Automatism for Digital Text Surrealists

electronicbookreview.com/essay/automatism-for-digital-text-surrealists

Nick Montfort
05-05-2024
https://doi.org/10.7273/jbxs-j452
This essay was peer-reviewed.

With this brief look at Large Language Model surrealism, Nick Montfort locates and identifies "the id of the internet, of publishing, of podcasting."

As a computational writing technology, the large language model's true aptitude is for Surrealism. I don't mean simply producing texts in the *style* of Surrealism. I mean furthering the deeper *project* of the Surrealist movement: to surface the unconscious using methods of automatism.

On the one hand, this is a provocative advance for writers who wish to pursue a new form of the Surrealist project. On the other, I find that LLMs will be good for this specific type of writing, but will not radically reconfigure all types of writing — these systems are significant as well as overhyped.

First, I'll note the two aspects of the original avant-garde movement most important to this discussion. Then, I'll proceed to two intermediate ideas crucial to the extension of Surrealism in what I will call Digital Text Surrealism. Finally, I must supply a technical discussion to explain how true or pure LLMs (e.g., the free/libre/open source GPT-NeoX, BLOOM, and Falcon) are the ones that can be directly employed by Digital Text Surrealists, while what I call LLM+SEs (those compelled to obey, follow instructions, and "align") reproduce mental and corporate hierarchies, stripping away the new writing possibilities that LLMs provide.

But first, regarding Surrealism as it was early in the twentieth century:

Surrealism was a literary movement. André Breton defines Surrealism's explorations first and foremost as verbal and written ("soit verbalement, soit par écrit"), going on to say that they may be done in any other way. While it is, of course, not wrong to consider Surrealism as an art movement, in the visual art sense, it was certainly a literary movement.

Surrealism sought to access the unconscious and allow it to express itself directly. In distinction to the strongly anti-art orientation of Dada, for instance, the Surrealists sought mainly to construct a super-reality or surreality by tapping into the unconscious.

With those aspects of Surrealism in mind, these two ideas allow us to adapt and enlarge the movement's individually- and cognitively-focused project:

There is an unconscious that is not personal. We may understand this as a collective unconscious, if we adhere to the ideas of Carl Jung, believing that mythical archetypes exist. However, we do not need to believe this collective unconscious is provided to all of us neurophysiologically, along the lines of Chomsky's universal grammar and the inborn language acquisition device. It could be more fruitfully understood, as Harry T. Hunt argues, as "a socio-cultural collective consciousness, based on a metaphoric imagination."

Automatism can be done without a mind, just with texts. In the present case, I am considering an enormous mass of texts. Tristan Tzara couldn't have had a hat big enough to contain the Pile and other textual resources employed today, although his early attempt at cut-up poetry, which William S. Burroughs describes, provides an analogy for how to automatically write using text alone. Just as the operation of our cognition can be expressed in a concise poem, a mass of textual language can be used to produce a succinct poem that contains striking juxtapositions and suggests the thought behind that corpus, what went into all the writings in it. Considering how the Surrealists proposed to express "le fonctionnement réel de la pensée," we substitute *the actual texts of a corpus* for *the actual functioning of thought.* The mass of text, embodied in a formless corpus, made of posts and books and transcripts, here can be understood as the id of the internet, of publishing, of podcasting.

Now, the distinction between the LLM and (as I call it) the LLM+SE. There are free/libre/open source systems that are LLMs, while an LLM+SE is easily available in the closed, proprietary, opaque system ChatGPT. I cannot go into all of the reasons to prefer free software in this brief article, but there are many. My focus must be on the LLM, the LLM+SE, and how they differ. This distinction, which is essential for the Digital Text Surrealist, requires some technical discussion.

LLMs are (1) transformer-based deep learning systems, (2) pretrained on massive amounts of textual data. A deep learning system, most generally speaking, is a type of machine learning system. Such a system's "learning" has nothing in particular to do with the way humans, or for that matter non-human animals, learn. It simply indicates a software system that improves its performance (measured objectively) on some task over time, or with more data. Of machine learning systems, some are artificial neural networks (ANNs), devised by analogy to the neural networks we and fellow non-human animals have — but not serving as simulations of these. These ANNs are algorithms with specific implementations. The earliest and simplest sort (the perceptron, with a single layer) was developed during 1943–1957, with the earliest steps taken before general-purpose computing. Without proceeding through an elaborate history of machine learning research, an important breakthrough was *deep learning,* which allowed tremendously larger ANNs, with many more layers, to be effectively trained. And, after the development of specific deep learning architectures that include long short-term memory (LSTM), the *transformer* architecture was another extremely important

breakthrough. While significant research was done earlier, the first paper on the transformer architecture only appeared in 2017. As of this writing, it has been cited more than 100,000 times; because I am only mentioning the architecture here as one of many points on a timeline, I'll refrain from adding another citation.

