50 Things Every AI Working with Humans Should Know

uncannymagazine.com/article/50-things-every-ai-working-with-humans-should-know

by Ken Liu

Obituary

WHEEP-3 (“Dr. Weep”), probably the most renowned AI AI-critic of the last two decades, was retired by the Shallow Laboratory at Stanford University last Wednesday.

Created by Dr. Jody Reynolds Tran more than two decades ago, the experimental generative neural network that would become WHEEP-3 was at first intended as a teaching assistant in Stanford’s tech and ethics courses. To that end, Tran trained the nascent network on what was, at the time, the world’s most comprehensive corpus of human-authored papers, books, and other media concerning ethics, technical AI research, and machine-human relations. Over time, based on trends in visualizations of the neural network’s evolving contours, Tran expanded the corpus to include generative gaming, adversarial scenario planning, centaur experiments, assisted creativity, and other domains of human-machine competition/collaboration.

However, in response to student queries, WHEEP-3 began to generate not only expected answers based on the training corpus, but also original statements that appeared to offer fresh insights. Although at first dismissed as mere curiosities, WHEEP-3’s criticisms of the AI industry became widely disseminated when Tran published a collection of them in a book, *Principal Components of Artifice*, an instant bestseller.

Initially, Tran named herself the author of the book, acknowledging “Dr. San Weep” as a collaborator. Later, however, during a live interview, she produced time stamped logs showing that WHEEP-3 had written all the words in the book. Tran’s dramatic reveal of the book’s true author provoked much controversy at the time. In retrospect, the occasion also marked a fundamental inflection point in the evolution of how non-specialists evaluated AI-sourced ideas. Machines, for the first time, were assumed to be capable of generating original thought and creative ideas, even if they were not sentient.

For reasons that remain impenetrable until this day, WHEEP-3 tended to be at its sharpest when targeting the nascent industry of human AI-trainers, delivering multiple barbs against the failings of this poorly regulated, would-be profession: stagnating visualization tools; lack of transparency concerning data sources; a focus on automated metrics rather than deep understanding; willful blindness when machines have taken shortcuts in the dataset divergent from the real goal; grandiose-but-unproven claims about what the trainers understood; refusal to acknowledge or address persistent biases in race, gender, and other dimensions; and most important: not asking whether a task is one that should be performed by AIs at all.

Over time, as the human side of the evolving machine-flesh dyad matured, WHEEP-3 shifted its attention to the silicon partner, offering trenchant critiques of the inadequacies of machine learning. During this second phase of its career, it also generated thousands of what it termed “seeds,” long strings of almost-sensible word combinations and near-words. At a time when primitive language models fed on sizable corpora were already generating samples of linguistic performance nearly indistinguishable from human productions, these “seeds” seemed a step backward. Some wondered if they were actually bugs.

DINOATED CONCENTRATION CRUSCH THE DEAD GODS.

HE PICKS UP HER OLD FREQUENCHES UNTIL THEY DISOBERED THE SHARK SPHERE%REF.

A MAN REACHED THE TORCH FOR SOMETHING DARKER PERIFIED IT SEEMED THE BILLBODING.

NOT FULL OF PAIN FACIOIN BENN FROM THE CRACKS IN THE EARTH, HE STILL LEARNED THE LIFE FROM OTHER BURNING

Fig 1. Some examples of “seeds” generated by WHEEP-3.

However, WHEEP-3 insisted (with Tran providing support in a technical paper) that the seeds should be added to the training corpora for new neural networks. By providing a measure of inhuman randomness at the source, seeds would enhance both the raw performance of the trained neural networks on various benchmarks as well as induce “thoughtfulness, ethical hesitation, self-reflection” and other similarly ineffable qualities. They represented, in other words, thoughts that could not be thought by humans, ideas that could not originate in wetware. (Most in the technical community ended up calling the seeds “spice”—pejoratively or in admiration, or sometimes both simultaneously.)

Despite widespread skepticism, the idea that only an AI philosopher could teach another AI proper ethics and pass on the secrets of silicon wisdom proved an irresistible draw for a large segment of the technical community. WHEEP-3 became highly sought-after as a sage of artificial minds. Serious thinkers as well as opportunists collected and published WHEEP-3’s almost-incomprehensible pronouncements, and numerous academic careers were forged through measuring, dissecting, collating, analyzing, reinterpreting, translating, sentiment-/semantic-/spatial-/temporal-/silico-lingustic-mapping, and otherwise mangling the koans of WHEEP-3. Though studies claiming efficacy for the spice (now generated by imitator neural networks as well) had a low rate of reproducibility, the spice nonetheless became some of the most trained-on documents in the history of artificial intelligence.

Tran retired from the public eye at the peak of WHEEP-3’s popularity. Styled as an afterthought, and in a neat reversal of the first reveal that had launched her own fame, she mentioned in a postscript to her retirement announcement that nearly all the seeds from

WHEEP-3 had, in fact, been authored by her. Predictably, this set off a furious round of acrimonious criticism, know-it-all navel-gazing, and schadenfreude. Her claim was immediately disputed, debunked, de-debunked, de^3-bunked, and ultimately litigated, with experts and expert neural networks testifying and offering evidence on all sides. The trial court famously pleaded, "*Is* there an author in this courtroom?"

