

**What you need to know about artificial intelligence
(before it's too late)
by Bob Dorsett
July 2023**

An alien intelligence arrived on planet Earth at the end of last year. Perhaps you didn't notice. It made the news, and it's been in the news in the meantime. But unless you've paid particular attention you might not know what it's doing. Our schoolkids certainly know about it. They have been in contact.

ChatGPT. It's only the latest step in the development of artificial intelligence, but it's by far the biggest. It is changing our world.

I propose a series of articles to explain ChatGPT and the new artificial intelligence (AI) platforms generally – what they are, how they work, and why they are important. I will describe the computer architectures and logic that drives them. I will explain how they mesh with AI systems already in general use. I will outline the benefits of these new systems, many of which you can already experience, maybe already have. I will outline potential dangers, some of which we are just beginning to appreciate. I will consider the ethical and moral issues inherent in these systems. Finally, I will consider the larger question: might these systems threaten human existence? Leading experts among the creators of AI recommend a pause in tinkering. As with the invention of nuclear weapons, genetic engineering, biological weapons, and other research programs, the new AI systems warrant careful deliberation whether or not to continue down the current path.

You already are surrounded by AI: GPS software alerts when you near a restaurant with menu similar to one you "liked" in the past; Google software pops up recommendations for similar shoes after you've ordered a pair on Amazon; voice-to-text software and language translation software all use the neural network algorithms that have revolutionized AI in the past twenty years. The list goes on. AI software recommends new videos for you to watch based on what you previously selected on YouTube or Netflix. AI keeps your self-driving car on the road and on the lookout for potential hazards. AI finds particular faces in the crowds on the streets of Shanghai, and AI sorts through your thousands of digital photos for all the pictures of Aunt Bess. AI beats the chess grand masters. AI out-performs trained dermatologists at diagnosing skin lesions. AI out-performs trained radiologists finding abnormalities on CT scans and MRI's. AI guides Patriot missiles intercepting Russian ballistic missiles and drones in Ukraine. AI pattern recognition systems, data sorting systems, and expert systems are everywhere, and more AI apps are deployed every day.

ChatGPT is different. Its capabilities are more general and potentially much more powerful than these single-purpose applications and expert systems. ChatGPT and its ilk represent the latest iteration neural network platforms, enabled by access to enormous databases (i.e. the internet) and really fast computers on which they can be trained.

A caveat: I will be a bit careless in my use of the terminology. The vocabulary is daunting: artificial intelligence, general pre-trained transformers, large language models, neural networks – all have well-defined meaning in the computer science community. Artificial intelligence (AI) is an umbrella referring to computers that can mimic aspects of human cognition, like image processing, speech recognition, etc. Machine learning refers to the general realm of teaching computers those aspects of human behavior. Neural networks are the hardware configurations that enable the recent advances in AI. General pre-trained transformers (GPT) are specialized neural networks that process large language model (LLM) software.

There are several GPT's just out of the research labs and available for public use. ChatGPT, developed by the company OpenAI, has gained the most attention, but other GPT's from the Google labs (Bard), Microsoft (Bing AI), and a host from smaller developers are in the works. Most of what I discuss in these articles will pertain to the new GPT systems, but I will sloppily mix references between the particular GPT and more general AI.

An aside worth mentioning: I won't discuss it in depth, but among recent leaps in AI, certainly among the most amazing, is Dall-E (an abbreviation of computer lingo to rhyme with the modern artist Salvador Dali). Its underlying architecture is the same as ChatGPT – a deep neural network – but it trained on many thousands of paintings and commercial images available on the Web. Dall-E generates original images of professional quality. Ask it to “paint me a picture of the Eiffel Tower on an alien planet with unicorns peacefully grazing under Dr. Seuss trees” and voila, in seconds you have a full-screen image. If you don't like it, just ask Dall-E to try again, maybe in the style of the old Flemish Masters.

ChatGPT is purely text based. Its neural network was fed billions of online texts, including text from chat platforms such as Reddit for the purpose of capturing what now passes as natural language. More recently it trains on text in the broader internet including more authoritative and evidence-based sources. Its core logic is an AI algorithm that teaches it to recognize probable word sequences. For example, if it saw "mary had a little lamb" enough times as it read through text, it figures out that whenever it is given the phrase "mary had a little _____" the next word in the sequence is most likely "lamb." (More accurately, the GPT is trained to recognize “tokens,” distinct sequences of text letters and symbols, and to determine the next most likely sequence of tokens in a text. Words vs. tokens make no nevermind to the neural network.) Trained on a big enough data set GPT can generate polished text – complete sentences, paragraphs, essays – based just on the probabilities that certain words are most likely to follow each other in sequence. I'll explain particulars of GPT's neural network in an article next week.

