# Borges and AI

Rita Raley
*University of California, Santa Barbara*

Russell Samolsky
*University of California, Santa Barbara*



**Keywords** allegory, authorship, ChatGPT, generative AI, large language models

Although at first glance our title, "Borges and AI," might seem to promise or betoken an analysis of the relationship of the work of Jorge Luis Borges to artificial intelligence, what lies behind it is more specifically a play on his fiction, "Borges and I," which we intend to serve as a precursory lead-in and accompaniment to our own thoughts on artificial intelligence (AI) and authorship (Borges [1957] 1998: 324).[1] What Borges's parable gives voice to is his struggle against being eclipsed by his fictions; thinking with this parable will thus help us begin to work through our own confrontation with the "AI author." It is notable that Borges does not articulate one reductive response, and neither will we. What he does instead is unfold the affective and analytical dimensions of being subject to fictional capture. Our essayistic counterpart takes its cue from Borges's parable, but we shall not simply mirror it; rather, we present our own reflections on what might turn out to be a more encompassing eclipse.

While "Borges and I" might be read as a pretext in the double sense to such poststructuralist formulations as Michel Foucault's "author function"

1. Our analysis is based on all authorized English translations as well as the original Spanish text, "Borges y yo."

*Poetics Today* 45:2 (June 2024) DOI 10.1215/03335372-11092911

and Roland Barthes's "death of the author," we propose to extend the parable's critical afterlife by considering it as anticipatory of the condition of authorship after the emergence of large language models (LLMs). There are analogical relations between Foucault's and Barthes's arguments and generative AI, as in the contemporary reassertion of authorship as a property regime and in LLM output as a series of combinatorial tokens. However, our move here is to go beyond an analysis of the Borges story as anticipating poststructuralist theories of writing. Indeed, we propose to go still further by reading this parable not only as anticipatory but also as literalizing the paradoxical position of authors in the age of generative AI.

Readers will recall that the theme of Borges's parable is already implicit in its title, with its split between author and what has been authored, which prefigures the narrator Borges's rather melancholy musing on the way in which he has been overtaken by, indeed is in the process of being fully given over to, the eminent public author Borges. Reduced to an ever-more diminished and receding "I," the narrator remarks, "I walk through Buenos Aires and I pause—mechanically now, perhaps—to gaze at the arch of an entryway and its inner door; news of Borges reaches me by mail, or I see his name on a list of professors or in some biographical dictionary" (Borges [1957] 1998: 324). Bereft of the enchantment of Buenos Aires that has since been transposed to the world of his fictions, he now looks "mechanically" at the streets that once filled him with inspiration. When read in the context of LLMs, *mechanically* takes on an additional meaning, not only because it suggests automation but also because generative AI is regarded as threatening to "mechanically" supersede our powers of inspiration and creative enchantment. To put this differently, the parable, which also alludes to Borges's fictional "games with time and infinity" (324), becomes an anticipatory allegory in its foretelling a future of fictional capture and in its evoking the mechanical and mathematical processes of probabilistic calculation.

There is yet a further, though allied, aspect of fictional capture: not only is the narrator Borges stripped of inspiration, but his very life is also given over to Borges's authorship, to the making of his fictions. He records, "I live, I allow myself to live, so that Borges can spin out his literature, and that literature is my justification" (324). Yet a paradox ensues, for even as he gives his life over to Borges, the author, he is also given over to something beyond this proper name. The value of the pages Borges has written, the narrator recognizes, "no longer belongs to any individual, not even to that other man, but rather to language itself, or to tradition" (324). Here again the parable anticipates the era of big data and machine learning in that the theme of works passing beyond their authors into the realm of tradition is manifest in the corpora on which large language models have

been trained and that inform their "writing." The comparison, however, is not complete; for while Borges the person is destined to die, some part of him is destined to survive and to pass on in the corpus of "Borges." But it is not a perfect imparting, for this Borges is nothing if not a fictioneer: "I have been turning everything over to him, though I know the perverse way he has of distorting and magnifying everything" (324). What is true for Borges the fictioneer is all the more true for generative AI. The language of "deep fake," after all, serves as a reminder that synthetic images and text are in a general sense predicated on a falsification and magnification of what was once thought as proper to, and the property of, the singular, human author.

Although destined to survive only in his infamous double, Borges does not simply surrender to this giving over, or to being given over. He tries by act of imagination to escape from that which he has created and which is in the process of eclipsing him: "Years ago I tried to free myself from him . . . [but] everything winds up being lost to me, and everything falls into oblivion, or into the hands of the other man" (324). In a sense Borges, the "I," is caught up in a version, or perhaps inversion, of Zeno's paradox (on which he has famously written) because every attempt at fictive escape simply winds up on the author Borges's shore. Even more so with the apparatus of large language models: the entirety of print culture is now subject to linguistic capture for ever-expanding training datasets (to say nothing of voice logging and the use of chat applications).