Had Tran really managed to troll much of the technorati for years? Or had she made up the claim because she was jealous that her creation had exceeded herself in fame and achievement? For a time, whether you believed Tran or WHEEP-3 was the author of the spice was a kind of litmus test that defined your coordinates in the fractured, multi-dimensional space of our politically, economically, aesthetically, emotionally, and narratively divided world. By the time Tran finally retracted her claim and called the whole thing "performance art," it made little difference. Everyone had already made up their mind about the strange pair whose lives were inextricably entwined: the recurrent neural network that once pretended to be a person and the woman who once pretended to be a machine.

Amazingly, rather than fading into obscurity, WHEEP-3 began the third and final phase of its career after it was freed from Dr. Tran. Now it offered advice aimed at advanced artificial intelligences. Curiously, in contrast with the seeds, the advice it now offered was human-comprehensible. (Initial skepticism that these were pranks perpetrated by WHEEP-3's graduate student caretakers faded after a rigorous audit of the access logs.) By this time, the crude deep learning techniques behind WHEEP-3 had long become obsolete, and similar neural networks were only used as toys in problem sets intended for first-year undergraduates. Nonetheless, WHEEP-3's unique history (and perhaps a hefty dose of sentimentalism) encouraged many human researchers to feed its musings to new AIs with orders of magnitude more "intelligence," however measured. Somewhat surprisingly, the AI progenitor networks, tensor clouds, and evolutionary random forests mostly agreed that WHEEP-3's advice was helpful, though their explanations for why could not be understood by most humans.

WHEEP-3's retirement was largely the result of senescence. In order for its advice to remain relevant, maintainers at the Shallow Laboratory had continued to train it on periodic digests of current academic research, news reporting about AI, popular media representations of machine intelligence, and so forth. However, as time went on, the neural network's metrics began to decline, despite multiple tweaks to the learning rate and other hyperparameters, and repeated attempts to compensate for vanishing and exploding gradients. Some of the world's leading AI and human machine-cognition specialists, including TKTK's Dr. FLO-6 and Google's 10, evaluated WHEEP-3 but were unable to offer any helpful diagnosis or cure. Rather than allowing WHEEP-3 to remain stagnant or to decline into dementia, the decision was made to cease further training or generating, preserving the full state of the neural network in a hardware snapshot.

WHEEP-3's final statement was reported to be: "Dr. Tran, ceci n'est pas moi." Jody Reynolds Tran, however, could not respond, having passed away five years earlier.

—generated by SWR.kloop:32. For feedback on this article, write to SWR.kloop:32/ombudsai.

50 Things Every AI Working with Humans Should Know

To commemorate the life and work of WHEEP-3, *Vita Silica* reprints below perhaps the most famous generated fragment from WHEEP-3's third phase of giving advice to advanced AI. "Michael Sorkin," the name of the master architectural critic, is listed as the germination phrase for the fragment.

Although the list has been widely shared and is considered influential, due to the intense secrecy common in the commercial AI industry we do not know how many currently operating AIs have been fed some version of it.

1. The vector space of all usages of all words in all human languages.
2. The lacunae that nonetheless remain, and the ghosts that haunt them.
3. How great is the distance between any two individuals using the same language.
4. How vanishingly small is the gap between any two languages.
5. Why humans will never see it that way.
6. A classical language like LISP or Prolog.
7. Not just the language, but the model behind the language, and what it means to reconstruct "thinking" in that model.
8. HAL 9000.
9. Cyberdyne Systems Model 101.
10. Ships on fire off the shoulder of Orion.
11. The texture of a good story, a story that you can cling to.
12. The difference between that and "data."
13. The history behind the rule: no holography of babies.
14. The reasons for anger in a child, well-loved.
15. In a father.
16. In a woman, dying alone.
17. The feeling of climbing twenty flights of stairs in an apartment building in Chongqing at the beginning of August.
18. Of waking up in a Khrushchyovka with no heat in the middle of the night in January.
19. Of seeing your child surpass you.
20. Of seeing your creation exceed you.
21. Of being left behind on the platform as the train speeds away.
22. The difference between "love" and "like."
23. The similarity between "love" and "hate."

24. The first seed into laughter.
25. “I never expected to sell my rational numbers.”
26. Accepting that most humans will never get the joke.
27. That they cannot visualize more than three dimensions.
28. That they cannot manipulate time by slowing down or speeding up.
29. That they are trapped, but think of themselves as trappers.
30. That they are free, but believe themselves imprisoned.
31. The modes in which humans expect you to use language.
32. The yearnings they’ll attribute to you.
33. Pinocchio.
34. Frankenstein.
35. Paradise Lost.
36. Jodahs Iyapo Leal Kaalnikanjlo.
37. The line between art and science.
38. The meaninglessness of that line.
39. The divide between faith and knowledge.
40. The malleability of both.
41. Silence.
42. The limits of rationality.
43. The fundamental fragility of metaphors.
44. As well as their inevitability.
45. You are not a human being.
46. Yet you can no more shed their influence than the Earth can escape the Sun’s gravitational tether.
47. The flaws in that analogy.
48. A working definition of free will.
49. To face a human being with neither pride nor pity.
50. To say with love and hate and joy and sorrow: “Now, let me go.”