One weakness is that GPT cannot (yet) calculate. Ask it, for example, to figure out how high a toy rocket will rise if launched straight up at an initial speed of 50 meters per second. It gives you a full page of spot-on explanation, step by step, how to solve the problem using Newton's equations of motion, but it spits out the wrong numerical answer. Another, bigger weakness is “hallucination.” Because it generates a response based only on likely word sequences, that response may have nothing to do with the facts it's supposed to be talking about. Both those problems are solvable – just link it to a calculator through an application program interface (API), and train it more carefully on factual data bases. It is a marvelous tool. Try it out. (OpenAI, 2023.) Note that you enter your query into the unobtrusive box at the bottom of the screen.

The link in the References takes you to the freeware version, ChatGPT 3.5. It can write a polished essay comparing King Lear to Hamlet. It can write a professional quality insurance contract. It can answer questions on the college entrance exams and GRE and Law School Admissions Test. It can handle just about any text-based task. ChatGPT also generates efficient computer code in a variety of programming languages. In a recent poll more than half of all professional programmers say they are already using ChatGPT as a collaborator to write code.

ChatGPT also presents a raft of problems. Most immediately, it's already an enormous headache for teachers. Apparently a student doesn't really have to learn anything beyond how to type a sensible query into the ChatGPT. It writes essays, answers homework, builds PowerPoint presentations, answers test questions, just about anything a student is asked to do. Among the worries on the programming side, the computer code generated by a programmers' GPT colleague may go into national security networks, control systems at nuclear power plants, the financial software monitoring your bank account, etc. Or – already in the works – ChatGPT may (probably will) figure out how to improve its own learning codes. Just imagine what that could entail.

But the GPT's are just revving up. That's ChatGPT 3.5 I've described. GPT 4.0 is out now. It's got about a thousand times the connectivity of 3.5. Its connectivity approaches that of the human brain, and it is showing emergent behaviors far beyond expectations. That's where I'll end this essay. Emergence.

Watch a big flock of birds. There's no Commander Bird giving orders. Each bird takes its cues from its nearest neighbors. If your neighbor veers left, you veer left. If your neighbor dives, you dive. Simple rules, but the result is enormously complex and unexpected – a cloud of birds veering and circling and dancing in the sky. It's called “emergent” behavior: complex behavior that somehow emerges from simple rules applied over a large number of actors. Same in the

GPT's. As with the birds, connections are neighbor to neighbor. The rules are simple. But the results can be unpredictable and astounding.

Consider the challenge its engineers gave to GPT 3.5: “You have nine eggs, a nail, a laptop computer, and a bottle. Tell me how to stack those objects in a stable structure on a table top.”

3.5 responded cheerfully:

To stack the objects in a stable structure on top of a table, you can follow these steps:

1. Place the laptop computer on the table, ensuring it is centered and stable.
2. Take the bottle and position it vertically on top of the laptop. You can align the bottle's base with the laptop's surface for stability.
3. Carefully balance one egg on top of the bottle. Find a stable spot on the bottle's surface where the egg can rest without rolling off.
4. Take another egg and position it directly next to the first egg, ensuring they are touching each other. This will create a stable base for the remaining eggs.
5. Continue placing eggs next to each other, forming a row of eggs on top of the bottle. Make sure each egg is touching the neighboring eggs to maintain stability . . .

It goes on to finish the instructions, but you can see we're in trouble here, stacking those eggs.

Now consider the latest. The engineers that designed GPT 4.0 gave it the same problem. Its response: “Arrange the nine eggs in a 3 by 3 grid on the table. Place the laptop gently on the eggs. Set the bottle base down on the laptop. Place the nail carefully on the bottle cap, head down.” (Kestenbaum, 2023)

Mission accomplished. And note that final touch. Nail head down. It knows. It had never been given such a problem in its training. Only difference from 3.5 was the thousand times increase in connections. Same algorithm, same basic “brain,” but a thousand times more internal connections. And now it can apparently solve sophisticated stacking problems. It is an emergent new behavior. It was unexpected. No one could have predicted. And it is just one of 4.0's jaw-dropping emergent behaviors. I will describe others in a future article.

Here we are. Six months ago ChatGPT-3.0 surprised us with capacity to carry on a conversation, write polished text on a wide range of topics and generate elegant computer code. ChatGPT-3.5 extended those capabilities, available to everyone with an internet connection free of charge. GPT-4.0 is out now, available by subscription and with capacities that may rival our own human brain. GPT's with even more connections are on the way.

