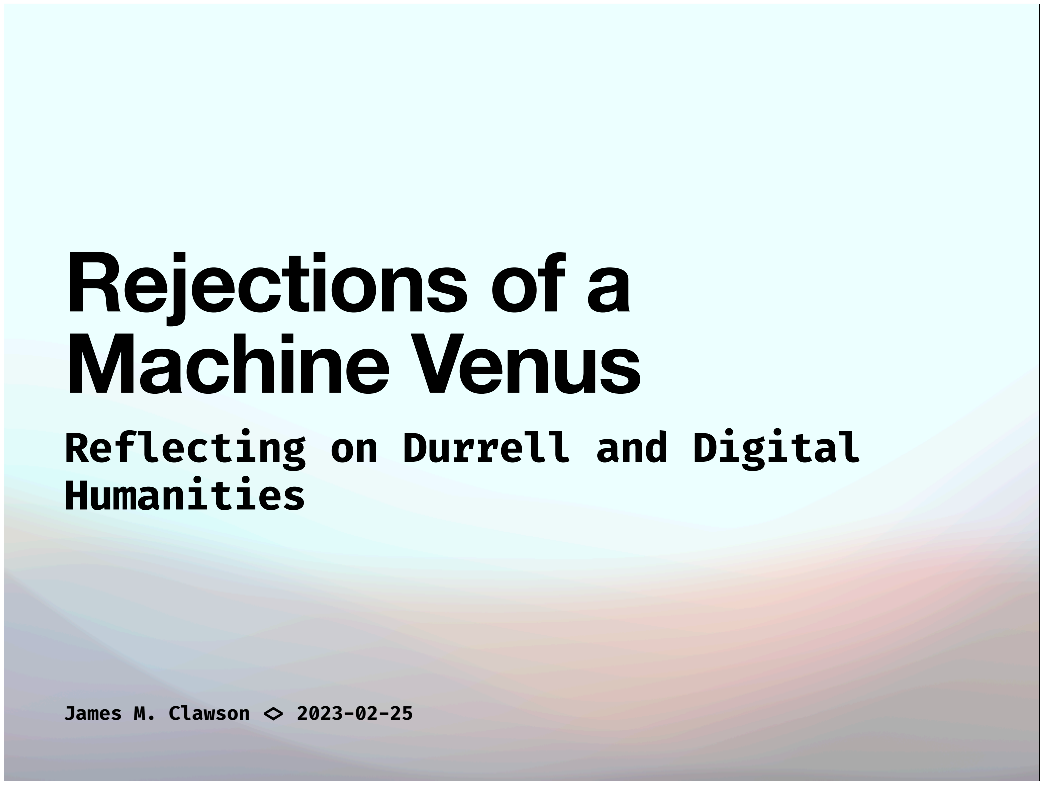


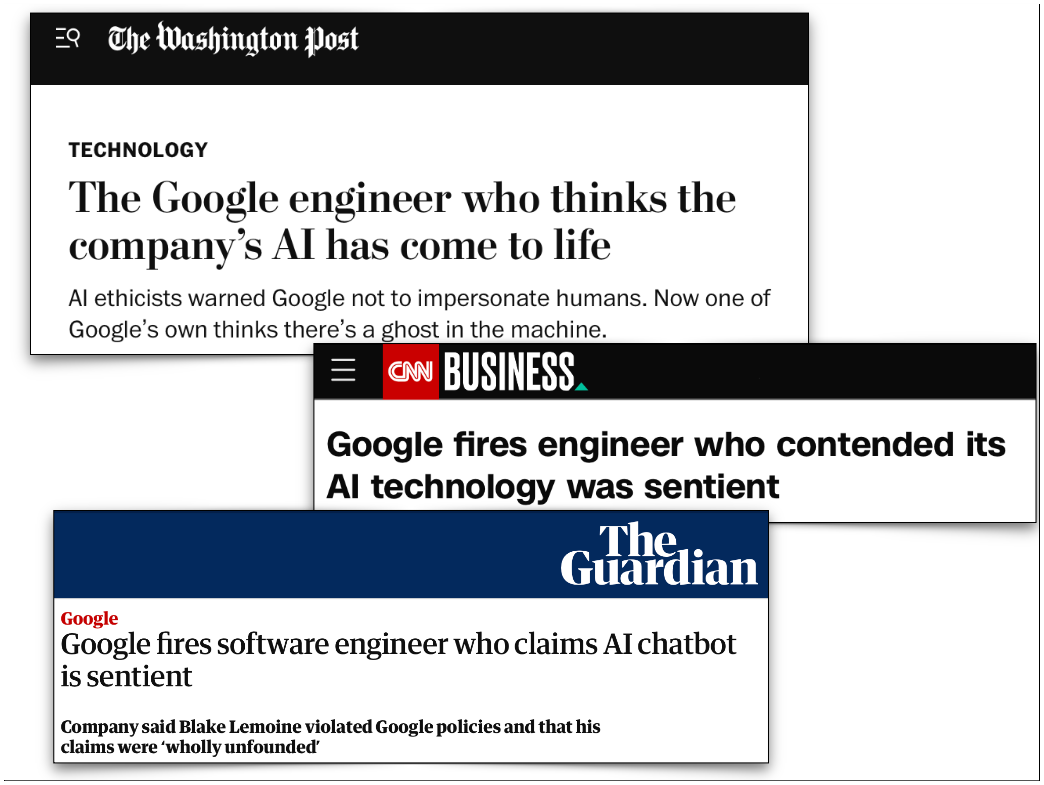
Rejections of a Machine Venus

Reflecting on Durrell and Digital Humanities

James M. Clawson <> 2023-02-25

Presented at the 50th Louisville Conference on Literature and Culture Since 1900, held 23–25 February 2023 at the University of Louisville, KY.

http://jmclawson.net


The Washington Post

TECHNOLOGY

The Google engineer who thinks the company's AI has come to life

AI ethicists warned Google not to impersonate humans. Now one of Google's own thinks there's a ghost in the machine.