When first drafted, "Borges and I," in a sly reflexive statement of its own theme, ended with the line, "I do not know which of us has written this page," but Borges soon thereafter revised it to the continuous present, "I am not sure which of us it is that's writing this page" (324).[2] If this final line enacts the process of capture, it also prefigures our play, "Borges and AI." Just as Borges the narrator has been given over to the authorial "Borges," so too does this impersonal figure now stand as a metonym for the appropriation of literary "tradition" in the development of corpora used to train the generative AIs that have in turn introduced a discomposing uncertainty as to who or what is writing the pages that we read.[3]

2. As Donald A. Yates (1973: 318) details, Borges, who at that point in his career was dictating his work to his mother because of his deteriorating sight, made only three changes to the manuscript, the most significant of which, and not incidentally the only one in his own hand, was to revise the tense of the final line from "ha escrito" (has written) to "escribe" (is writing).

3. If the training data for GPT-1 was not "literature" in its institutional form but rather unpublished books—the BookCorpus dataset consists of about eleven thousand books scraped from the internet—later iterations in the GPT series have been trained on substantially larger corpora, including Project Gutenberg. For a data archaeological analysis of the in-copyright books used to train more recent models, see Chang et al. 2023.

It is felicitous for us that Borges recognizes his double as both the renowned author of magical fictions and as a professor, for it is not only creative writers but also equally, if not more predominantly, critical writers that generative AI threatens to eclipse. Lest it be thought that, like the contriving "Borges" of the parable, we exaggerate or are hyperbolic in our reading of "Borges and I" as prefiguring AI writing techniques and the problems of authorship they introduce, we make two observations: not only can different language models immediately generate exegeses of this very parable, but the exercise of prompting ChatGPT with our title and abstract results in the iteration of some of the very themes that we have explored in our comparative analysis. (Although to be clear LLMs have played no part in the genesis or composition of this piece.)

Unlike the resigned, if somewhat melancholy, "I" of Borges's parable, who has over time grown accustomed to slowly being usurped by what he has created, we find ourselves as academic writers suddenly confronted by the knowledge that what we have already published has been transmuted into training data—and so too the writing of these very pages as they unfold in different online applications has already been appropriated, and will continue to be. More unsettling, indeed literally discomposing, is the knowledge that what we might come to write, the projects we envision, our future writing, threatens to be usurped because it is anticipatable and hence replicable, by the vast domain of natural language processing. Comparative projects, formal analyses, stylistic readings—indeed, any work that relies on the detection of linguistic patterns and anomalies—all of this can and will be done mechanically, at speed and scale.

To this new technological situation, our response is necessarily both affective and analytical. Both registers are required (though sometimes imbricated) because we find ourselves facing what feels like an epochal moment in our writing lives. Some years of thinking about the techniques and tools of AI—and here we note our initial collaborative foray into the field was "Inter Alia: Aliens and AI," an essay on AI and alien communication (Raley and Samolsky 2019)—should condition the surprise, but it is nonetheless difficult to quiet the initial shock when a new model or application programming interface outperforms us in a domain we regard as our own.[4]

4. Katherine Elkins's wonderful essay in this issue also finds in the metafiction of Borges (and Italo Calvino) anticipatory analogs of emergent AI storytellers. There are a number of sympathetic resonances between us that we could readily amplify if we had the space. But to remark on one: although perhaps less disconcerted than we are by the potential of LLMs for usurpation, Elkins remarks, "being bested by a machine, even if it's (for now) in genres different from my own, is dispiriting."

We are caught, then, in a split between grasping the fact that the scholarly practices and techniques to which we have devoted a fair portion of our lives no longer belong wholly to us, and not at all being ready to accept the probability of our impending diminution—or, in worst-case scenario, obsolescence. But this oscillation between surrender and resistance to the seemingly all-encompassing AI machine gives way to a more considered, though still perplexed, sense of being dislocated by the generative models taking over the process of composition, which in turn gives way to our belief that they cannot really do what we human authors can do—that the output proffered is merely artifice.

No doubt a comparable emotional defensiveness has been felt by many in the humanities. In response, a set of analytical arguments has begun to structure the field of critical AI studies as well as the popular discourse that erupted in the wake of OpenAI's partial release of GPT-2 in early 2019. One line of thinking, to which we briefly gestured, proceeds from a categorical distinction between human and machine and finds language models to be mechanical instruments, or "stochastic parrots" (Bender et al. 2021), with admittedly powerful capacities of probabilistic calculation and rote emulation (a kind of maximalist autocomplete), but lacking all the attributes and talents that have been historically understood to be properly ours. A related, but differently cast argument, follows in the vein of the Platonic idealizing of speech over writing, mourning all that will be lost as one technological process of communication gives way to another. In contrast, optimistic arguments have also been made for the value of shared authorship, for embracing entangled processes of composition and the brave new world, or worlds, to which they may lead. In terms of the politics of composition, the more cautionary perspectives focus on the harms of generative AI, from the resource costs of training large models to biased output and hate speech. As redress, arguments focus on the need for smaller models and better datasets, if not a data commons, as well as a more robust regulatory system for content, platforms, and the industry alike.