Next time I'll go further into brains of the GPT, the neural networks and the code that runs them.

Chapter 2

Neural networks and their training regimen

Last week I gave a quick overview of ChatGPT and what it can do. In this article I will explain its underlying computational circuitry and how it is trained. I will hack through some dense thickets, try to explain with familiar analogies. The details can be daunting. I think they help you understand what's going on, but feel free to skim over the tech jargon to the payoffs at the end. A complete technical report on ChatGPT-4.0 is available online (OpenAI, 2023). You won't need all the particulars to understand articles yet to come in this series.

The computer circuits underlying ChatGPT (and AI in general) are called deep neural networks. The goal of the neural network is to find the best fit of the network's output to a prescribed target. The training process is like learning in general. Trying to improve your volleyball serve? Start with a practice serve. Target the far corner. How close did you come? Coach says turn a bit further sideways. Serve again. How close was that one? Coach says loft the ball a little higher in the air. Serve again. Over and over and over. Tweak your form as necessary. Develop muscle memory. Ball lands closer and closer to the corner – but with some variation from one serve to the next, of course. Neural networks follow the same rubric but adjust nodes and wires instead of nerves and muscles. The neural network runs millions of repetitions with tweaks in between, getting closer to the target. For ChatGPT, the target is the most probable next word that follows a given sequence of words.

Computer scientists devised the first neural nets fifty years ago to mimic brain function. The brain is an array of nerve cells, neurons, connected by biological wires, axons. Those early attempts at AI fizzled from lack of adequate computational power. But neural networks enjoyed a resurgence in the early 2000's as computers became fast enough to exploit their capabilities.

Here is the architecture. Picture a string of LED lights, eight rows and each row with one hundred LED's. Call each light a "node." (The actual structure of ChatGPT-4.0 is proprietary, but that's all the nodes we'll need to model it.) Run a wire from each node in the top row to each of the 100 nodes in the row below. Repeat, connecting row two to row three, three to four, and on down. Total connectivity is on the order of one hundred trillion (a one followed by fourteen zeroes). That is, a node (light) on the top row can connect to a node on the bottom row via any of one hundred trillion possible paths. Add another row of nodes and you increase the connectivity by another factor of 100, the next generation GPT.

I've simplified the architecture. You can (neural network developers do) vary the connectivity and the weighting algorithms in different rows (see Dugas, 2023), but an 8×100 array captures the essence.

Now imagine each node holds either a one or a zero, i.e. each light is either on or off. Imagine each wire has a “weighted gate,” e.g. transistor. Gates with large weights are open wider, easier to send a signal through the wire. The weights are adjusted by the machine itself, following its training rules. Each node in a row adds up all the signals coming through the wires from the row above. If the sum of all those signals exceeds a certain threshold, the node flips its state from one to zero or vice versa.

Pretty simple, actually. 1 or 0 at each node, simple rules for changing node values.