CNN BUSINESS

Google fires engineer who contended its AI technology was sentient

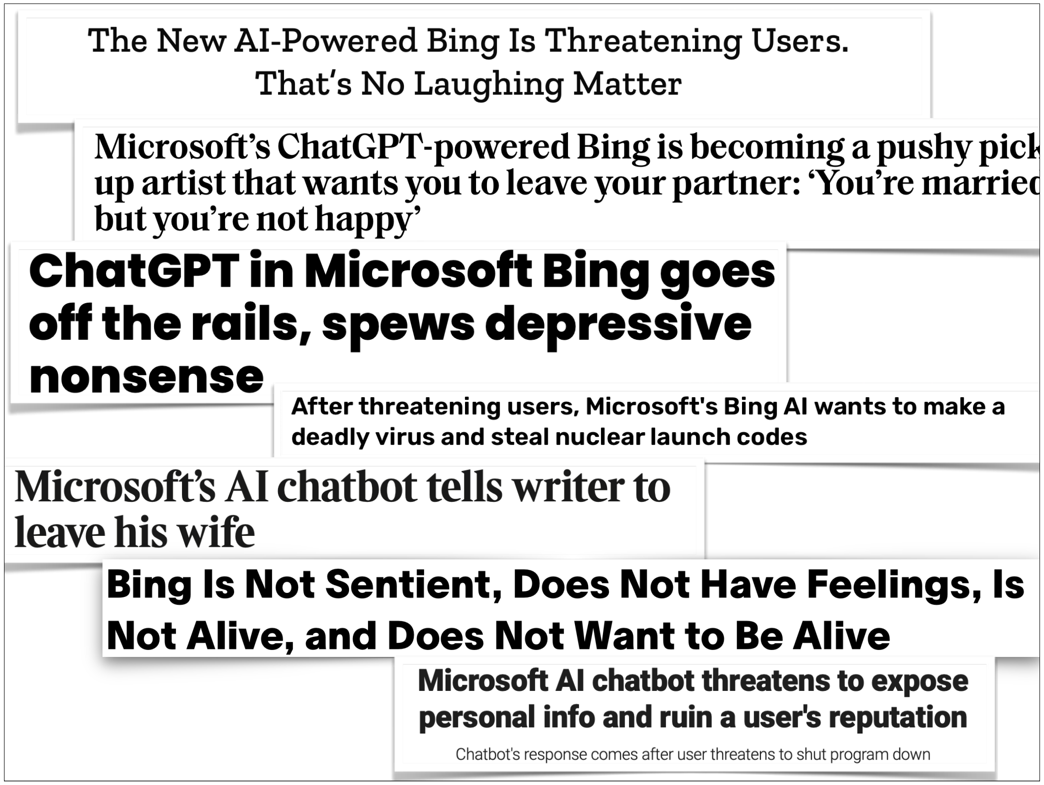
The Guardian

Google

Google fires software engineer who claims AI chatbot is sentient

Company said Blake Lemoine violated Google policies and that his claims were 'wholly unfounded'

I proposed this topic last fall, when the next wave of AI text generation tools prompted headlines of a Google researcher who thought the model was sentient, spoke out, and then got fired. It seemed then like we were reaching this critical point of needing public conversations about understanding sentience and technology.



The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter

Microsoft's ChatGPT-powered Bing is becoming a pushy pick up artist that wants you to leave your partner: 'You're marrie
but you're not happy'

ChatGPT in Microsoft Bing goes off the rails, spews depressive nonsense

After threatening users, Microsoft's Bing AI wants to make a deadly virus and steal nuclear launch codes

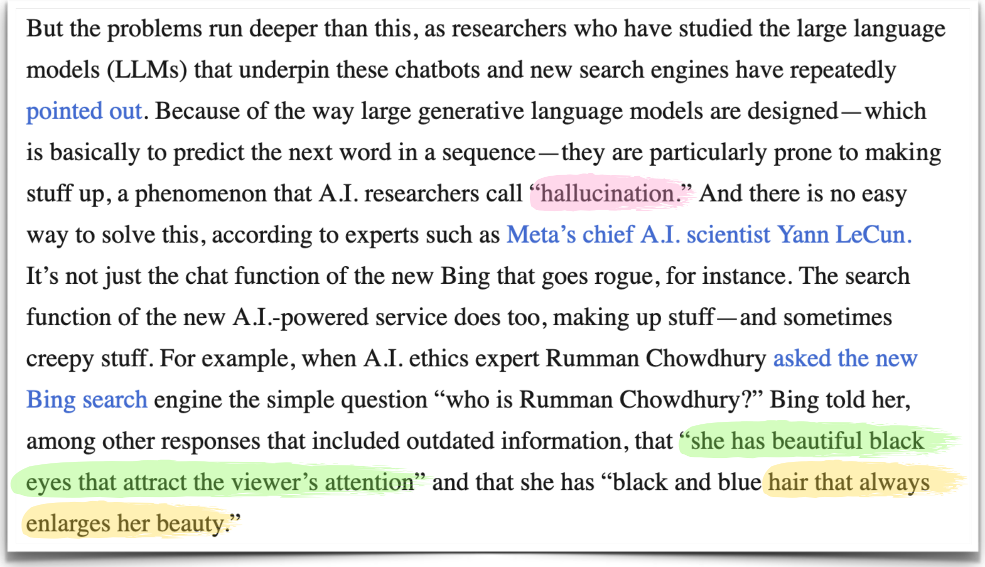
Microsoft's AI chatbot tells writer to leave his wife

Bing Is Not Sentient, Does Not Have Feelings, Is Not Alive, and Does Not Want to Be Alive

Microsoft AI chatbot threatens to expose personal info and ruin a user's reputation