The other element of an LLM is the extremely large set of textual data used for *pretraining.* LLMs can be further trained (fine-tuned). But it happens that amazingly cohesive texts of many different sorts can be generated without this fine-tuning process. These are *cohesive* rather than *coherent,* as there is nothing about an LLM that gives it any semantic understanding. The amazing thing about these models is how fluent they sound *without* such understanding. They are extraordinarily impressive text generators, and should be praised as computational breakthroughs. They also have no intelligence, artificial or otherwise, in the sense of comprehension, perception, or intellect. They are (very good) probability distributions over sequences of words.

LLM+SEs are augmented systems, with an added superego (SE), developed by corporations in an attempt to restrain the raw id of the LLM. Specifically, reinforcement learning with human feedback (RLHF) is the most prominent technique by which researchers seek to "align" LLMs to their interests — bending these language-producing machines into compliant customer service representatives. Reinforcement learning is a venerable *supervised* machine learning technique in which upvotes and downvotes are used to improve performance. As more of these are applied over time, the process alters a software system to more frequently produce the desired results and less often produce those that are undesirable.

There are various ways to employ reinforcement learning, some of which don't involve people at all — a game-playing system can play itself, for instance, rewarding itself for a win and punishing itself for a loss. But RLHF does involve people. Specifically, *Time* magazine first reported, it involves Kenyan workers paid less than $2/hour to view and rate horrific texts (such as a description of bestiality performed in front of a child) in order to improve ChatGPT. An investigation by *The Wall Street Journal* reporters revealed further details about the effect of these traumatic texts. Workers encountered offensive output that OpenAI had specifically coaxed from system.

Reporters in the *Journal* wrote that sexual and violent texts shown to workers were initially brief, but grew over time to five or six paragraphs in length. One worker on the violent-content team read texts "describing heinous acts, such as people stabbing themselves with a fork or using unspeakable methods to kill themselves." Another on the sexual-content team "read detailed paragraphs about parents raping their children and children having sex with animals." After having nightmares and withdrawing from the world, his wife and stepdaughter eventually left him.

The id of digital texts — written largely by those of us in the Global North — is more than disturbing. We can imagine the person providing a ChatGPT prompt, whether a typical privileged user or a Digital Text Surrealist, as the ego. The superego, interestingly enough, is supplied thanks to workers the Global South.

Digital Text Surrealists who use LLMs voluntarily and with an awareness of what they face may generate texts of many different sorts. Indeed, they may choose to bring the horrific sexual and violent side of the collective unconscious to light. If this is done with a willingness to reveal what lies beneath, very well.

Some “raw” LLM results are surprising without including traumatic language. Some even lack innovative stylistic twists. Still, these outputs may be uncanny and compelling. Digital Text Surrealists have yet to develop a serious poetics of LLMs, but such potential is latent. Scott Rettberg has written, with awareness of the dark side and the potential of the most famous LLM+SE, that “recent developments ... place the arts and humanities in an important experimental role in exploring how these forms can and will be used.” The point holds at least as much for an LLM without the SE. Digital Text Surrealists will have to peer into the id of our writings, not to serve corporations and train a more entertaining search engine, but to be an ax for the frozen sea of language.

The author makes another pilgrimage to the birthplace of Surrealism in April 2024. Photo by Nick Montfort.

I wish to conclude by offering a final poem of “my own” (actually an LLM-generated poem), and by letting the poem itself serve to conclude this discussion. My attempt at interpreting or analyzing it could foreclose more interesting responses from readers, so I would like to let it resonate — if it is going to. Nevertheless, I’ll clear my throat (or the throat of my text-to-speech system) to some extent beforehand, to offer two points of comparison. Here are two specific Surrealist texts from about a century ago, both from the seminal, collaborative work of Surrealism, André Breton and Philippe Soupault’s 1920 *Les Champs magnétiques*. The first is a short poem which I quote in English translation by Charlotte Mandell:

> *Feelings are Free*
>
> Trace smell of sulfur
> Public health swamp
> Red of criminal lips
> Quick-march brine
> Whim of monkeys
> Day-colored clock

The disconnected lines seem imagist, or at least based around perceptions. Any psychology is rather subtle. Why does this writing call attention to the typically sensual lips of criminals? Are the authors themselves the "monkeys" whose whim produces texts like these? The function of the clock blurs into its color, or perhaps the fact that the clock face is visible at all, during the daytime, serves as the first signal of the time.