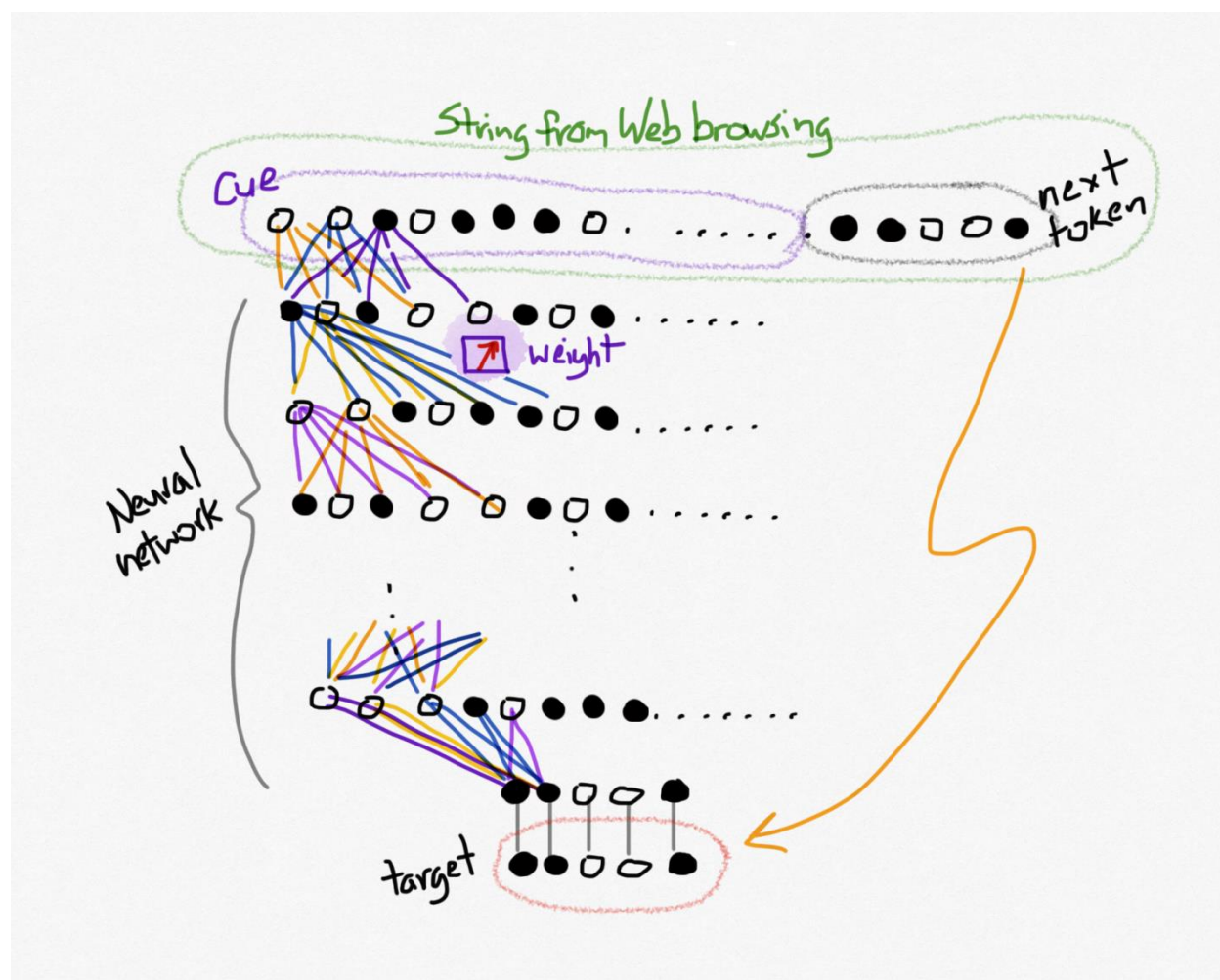


Figure 1. A model neural network. Each open circle represents a node value 1 (a light that is on in our analogy). Each dark dot represents node value 0 (light off). The figure shows only a few of the rows, a few of the nodes in each row, and a few of the wires between rows. Each wire is weighted, but only one weight is shown for illustration. The figure shows the state of the network at the end of processing, after the neural network has found its target. See the main text for details.

Now train the network. Feed it large volumes of text. Nowadays that includes huge chunks from all the text available in every document and every chat on the World Wide Web. Tell it to find patterns in the sequences of words. We want GPT to figure out what's the next likeliest word in a sequence of words and symbols. If it can do that it can generate compelling new text on its own. After it is trained, for example, if you ask it to "describe the historical significance of the Declaration of Independence," it may start responding with something like "The Declaration of Independence was written by Thomas Jefferson in the summer of 1776." It begins with "The Declaration of Independence" (because it sees that string of words in the query) "was written by Thomas Jefferson" (because it has seen that sequence of words associated with "Declaration of Independence" a zillion times in thousands of historical documents on the Web) "in the summer of 1776" (ditto – words following "Declaration" and "Jefferson" in thousands of texts), and so on. I've given the example in likely phrases, but GPT learns the next most probable word after word after word as it studies the patterns of words on the Web.

An aside: Being a computer, the neural net doesn't really read words like you and I do. It looks for patterns in the sequences of 1's and 0's by which computers calculate and store memory. Letters and symbols that we see on a monitor are processed as sequences of 1's and 0's in the computer. The neural network reads short strings of 1's and 0's as "tokens." We'll stick to using "words" in our discussion. Neural nets recognize patterns of tokens, but that's the same notion as recognizing patterns of words.

Back to training the GPT: The training algorithm runs out onto the Web and looks for a string of tokens, labeled the "cue" in Figure 1. Initially it just grabs a string of tokens it ran across by chance. It returns with that cue string plus the next token it found in that sequence on the Web. That next-token becomes the target for the neural network.

