind* 49: 433–60.
- Turner, Jack. 2023. "Whitman's Undemocratic Vistas: Mortal Anxiety, National Glory, White Supremacy." *American Political Science Review* 117 (2): 705–18.
- Whelan, Frederick. 1983. "Democratic Theory and the Boundary Problem." In *Nomos XXV: Liberal Democracy*, edited by J. Roland Pennock and John W. Chapman, 13–47. New York, NY: New York University Press.

- Wolin, Sheldon. 1969. "Political Theory as a Vocation." *American Political Science Review* 63 (4): 1062–1082.
- Zerilli, Linda. 2005. "'We Feel Our Freedom': Imagination and Judgment in the Thought of Hannah Arendt." *Political Theory* 33 (2): 158–188.
- . 2016. *A Democratic Theory of Judgment*. Chicago, IL: University of Chicago Press.

Author Biography

Emma Rodman is an assistant professor of political science at the University of Massachusetts Lowell. She is a scholar of American political thought with particular interests in the meanings of political concepts, thinkers in the tradition of Black political thought, and computational linguistics. Her research has appeared in *American Political Thought*, *Political Analysis*, and *Law, Culture, and the Humanities*.