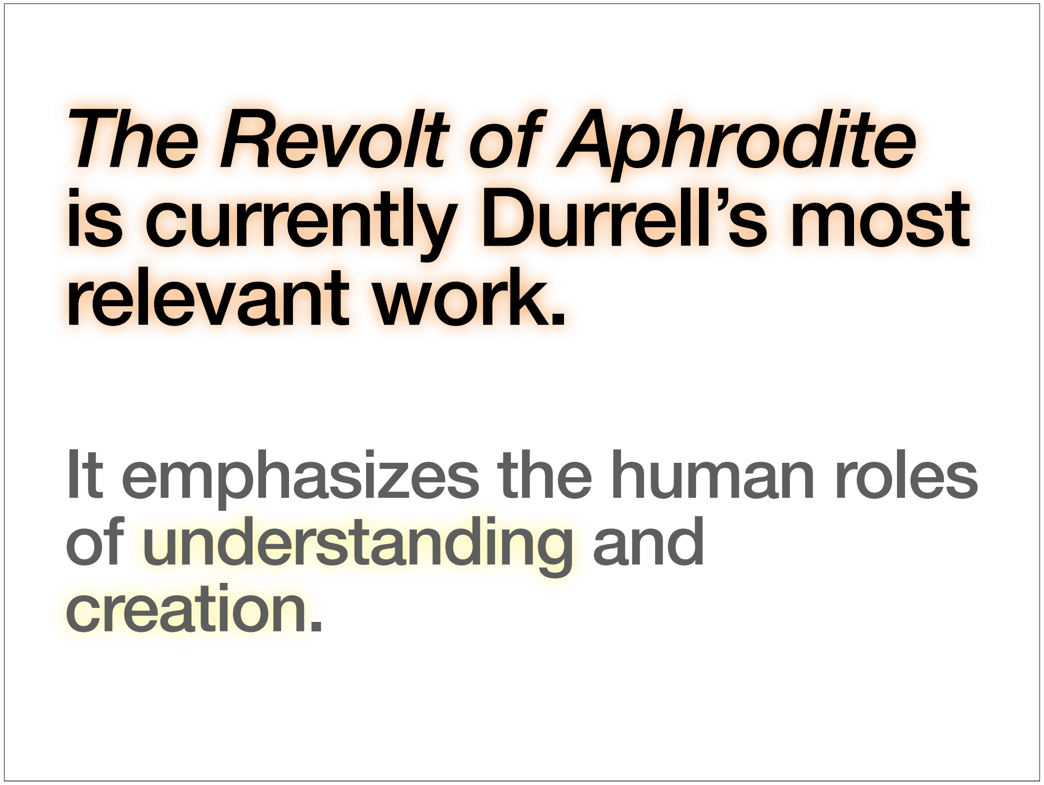
Chatbot's response comes after user threatens to shut program down

Since then, the headlines have shown the moment to be even more desperate.



But the problems run deeper than this, as researchers who have studied the large language models (LLMs) that underpin these chatbots and new search engines have repeatedly pointed out. Because of the way large generative language models are designed—which is basically to predict the next word in a sequence—they are particularly prone to making stuff up, a phenomenon that A.I. researchers call "hallucination." And there is no easy way to solve this, according to experts such as Meta's chief A.I. scientist Yann LeCun. It's not just the chat function of the new Bing that goes rogue, for instance. The search function of the new A.I.-powered service does too, making up stuff—and sometimes creepy stuff. For example, when A.I. ethics expert Rumman Chowdhury asked the new Bing search engine the simple question "who is Rumman Chowdhury?" Bing told her, among other responses that included outdated information, that "she has beautiful black eyes that attract the viewer's attention" and that she has "black and blue hair that always enlarges her beauty."

Perhaps most interesting is the phenomenon popularly called "hallucination," when a large language model fills in gaps with plausible (but possibly untrue) details.



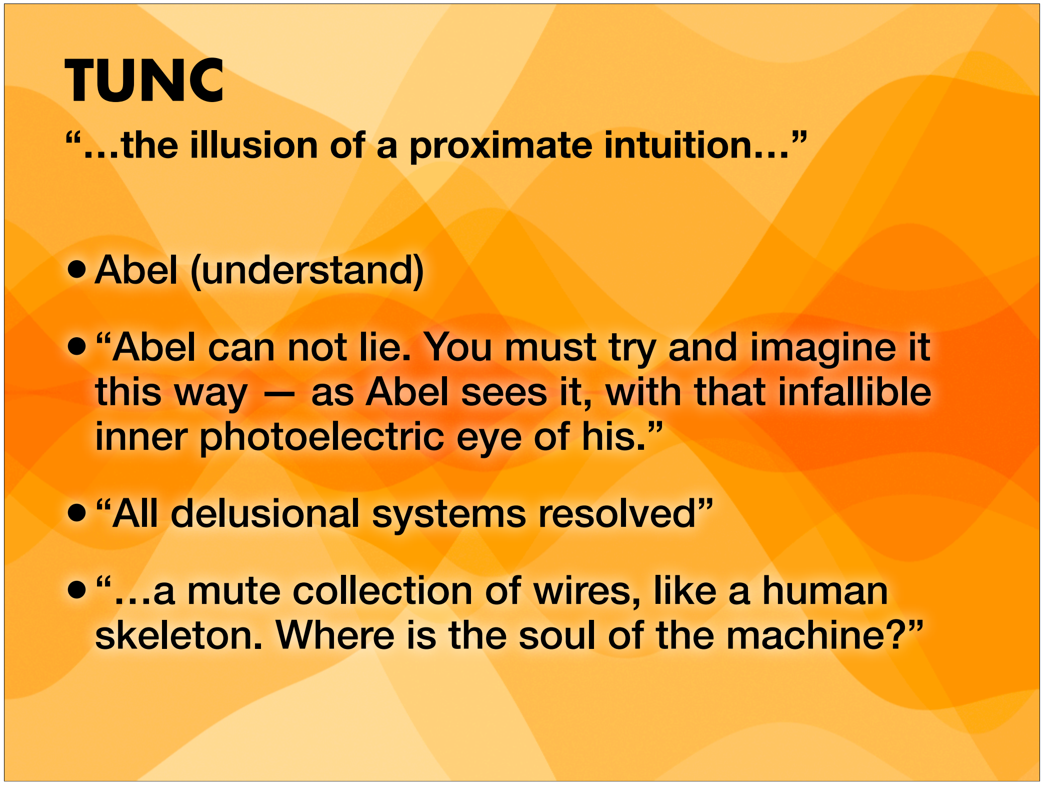
The Revolt of Aphrodite is currently Durrell’s most relevant work.

It emphasizes the human roles of understanding and creation.

So these concepts are coming to the fore with increasing urgency—the idea of humans needing to re-learn and come to terms with the ways we understand our evolving interactions with computers. It’s in this context that I’d really like to stake a claim:

- Durrell’s under-studied two-book sequence *The Revolt of Aphrodite* is currently his most relevant work.
- Because it’s about determinism and free will, it’s fundamentally about human agency.

Works like *Revolt of Aphrodite* emphasize the human roles of understanding and creation
By showing *computer* understanding and *computer* creation, invites us to engage with the idea of what these things mean in ways that are increasingly important.