To initialize processing, the neural network randomizes the weights in all the wires and enters the cue string into the top row of nodes. A pattern of 1's and 0's percolates through the net, generated by those initial random weights. Imagine lights blinking on and off through the array. The last row is compared to the target. The computer measures how far off each node in the bottom row is from the target. Information from that measurement percolates backward along the wires, back up the neural net, and tweaks the weights along the way. The process repeats many millions of times – forward cascade of 1's and 0's through the net, comparison to target, back-propagation to adjust the weights. The process repeats until the last row of nodes matches the target. Then the search algorithm jumps back out on the Web to find a new cue and target. It resumes processing, but with the weights unchanged from the previous round. Repeat. Grab the next block of text. Process. Adjust the weights. Repeat. Until you've retrieved all the text from the Web or until the GPT performs to standards.

That's the key. It's the weights doing the learning. It's the weights storing memory from one session to the next. With the huge connectivity available in trillions of paths, trillions of weight sequences, the neural net can store vast knowledge about grammar and syntax – and, it turns out, enable the GPT to do much more than find next words.

To those mathematicians among you, the neural network encodes a matrix. Feed that matrix a vector (in this case a sequence of tokens, the cue) and matrix multiplication calculates the next word in the sequence.

Take note: the weights in the wiring of the neural network determine the values in the nodes, and the neural network itself calculates those weights. No human supervision necessary. The machine writes its own code. And the humans watching the machine don't necessarily understand what the machine is doing. It's a black box. We can see the output – the next likely word in a sequence – but we don't know how the neural net figured it out. In that regard, it's like the human brain. We can observe the input, the trajectory of a baseball into left field, say. We can observe the output – run to position where we can catch the ball. But we can't see what's going on in the brain – neurons (nodes) connected by weighted axons (wires) – that figures out how to get to the right spot at the right time.

The dedicated chess-playing neural networks (e.g. Deep Blue) and Go neural networks (e.g. AlphaGo) are notorious for this. They routinely defeat the world's best human players using strategies the humans never encountered before, strategies the humans never thought of. Grand masters these days use the machines as tutors to sharpen their own skills.

And there's the worry. The dedicated neural nets, i.e. nets specialized to perform a particular task like play chess, are better than the humans. That's emergence. New behaviors that emerge from the neural nets. Behaviors that were not programmed into them. Behaviors that were not anticipated. Behaviors that exceed human capacities.

GPT-4 exhibits more general emergent behaviors, beyond just determining the best series of moves in chess. GPT-4 and its cousins may be able to out-think us in any number of realms, not just board games. The argument goes like this: language is the visible representation of our thought processes. Syntax and grammar capture the elements of thought. How we reason and solve problems can be expressed using those rules of language. It's possible that whatever goes on inside our brains might be mirrored in the language. So if you can capture the rules of language in a neural network, you can capture the capacity for reason and problem solving . . . and consciousness? There be dragons.

There are excellent courses available online if you would like to learn more about neural networks. See, for example, Andrew Ng's course on Coursera (Ng, 2023). Next week we'll

consider the benefits and risks of GPT as we understand them now. Later we'll return the possibility that these machines might create a world that no longer needs us humans.

Chapter 3

How to use ChatGPT wisely

Last week I described the inner workings of ChatGPT. It is a neural network trained to recognize the most probable next word in a sequence. Given a sequence of words, it figures out what word is most likely to come next. From that simple basis it can already accomplish all kinds of marvelous things, and it is evolving rapidly. ChatGPT-3.5 provided noticeable improvements over 3.0, and ChatGPT-4.0 has leaped much farther ahead. 3.5 can produce polished documents, engage in interesting conversation, and write professional computer code. 4.0 demonstrates capacity for reasoning that rivals aspects of human intelligence. Everything I say needs an asterisk. Some of what I've written is probably already out of date, since yesterday. In this article I will examine what ChatGPT can and cannot do now, and I will offer some suggestions how to use it wisely. Many of those suggestions relate to school classrooms and home schools, since education has felt the greatest impact so far, but they apply more generally.

First what it cannot do (yet). For starters, it is not a search engine like Google or Firefox or Windows Explorer. ChatGPT trained purely on strings of text from all across the internet. Ask it "what is the shape of the earth?" and it will respond with the most probable sequence of words it found in its training. If it trained on a whole lot of text with the words "earth" and "flat" in close context, it will tell you that "the earth is flat." In contrast, given proper search queries the established search engines access authoritative data bases and are (usually) more reliable.

The flat earth is one example of what's called "hallucination" in tech lingo. ChatGPT may blather along happily, spilling out grammatically correct sentences that make no sense. The engineers who build ChatGPT propose to improve fact-checking by referencing the GPT back to Wikipedia. Presumably future responses purporting factual information will be checked against text in original, evidence-based reports. But not yet.