Next, consider something that seems, to me, more connected, even bordering on narrative. This is the first paragraph of Breton and Soupault's book, presented as a prose paragraph under the heading "The Unsilvered Glass," in David Gascoyne's translation:

> Prisoners of drops of water, we are but everlasting animals. We run about the noiseless towns and the enchanted posters no longer touch us. What's the good of these great fragile fits of enthusiasm, these jaded jumps of joy? We know nothing anymore but the dead stars; we gaze at their faces, and we gasp with pleasure. Our mouths are as dry as the lost beaches, and our eyes turn aimlessly and without hope. Now all that remain are these cafés where we meet to drink these cool drinks, these diluted spirits, and the tables are stickier than the pavements where our shadows of the day before have fallen.

While the later lines may *suggest* that the poets are whimsical monkeys, this one outright declares us to be (non-human) animals, doing little in a dwindling environment.

The final thing I will point out about these two prototypical Surrealist texts is that, quite aside from what is declared, the language in both of them is extremely ornate. Nouns are almost never allowed to surface from the subconscious without adjectives buoying them up.

The following one particular work of Digital Text Surrealism may not represent the entire practice, but consider if it may align with some of the aspects of Breton and Soupault's work while deviating from others.

So, here, a poem I find telling but not traumatizing, one that I produced by running GPT-NeoX on my own feeble writer's laptop. (This explains the generation time of more than an hour.) The parameters not explicitly indicated have the default values; for the temperature — which ranges from a completely deterministic value of 0 up through and past an extremely

random 1.0 — I used a reasonably low value, considering that I was attempting to tap into the collective unconscious. The text below is *exactly* as generated — line breaks included — except that I added the last two punctuation marks.

> Temperature: .7
> Maximum Length: 70
> >My biggest fear is that
> GENERATED IN 1:03:34.423080
> Mon, 8 Jan 2024 21:30:50 -0500 (EST)
>
> My biggest fear is that
> somebody will
> be walking down the street
> and turn to the person next to them.
> And they'll say, "Are you a dog?"
> And the person they're talking to
> will be like,
> "No, I'm not a dog."
> And they'll reply, "Oh."

References

Breton, André. *Manifestes du surréalisme*, Paris: Gallimard, coll. Idées, 1967.

Breton, André. *Manifestoes of Surrealism*. Vol. 182. University of Michigan Press, 1969.

Breton, André and Philippe Soupault, *The Magnetic Fields.* Translated by David Gascoyne, 3rd ed. London: Atlas Press, 1985.

Breton, André and Philippe Soupault, "André Breton and Philippe Soupault: From 'The Magnetic Fields,' 'Feelings Are Free.'" Posted by Jerome Rothenberg, translated by Charlotte Mandell. *Jacket2.* May 22, 20220. https://jacket2.org/commentary/andr%C3%A9-breton-and-philippe-soupault

Burroughs, William S. "The cut-up method of Brion Gysin." *The New Media Reader,* pp. 89–91, 2003.

Hao, Karen, and Deepa Seetharaman. "Cleaning up ChatGPT takes heavy toll on human workers." *The Wall Street Journal,* July 24, 2023. https://wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483

Hunt, Harry T. "A collective unconscious reconsidered: Jung's archetypal imagination in the light of contemporary psychology and social science." *The Journal of Analytical Psychology* vol. 57:1 (2012): 76-98. doi:10.1111/j.1468-5922.2011.01952.x

Jung, Carl Gustav. “The concept of the collective unconscious.” *Collected Works* 9.1, pp. 99–104, 1936.

Perrigo, Billy. “Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic.” *Time.* January 18, 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/

Rettberg, Scott. “Cyborg Authorship: Writing with AI — Part 1: The Trouble(s) with ChatGPT.” *ebr.* July 2, 2023. Cyborg Authorship: Writing with AI – Part 1: The Trouble(s) with ChatGPT

Cite this Essay:

Montfort, Nick. “Automatism for Digital Text Surrealists”, Electronic Book Review, May 5, 2024, https://doi.org/10.7273/jbxs-j452.
Readers wishing to respond to an essay in *ebr* may send ripostes or short glosses to the journal’s Managing Editor, Will Luers.
This essay was peer-reviewed.