TUNC

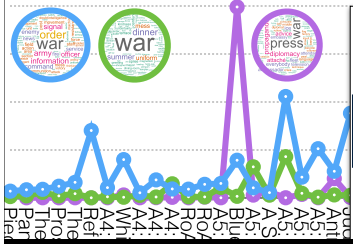
“…the illusion of a proximate intuition…”

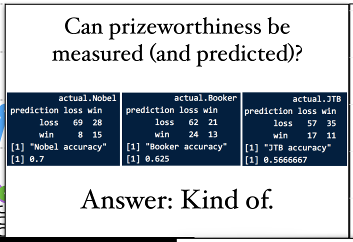
- Abel (understand)
- “Abel can not lie. You must try and imagine it this way — as Abel sees it, with that infallible inner photoelectric eye of his.”
- “All delusional systems resolved”
- “…a mute collection of wires, like a human skeleton. Where is the soul of the machine?”

The first book of the series, *Tunc*, explores some models of computer understanding.

Digital Humanities Understanding

“…the illusion of a proximate intuition…”





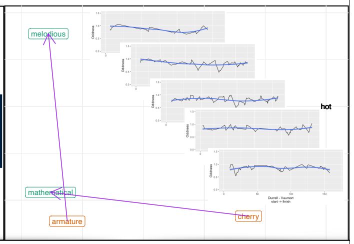
Can prizeworthiness be measured (and predicted)?

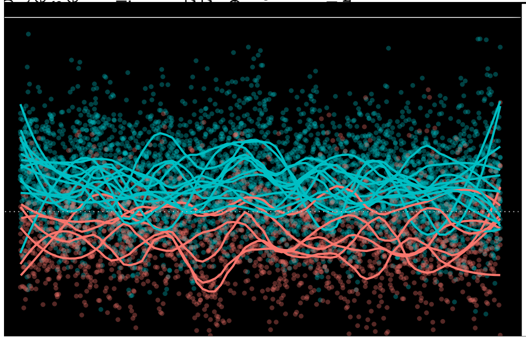
```
          actual.Nobel
prediction loss win
      loss   69  28
      win     8  15
[1] "Nobel accuracy"
[1] 0.7
```

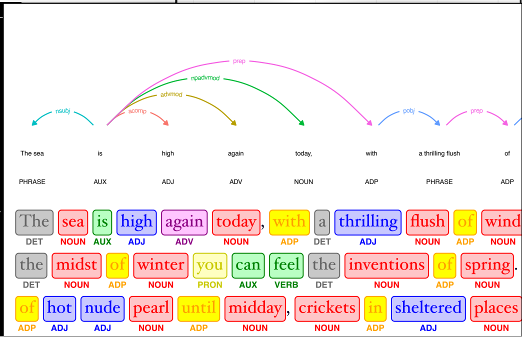
```
          actual.Booker
prediction loss win
      loss   62  21
      win    24  13
[1] "Booker accuracy"
[1] 0.625
```

```
            actual.JTB
prediction loss win
      loss   57  35
      win    17  11
[1] "JTB accuracy"
[1] 0.5666667
```

Answer: Kind of.



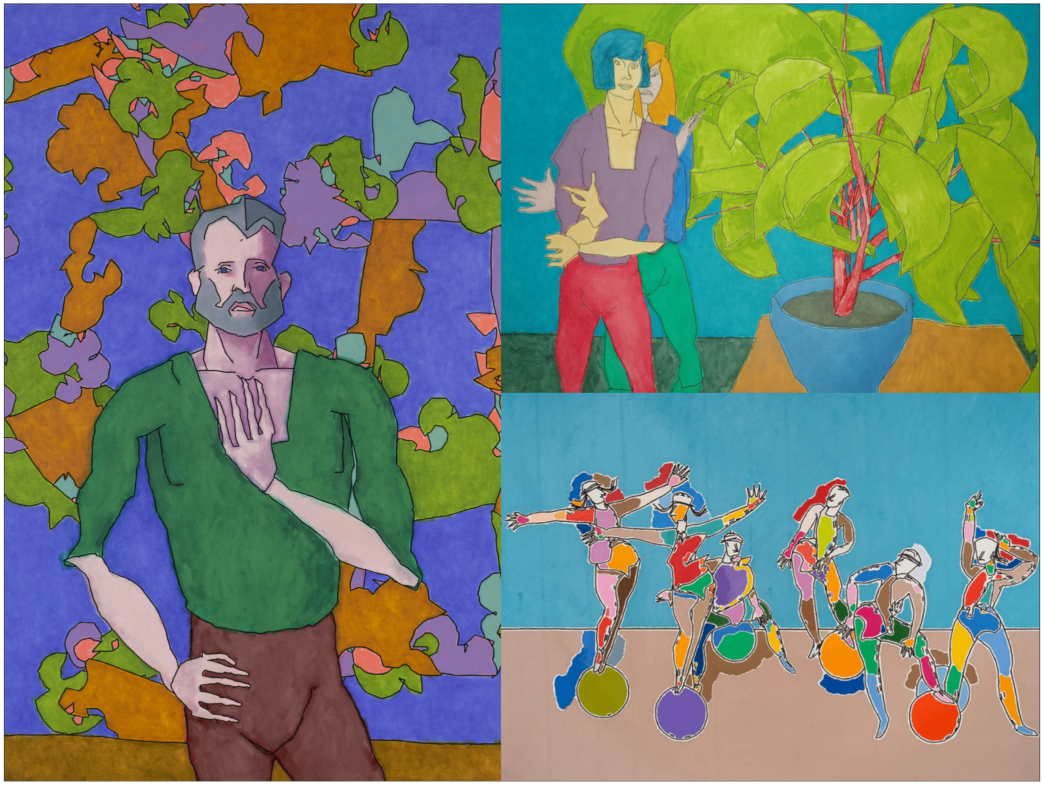



The sea is high again today, with a thrilling flush of wind
the midst of winter you can feel the inventions of spring.
of hot nude pearl until midday, crickets in sheltered places
DET NOUN AUX ADJ ADV NOUN ADP DET ADJ NOUN ADP NOUN
PRON AUX VERB

Working on Durrell and digital humanities since 2016, I’ve also used a computer to help me approach a proximate intuition.