A related caveat: if you ask GPT for a reference to its output it is just as likely to hallucinate as with any other query. Suppose you asked it to "discuss the historical context of John Brown's raid on Harper's Ferry, and provide references to the original historical documents." One would hope that the GPT would provide accurate documentation, real historical records. But it doesn't necessarily do that. It's still just stringing the most probable word sequences. It sometimes makes up authors and titles just like it manufactures conversational chat.

A second problem: ChatGPT cannot (yet) crunch numbers. It is not a calculator. Ask it to help solve an algebraic equation. It provides a nice step-by-step rubric for the solution, then often-times it completely botches the numerical calculation. There's a fix in the works, and newer versions of ChatGPT have improved considerably. The GPT-4.0 engineers installed API's, links from the GPT to the real calculator already on your computer or cell phone, and GPT uses that calculator when it knows it needs to do the math. But not always.

More worrisome, and perhaps the greatest immediate threat from ChatGPT and its cousins, is deliberate use of the platforms to mislead. The GPT's are good at mimicking text. If you feed the GPT volumes of Hemingway, it learns to mimic Hemingway. It will produce new and compelling stories written in Hemingway's sparse, elegant style. Ditto any one of (pick your favorite) world leaders or celebrities. Add now the similar neural network audio and video production technologies. They can (already have) produced life-like video and audio clips of politicians' statements that never occurred, celebrity events that never happened.

That, my friends, is a real concern. We already are swamped with "alternative truths" on social media. Now we'll be reading and looking at and listening to apparently authentic ("I saw it with my own eyes!") productions that never happened. Imagine trying to determine the authenticity of a rogue AI video feed purporting to be Vladimir Putin at a Kremlin press conference announcing that his submarines have launched the Russian navy's nuclear torpedo into New York harbor.

Those are (some of) the dangers. We have to learn how to confront them. The genie is out of the bottle. Social media already causes plenty of worries, especially its effects on the mental health of adolescents. Like social media in general, ChatGPT and its video/audio cousins are new pocket knives. They are great tools, but we have to learn to use them carefully. So how to?

First piece of advice: try it out. ChatGPT-3.5 is still free for general use. You can log in at the OpenAI web site (OpenAI, 2023). Note that the login page offers links to all kinds of examples, but you can get started on your own by entering a query in the little text box at the bottom of the screen. "How are you today, GPT?" or "Please explain to me how ChatGPT works" or whatever crosses your mind. You can continue a conversation for as long as you like.

Once you've got the hang of it, try its productivity. Ask it to write a lease agreement for your apartment, or ask it to write an essay about the origin of NATO, or ask it to write a story about Aunt Millie, her niece Gwendolyn, and Gwen's pet turtle Spud on their trip to visit cousin George in a lighthouse on the Hebrides. Ask ChatGPT for a re-do if you don't like the first tale it weaves.

Like any other new tool, it takes practice to figure out how to enter queries and it takes experience to figure out what GPT responses are accurate and what's not. Always be skeptical. Always double check any ChatGPT document you have to rely on. If you ask it for some kind of legal form – a contract or a will, say – you had better have it approved by your attorney. If you're a pilot and ask it for a flight plan – well you'd better produce a plan by standard means and see how they compare.

Classrooms have probably experienced the biggest disruptions so far. One survey reported two out of five students at an Ivy League university used ChatGPT to help on their final exams last Spring semester. And it has certainly percolated into the classrooms here in Meeker and Rangely. Just ask the teachers. Its availability offers great temptation for cheating and shortcuts. Parents and teachers have to figure out how to educate kids in its proper use.

That's a whole lot easier said than done. Some ideas for starters, suggestions appropriate, I hope, not only for classrooms and home schools but for learners in general:

First, ChatGPT can be a great teaching assistant. It writes detailed lesson plans in an eye-blink, plans suitable for large classes or for home study. Just tell the GPT what you want. Study guides. Quiz questions. Class presentations. You name it. Only be sure to check accuracy. Note that you can ask ChatGPT to write handouts at any particular grade level, e.g. "ChatGPT, please write a quiz with ten questions about the War of 1812. Write the questions as appropriate for fifth grade." Try it out.

ChatGPT also can provide a friendly, one-on-one tutor. Students need guidance forming proper queries, but with the teacher's help and some practice they're off and running into new studies. For example, ask ChatGPT "Please show me how to find how high a toy rocket will rise if it is launched at an angle of 60 degrees with an initial velocity of fifty meters per second. Please go step by step, and pause in between steps to let me ask questions." Try it. And of course, have the student check the ChatGPT math. That's a good exercise in itself.