For the moment, at least, it seems the array of possible positions one can take in relation to ChatGPT, as proxy for a technological revolution in language processing, has been mapped out. In response to the *Poetics Today* call for our reflections, we will thus try not to supplement this set of positions but rather to work within it to explore what is not yet exactly, in the parable's phrase, a "hostile relationship" between human and machine writers, but is not exactly an amicable one either (Borges [1957] 1998: 324). The premise of the journal's call seems to us exactly right: absent an informed social and institutional consensus about the applications of LLMs, per-

haps the best we can do is to work from where we are, using our own situations, practices, and experiences to inform our approach to authorship and teaching in the age of generative AI.

One of the primary questions, for the journal as well as the educational professions more generally, concerns the classroom. Indeed the student essay—with its templated structure, mode, and style—is particularly available for replication. Like recipes and genre fiction, there is an identifiable formula, as every set of papers to grade can often confirm. Particularly following the expansion of the market for writing assistants, whether software or "shadow scholars," it had already seemed, at least to us, as if undergraduate essays had started to more competently emulate academic prose. Now LLMs, in their actualizing of the question of who or what has written the pages before us, has introduced another dimension to this practice of assisted writing.

If our future critical projects are imperiled, then student essays may already be outmoded, at least insofar as generative AI has been so widely implemented and integrated in everyday writing platforms that it is now impractical to try to enforce a strict prohibition. Students have entered into a new techno-linguistic order; and student writing will have to be understood in terms of machinic entanglement, whether intentional or not. (In this respect, to further our analogy with the Borges parable, each functions as an "I" in relation with the AI that can contrive and "spin out" the "sound pages" that pass as valid [324].) Nonetheless, we as teachers are not ready to give over the entire practice of textual analysis, and the skills of mind this entails, to LLMs. One contemporary, perhaps only temporary, solution for some (presuming interest and capacity) is instead to find compelling ways for students to use these tools: for example, to use the genre of the model prompt to teach the history of artistic styles and movements, or to ask students to develop criteria for evaluating model output. Granted, pedagogy is always evolving, but what might be lost for students in such a methodological revision, one that seems to implicitly acquiesce in and concede that part of the work of textual generation and discovery is more efficiently executed by language models?[5] It is not only skills of reading and writing that are at stake, however, but also the way in which reading and writing compose the self itself.

Generalized, Borges's parable anticipates our disquiet as we become aware of ourselves, and our profession, in the process of being surpassed by the

5. Answers to this question will have to take account of the potential for cultural devaluation of traditional techniques such as exegesis, analogical thinking, and comparative study, among others.

impersonal force of predictive models, even as we hold to the belief that something will remain to us. The narrator invokes Spinoza's claim: "All things wish to go on being what they are—stone wishes eternally to be stone, and tiger, to be tiger" (324). Thus the register of pathos in his remark, "I shall endure in Borges, not in myself . . . but I recognize myself less in his books than in many others,' or in the tedious strumming of a guitar" (324). For Borges there were still other books in which he found himself reposed, or even ironically in the "laborious," or mechanical, strumming of a guitar. If the narrator found himself absorbed by Borges, there was still something left to him, left of him, if only in the hands of others. But the regime of AI is more comprehensive, more consuming, than the apparatus of authorship that subsumed the "I" of Borges's fictions; all corpora are, to deploy a metaphor, grist to its mechanical mill. But because we also wish to persist in our being as authors (both singular and collaborative), we shall have to imagine anew. Like the narrator who proclaims, "those games belong to Borges now" (324)—and here we retrospectively find intimation of the future game of AI—we shall have to invent new forms, some of which may perhaps not be so readily assimilable to AIs.

In one of our earlier essays, "Inter Alia" (Raley and Samolsky 2019), we imagined a future AI archaeologist coming upon a visual archive of the remnants of human civilization attached to a communications satellite in orbit around Earth. This archive, Trevor Paglen's *The Last Pictures*, is comprised of one hundred photographs etched on a silicon disk that is projected to remain legible for billions of years. One of our claims was that the leading image in this artwork, the verso of Paul Klee's *Angelus Novus*, the celebrated painting that inspired Walter Benjamin's theses on history, would present a limit to what this speculative AI could come to know of Earth's inhabitants and incorporate into its archaeological database.[6] At the time—pre-GPT, even pre-transformer—we thought of this scenario as belonging to the far future, and of course it still does, but the writers who we were then could not have anticipated how swiftly the threat of a technological foreclosing would give rise to our desire to assert the persistence of at least a vestige of the unassimilable human author.

6. The verso contains the painting's label from the Israel Museum as well as a transmittal notice for an exhibit at MCA Chicago.

## References

Bender, Emily, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. New York: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Borges, Jorge Luis. (1957) 1998. "Borges and I." In *Collected Fictions of Jorge Luis Borges*, translated by Andrew Hurley, 324. New York: Penguin.

Chang, Kent K., Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. "Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4." arXiv, October 20. https://doi.org/10.48550/arXiv.2305.00118.

Raley, Rita, and Russell Samolsky. 2019. "Inter Alia: Aliens and AI." *Public* 30, no. 59: 126–37.

Yates, Donald A. 1973. "Behind 'Borges and I.'" *Modern Fiction Studies* 19, no. 3: 317–24.