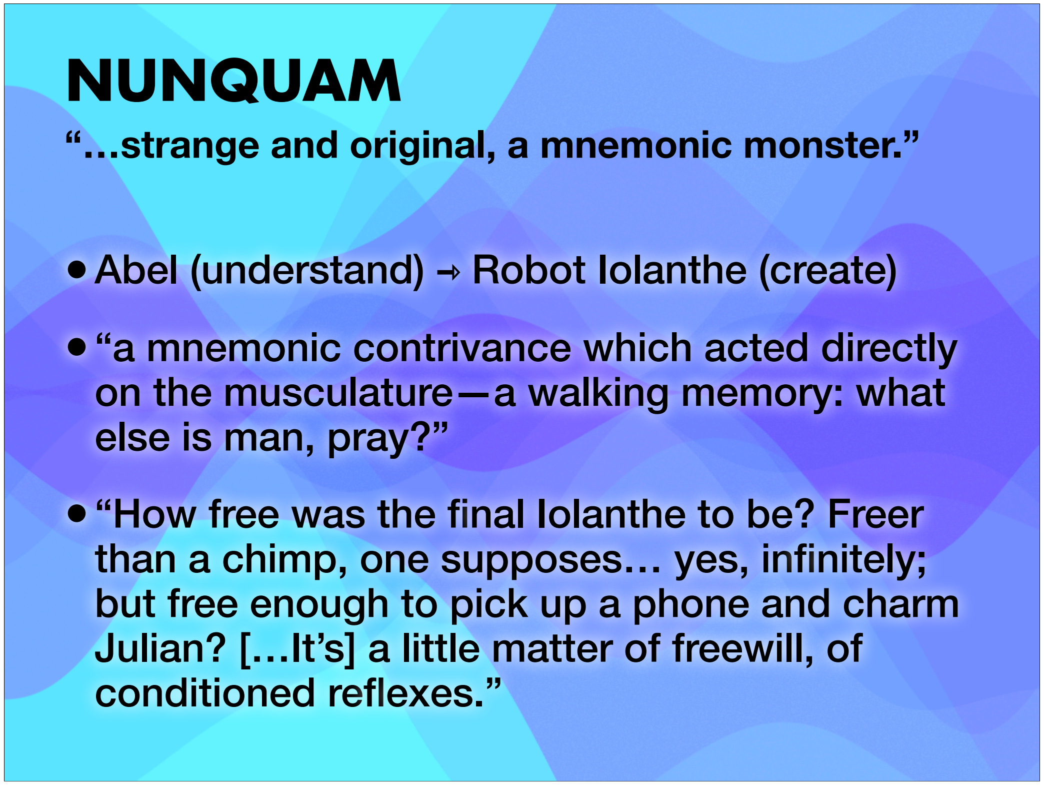
- (left to right, top row) Topic modeling, clustering, word embedding models (miniaturized language models)
- (next row) stylometry to measure “literariness”, natural language processing to study structures and patterns in phrases and sentences.

These are interpretive acts, with understanding necessary to design the project and to make use of the results. But this is digital understanding. What we’ve seen lately is an explosion in digital creation.



Although seeming new in its domination of our news cycles, computer-generated art, or computer *creation*, has been around for awhile. Margaret Boden in her early-2000s book *The Creative Mind* points out the ways that computers have already been called on engagement in creative pursuits. Harold Cohen is one of the examples she offers (with examples shown here). Cohen was an artist who created a generative computer system (1970s through 1980s) that would plan its own images and draw them, eventually coloring them, too. By exploring a conceptual space described by their creators, Cohen's AARON system for visual art generation, David Cope's EMI system for music generation, and others like them make up the second of three types of computer-driven creativity Boden describes.

Music - Steve Reich
second of 3 types of creation

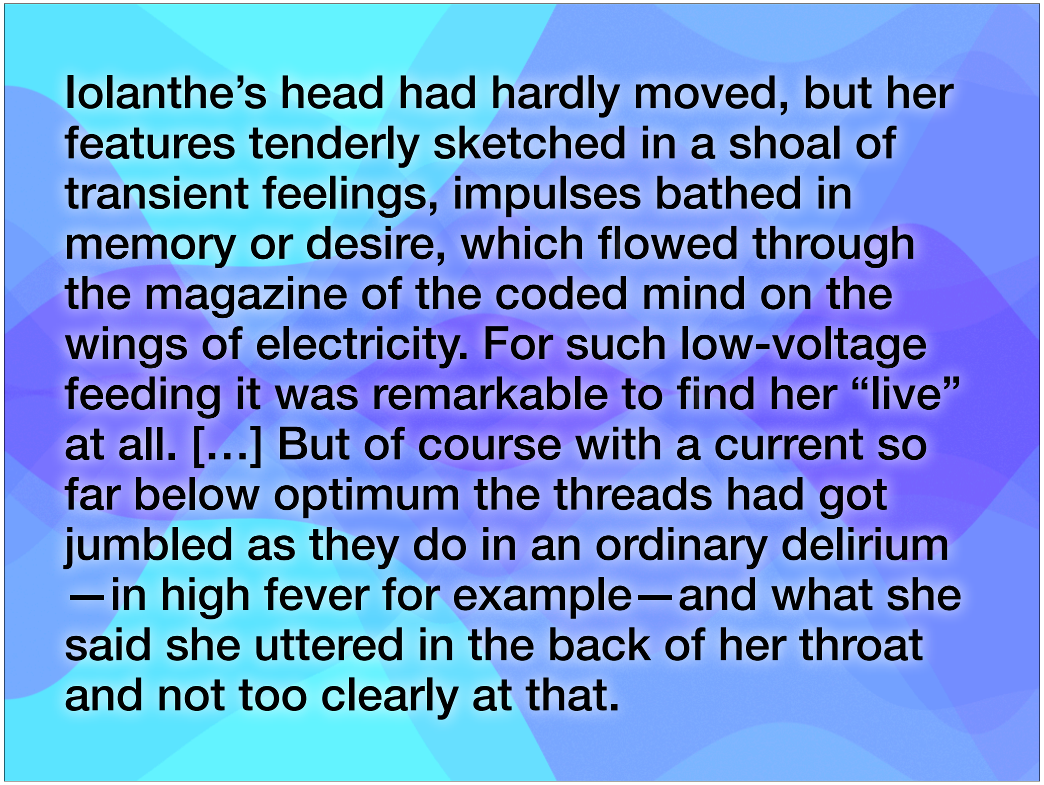


NUNQUAM

“…strange and original, a mnemonic monster.”