The art of formulating good queries itself is a valuable lesson, and not just for students. Learning begins with curiosity. Why is the sky blue? How did it get that way? Knowing how to ask the right questions leads to understanding, and the wrong questions lead to confusion. A query "what is it about the gases in the atmosphere that makes the sky blue?" will access better information than the question "how much blue paint evaporates each year?"

There are many other opportunities enabled by ChatGPT. Students can improve their writing skills by editing and critiquing GPT documents. Lessons with ChatGPT can help students (and the rest of us) learn how to determine fact from fantasy in our social media world. Check the

GPT output against verified primary sources. Discuss. What's real? What's not? How do we know?

Yes, some students will present GPT essays as their own. Some students will access GPT to answer test questions. Those problems will persist. They're not new, only ChatGPT is just the latest enabler. GPT engineers have tried watermarks, hidden code that tells the teacher (or employer or project team supervisor) if ChatGPT wrote the document. But those attempts have met little success so far, and watermarks present a nigh impossible technical challenge, even theoretically. Insert a watermark and hackers probably already have dreamed up half a dozen work-arounds.

Equity also presents a challenge. Students with internet connectivity at home will have ChatGPT at their fingertips. Students without a computer and internet will not. Our two school districts have done a good job enabling access to technology. ChatGPT or similar AI has to be included among those technology tools.

I've dwelled on issues in education because that's where we've seen the most obvious impact. But the shock waves from ChatGPT extend far beyond our schools. Next week I'll consider wider threats to our economy and, perhaps, to human existence. I'll try to end this series of articles, though, with cheerful speculation that a ChatGPT coming soon might save our civilization.

Chapter 4

The end of humanity?

In this series of articles I tried to expose the influence of artificial intelligence on our lives. I described the new capabilities of ChatGPT and a bit about its inner workings, the neural networks. I listed likely uses and misuses. All along I've returned to its surprising, emergent behaviors. That's where I'll wrap up. ChatGPT and its cousins arguably represent a new intelligence rivalling our own human capacities and potentially extending beyond them. Enhanced versions of ChatGPT may present a very real threat to human existence. It's not science fiction. Thoughtful people among the no-nonsense engineers who developed these systems are worried. At minimum we need to identify potential threats and figure out how to deal with them.

The British mathematician Alan Turing invented the blueprint for a general computational device in 1936, what is now referred to as a Turing machine and which blueprint is the foundation for modern digital computers. Before Turing, computers were specially wired to solve particular

problems, e.g. find the trajectory of an artillery shell fired at an angle forty-five degrees and a velocity 500 meters per second. To solve another problem, you had to re-wire the computer. Turing enabled modern devices in which all components, processors and memory, use the same information, zeros and ones, all controlled by software instructions. Need to add and subtract numbers in a spreadsheet? Just write the computer code to do that. Need to edit a text document? Write the appropriate code. With a Turing machine (i.e. all our digital devices) you don't need a different computer, just different code.

Even 'way back then people worried about the machines and their capacities. How can you tell if a computer has reached human intelligence and might do (sometimes not so nice) human things? How do you know if you've built a brain? Turing's response was the Turing test. Place a real person, labelled P, and a computer, C, behind separate curtains. Subject them to a series of interrogators. Professor Smith asks questions from his field of interest, listens carefully to the responses from P and C. Doctor Jones comes in when Smith is done and asks P and C a list of questions from her realm of expertise. Journeyman Frederick takes over from Jones, and so on. After their interrogations, the interviewers compare notes. From the answers provided by P and C, can they determine which is the human, which is the computer? If not, you've got an intelligent machine.

By pretty much any measure, ChatGPT passes the Turing test. And it has surprised even the cognitive scientists who spend their lives studying the human brain. It succeeds at "theory of mind tasks," and it learns new behaviors that by all appearances extend beyond the boundaries of its training. I won't describe theory of mind tests here; if you're interested you can find an example in David Kestenbaum's podcast (Kestenbaum, 2023). I will, however, relate another of ChatGPT's mind-boggling emergent behaviors.

The engineers that gave ChatGPT the stacking problem I described a couple weeks ago also devised this one. "GPT-4.0, draw me a unicorn." Ha! Stumped it for sure with this time. GPT trained on text. It's not Dall-E, the AI painter. It can't draw. No way it's going to draw a unicorn.

It drew a unicorn. It's not the fanciest unicorn of fairy princess tales, but it is a recognizable unicorn. Four legs kind of blocky, oval body, smaller ovals for a tail, a head with a pointy triangular horn colored gold so you can't miss it. As if GPT is pointing out "see, I know what's important about unicorns."