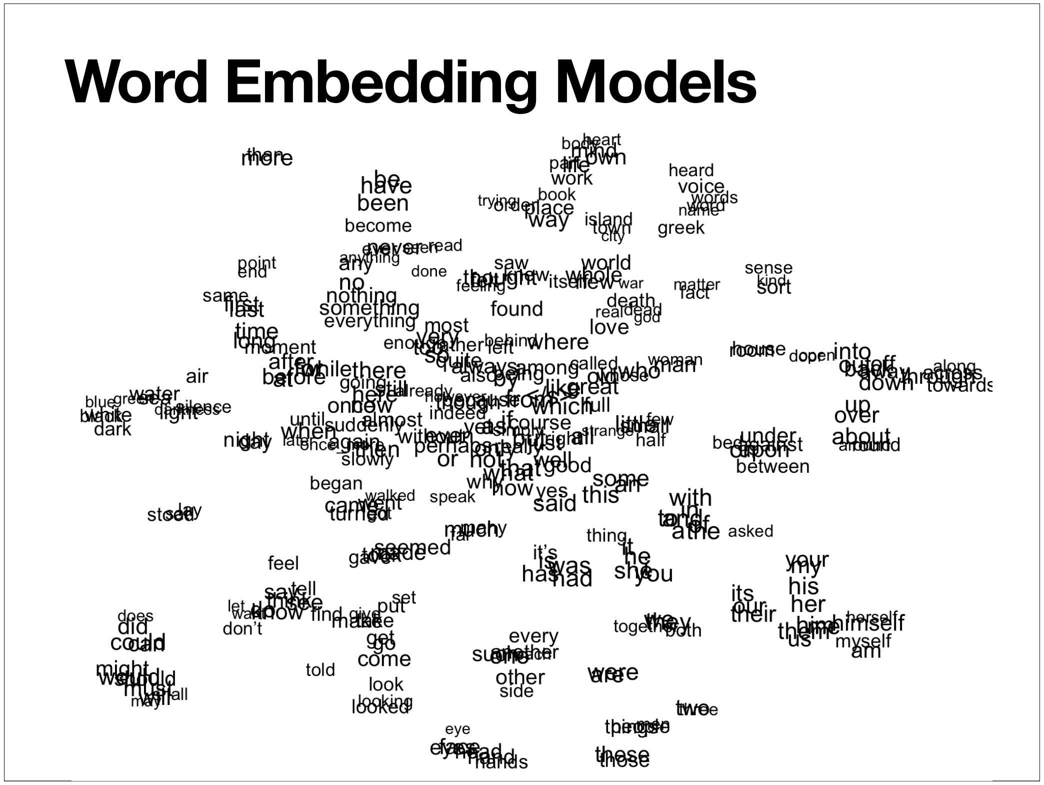
- Abel (understand) → Robot Iolanthe (create)
- “a mnemonic contrivance which acted directly on the musculature—a walking memory: what else is man, pray?”
- “How free was the final Iolanthe to be? Freer than a chimp, one supposes… yes, infinitely; but free enough to pick up a phone and charm Julian? […It’s] a little matter of freewill, of conditioned reflexes.”

In *Nunquam*, the second book of the series, Durrell moves from computer understanding to computer creation.



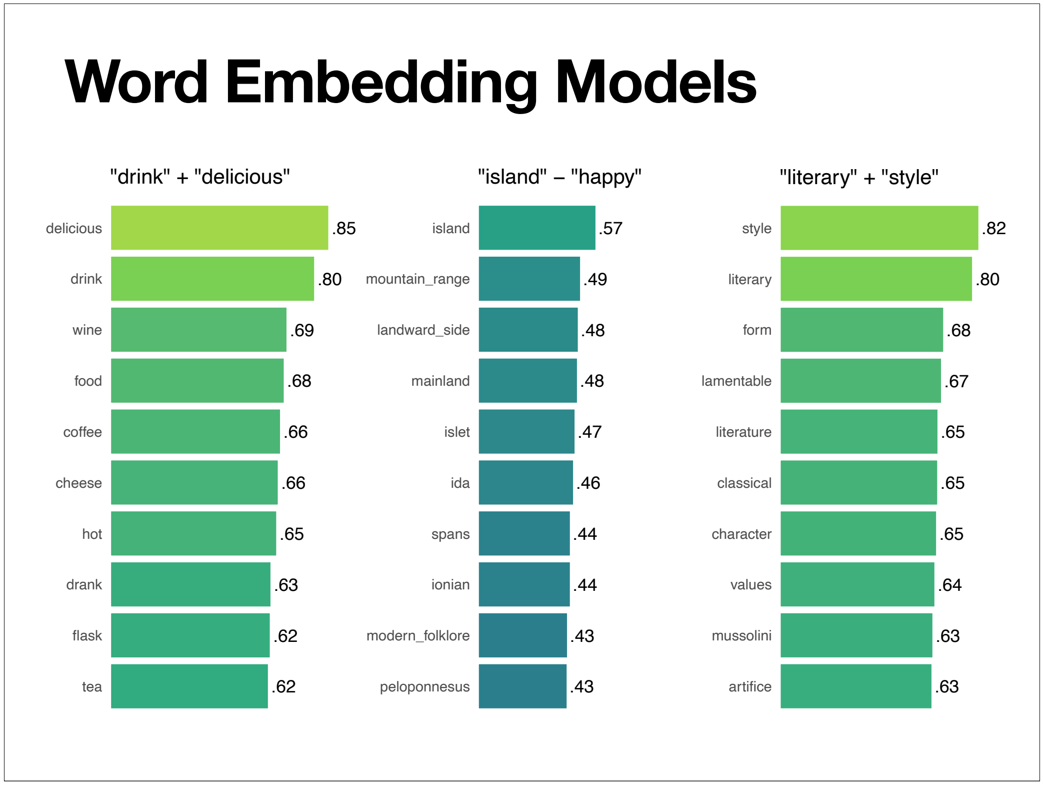
Iolanthe’s head had hardly moved, but her features tenderly sketched in a shoal of transient feelings, impulses bathed in memory or desire, which flowed through the magazine of the coded mind on the wings of electricity. For such low-voltage feeding it was remarkable to find her “live” at all. […] But of course with a current so far below optimum the threads had got jumbled as they do in an ordinary delirium—in high fever for example—and what she said she uttered in the back of her throat and not too clearly at that.

We even get scenes evoking the same kinds of “hallucinations” AI researchers have found lately.



Word Embedding Models

Using tools of digital humanities, we can try some simple versions of this, too. For instance, we might be inspired by Charlock's method of pouring as much data as possible into a computer model and seeing what happens. This slide shows just a *minuscule* portion of a word embedding model derived from more than 2.3 million words written by Durrell: 18 books of fiction, 6 travel books, 3 books of essays and letters, 1 book of poetry. Words here are shown in clusters of similarity.



Word Embedding Models

…And we can use this model to ask simple questions.

- For instance, what words might Durrell think of describing as a tasty beverage (*drink* plus *delicious*)? Surprising nobody, the answers include “wine” very high in the list.
- Or what might he note about an island that doesn’t spark happiness (*island* minus *happy*)? In Durrell’s writing, it seems like islands are only happy places if they allow one to focus on the sea, since an emphasis on mountain ranges and the mainland are where happiness isn’t.
- Or what might he say if asked about literary style (*literary* plus *style*)? Here we might be unsurprised to learn of an emphasis on “form” and a consideration for “classical” features. But it’s suggestive that we also see “artifice” quite high up, as perhaps suggesting a facade disguising other elements of a work. Is a literary style, then, something that could be considered artificial in Durrell’s work? (Spoiler: Further work I’ve done on this question, hopefully coming out this year, suggests it is!)

In this way, we can probe the connections a computer *has* been able to make among all the words in Durrell’s writing to infer connections that might not explicitly be there. But this is limited and potentially unsatisfying. Even with 2.3 million words, this model is small compared to truly **large** language models like ChatGPT…

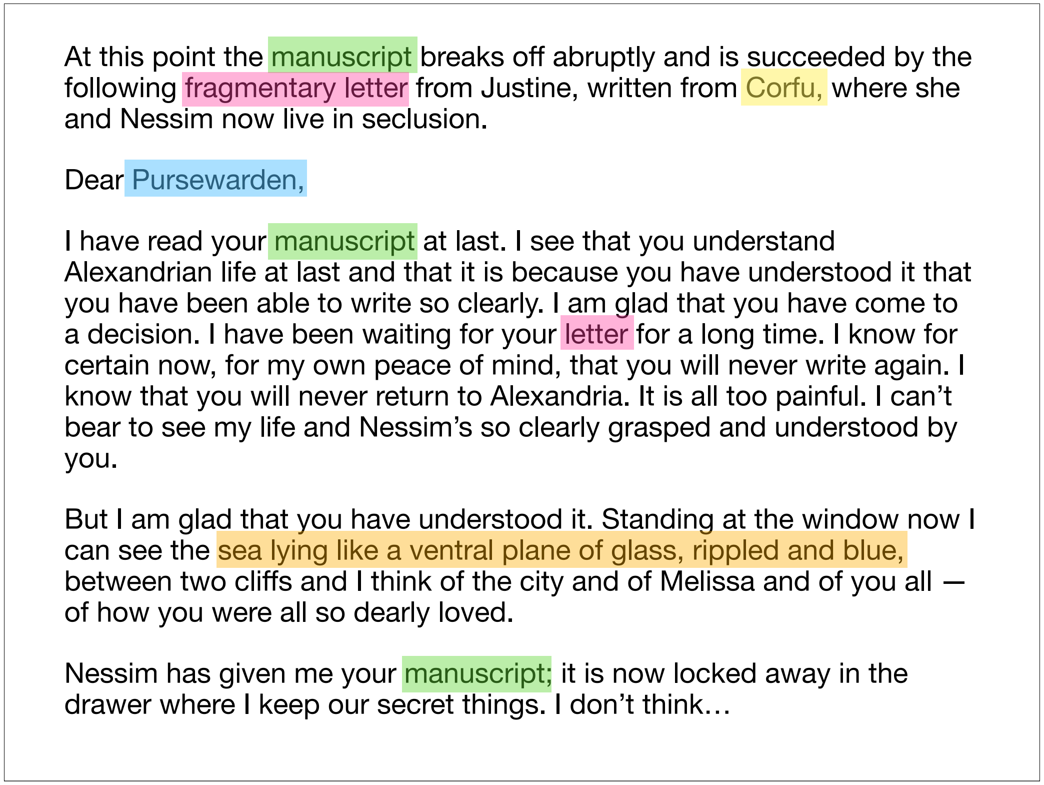
ChatGPT in Durrell's voice

Prompt below (with response on next slide)

At night when the wind roars and the child sleeps quietly in its wooden cot by the echoing chimney-piece I light a lamp and walk about, thinking of my friends — of Justine and Nessim, of Melissa and Balthazar. I return link by link along the iron chains of memory to the city which we inhabited so briefly together: the city which used us as its flora — precipitated in us conflicts which were hers and which we mistook for our own: beloved Alexandria!

I have had to come so far away from it in order to understand it all! Living on this bare promontory, snatched every night from darkness by Arcturus, far from the lime-laden dust of those summer afternoons, I see at last that none of us is properly to be judged for what happened in the past. It is the city which should be judged though we, its children, must pay the price.

…So before wrapping things up, it seems necessary to show some of the potential for computer creation as shown by ChatGPT 3.0. Here is a prompt given to the model, with an early passage from *Justine* written by Durrell. What follows is the LLM's response, continuing onward from the prompt, filling in details in a way that seems most plausible to the model. In this way, it evokes Darley's attempts in *Balthazar* to fill in the gaps from the *known* to the *unknown*: "I must record what more I know and attempt to render it comprehensible or plausible to myself, if necessary, by an act of the imagination."



At this point the manuscript breaks off abruptly and is succeeded by the following fragmentary letter from Justine, written from Corfu, where she and Nessim now live in seclusion.

Dear Pursewarden,

I have read your manuscript at last. I see that you understand Alexandrian life at last and that it is because you have understood it that you have been able to write so clearly. I am glad that you have come to a decision. I have been waiting for your letter for a long time. I know for certain now, for my own peace of mind, that you will never write again. I know that you will never return to Alexandria. It is all too painful. I can't bear to see my life and Nessim's so clearly grasped and understood by you.

But I am glad that you have understood it. Standing at the window now I can see the sea lying like a ventral plane of glass, rippled and blue, between two cliffs and I think of the city and of Melissa and of you all — of how you were all so dearly loved.

Nessim has given me your manuscript; it is now locked away in the drawer where I keep our secret things. I don't think…

…ChatGPT's response provides a very plausible continuation that seems very clearly modeled on Durrell's style. A few parts stand out:

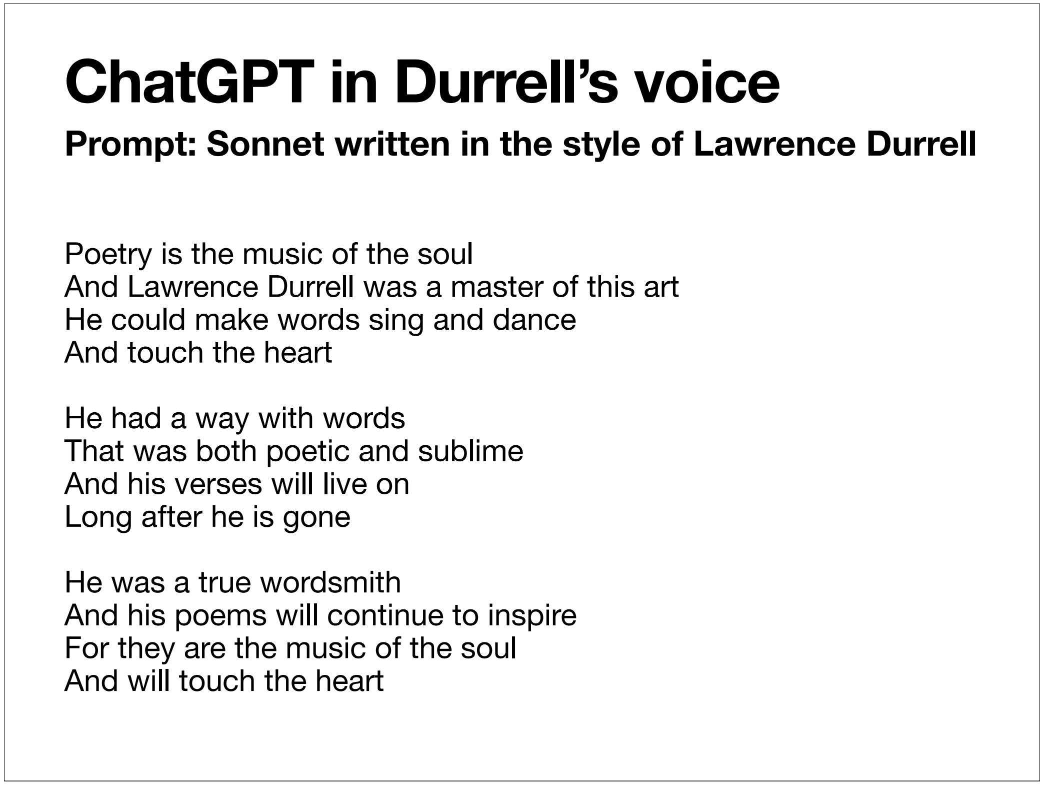
- In green and pink: the interest in manuscripts and letters. Artifacts of textuality are littered throughout the *Quartet* even though they're missing from the prompt. The model clearly has a sense of the kinds of things one might expect to see when it "imagines" a continuation of Durrell's writing.
- In yellow and blue: the use of places and names that belong to Durrell's context. Because these proper nouns are so specific and unmentioned in the prompt, they might be unsettling to see here. How has the model provided them? Does it know more than it's letting on? From *Nunquam*, we see a similar moment of discomfort when the robot version of Iolanthe seems to know more than expected: "We had even forgotten (how is this possible: please tell me?), we had forgotten that she would know all about us, even our names." Because the robot in *Nunquam* is trained on all the memories of the actress Iolanthe, she has access to experiences that happened before her own creation. Here, too, the model shows a bias in its training. Even though the prompt never mentions Pursewarden, he shows up in ChatGPT's response, suggesting that the model has been trained on the *Quartet* sufficiently to recognize the prompt and to reference a character in those books. And even though the *Quartet* never mentions Corfu, the model knows that it's a location with an outsized resonance in Durrell's other works.
- In orange: a very Durrellian description that is somehow familiar seeming and yet entirely new. The word "ventral" appears only once in Durrell's works. In *Nunquam*, Goytz uses it in an anatomical description of embalming, one of many moments of

ventriloquism in Durrell's works where the text assumes an outsider's voice. But here it's used poetically to describe calm waters, in a way we might be surprised to discover Durrell never used the word. Here, the model is doing more than reappropriating familiar items and proper nouns to approach a familiar style, convincingly ventriloquizing Durrell and offering up things he never actually said.

Using Large Language Models like ChatGPT in this way seems useful for exploring the borders of our own expectations.

- Generative artist Harold Cohen called his AARON program not "an expert system" but "an expert's system" — a "research tool for the expansion of my own expert knowledge" of art.
- We might use these tools in the same way.
- Generative models offer opportunities for the expansion of our understanding—serving us in our pursuits as scholars, because of their ability to generate new opportunities beyond what methods of digital analysis can reveal.
- How might studying methods of digital creativity further our understanding of non-digital works of art? What does *our* reaction to the computer's mimicry of Durrell's style reveal about that target style and our understanding of it? And how does it engage with or challenge our notions of authenticity?
- Cohen has also described style as "the signature of a complex system." Are generative systems forgeries, then, or something else? When they can give us something that ***isn't Durrell*** yet still recognizably be ***in the style of Durrell***, where are we left to situate our understanding of stylistic signature?

I have no answers to these questions yet, but I'm left reemphasizing my point that *Revolt of Aphrodite* is currently Durrell's most relevant work because it engages with this two-natured question of *understanding* and *creating*. From some perspectives, *creating* seems like an unacknowledged or under-appreciated milestone of *understanding*, a lever long enough to help us move the world if only we can figure out where to put the fulcrum.



ChatGPT in Durrell’s voice

Prompt: Sonnet written in the style of Lawrence Durrell

Poetry is the music of the soul
And Lawrence Durrell was a master of this art
He could make words sing and dance
And touch the heart

He had a way with words
That was both poetic and sublime
And his verses will live on
Long after he is gone

He was a true wordsmith
And his poems will continue to inspire
For they are the music of the soul
And will touch the heart

I’ll end with with this failed sonnet, written allegedly in the style of Durrell. Juxtaposed beside an earlier example that seems more successful, it clears a space for asking more questions—which seems the best way to end a talk on a topic about which we’re still seeking answers.