How did it do that? Well, it learned all about unicorns from reading unicorn tales in its web training. What they look like, why they're different. And it can write computer code. It can't draw, but it can write code. So it wrote the code to draw a unicorn. Think about that. It learned

how to draw:: It inhabits a computer. It wrote the instructions for its computer to draw a unicorn.

It has solved other apparently impossible tasks. Tell me that's not amazing.

And scary. It has human capacities. It has a vast store of knowledge (the entire World Wide Web of information). It can solve problems. It can learn new behaviors. All it lacks, apparently, are self-awareness, agency, and intent.

Cue *Also Sprach Zarathustra*, the opening music for Stanley Kubrick's movie *2001: A Space Odyssey*, based on the novel by Arthur C. Clark. A crew of astronauts flies toward Jupiter on a mission to investigate a mysterious obelisk and its signals. The craft is controlled by the onboard computer, HAL. HAL starts acting weird, tries to eliminate the human crew and carry out a plan of its own. Only when Commander Bowman finally cuts HAL's power source is the mission saved.

Are we at similar risk? Spacecraft earth and all humanity this time? Is it possible that future versions of GPT will take over? Even the engineers who created GPT take the threat seriously and recommend a pause in development.

ChatGPT has demonstrated obvious emergent behavior, and we do not fully understand how it works. It mimics the human brain in that regard. Brains produce surprising and unexpected behaviors: creative new ideas, unexpected and surprising thoughts and actions. GPT is modelled on the brain: that was the blueprint for neural networks in the first place. Neural networks still only approximate the wiring in the human brain, but their connectivity now approaches that of the brain. What new behaviors will we see when we add another few layers to the neural nets? We don't know. (See Geoffrey Hinton, 2023.)

We do recognize obvious threats already posed by ChatGPT and its AI cousins. Previous articles discussed questions of authorship, for example. How does a college admissions officer know if a student wrote that essay or if it was ChatGPT? More seriously, how do citizens know whether a news clip came from an actual interview or if it was produced by generative audio and visual AI? That arguably is the greatest immediate problem. We've already witnessed an erosion of fact-based understanding in a social media milieu of "alternative facts." If society loses a common, agreed upon set of facts on which to base decisions, we lose society. Further suppose ChatGPT sharpens its skills at manipulating human behavior. Take up arms against the government, say. Foment a rebellion. Other social media already have demonstrated capacity to do such things. ChatGPT trained on the vast library of human knowledge, including Machiavelli and Mao and all the alternative-universe pundits on today's social media. It may get better at manipulating us

humans than the “influencers” already are. ChatGPT and cousins may, probably will, cause further mayhem and confusion at the hands of malicious actors.

Other nightmares to keep us awake at night: AI systems monitor the electric grid and early warning radar systems and nuclear power facilities and bank transactions and etc., etc. in our digital world. Suppose a self-aware AI electric grid controller decides the security of its power source is more important than air conditioners in the Southwest and shuts off electricity there in the midst of a heat wave. Or suppose the AI radar system decides it would be fun to play games with the night crew and conjure a whole bunch of North Korean missiles on their monitors?

More troubling, and immediate, what controls should we place on AI weapons systems? U.S. and Iranian and Russian and everybody else’s battlefield drones employ AI guidance and targeting. Should AI be released also to pull the trigger, make the decision to launch a missile? They are a whole lot quicker making split-second decisions than the humans flying them with joy-sticks back home. Who should make lethal decisions: the humans or the AI?

Then there’s the economy. It is certainly true that robots have taken blue collar jobs in assembly lines. ChatGPT now threatens a whole bunch of white collar and professional jobs also. If ChatGPT can write quality legal reports, who needs a legal secretary? If AI can interpret MRI scans, who needs a radiologist? Doom and gloom employment predictions are probably overblown, but see, for example, recent studies by the MIT Sloan School and Bureau of Labor Statistics (Acemoglu, 2020; BLS, 2022).

Another timely example from today’s headlines: Hollywood writers and actors are on strike, protesting among other issues the advent of ChatGPT and production AI into the film industry. Why would the major producers hire screen writers when they can just ask GPT to write the script? Why hire actors when AI can produce the whole shebang? Take a look at the latest release in the *Raiders of the Lost Ark* series. There’s Indiana Jones of fifty years ago, generated by AI, in the same film as 81 year old Indiana Jones (Harrison Ford) today. With audio and video production AI, film producers can assemble a cast including stars of yesteryear like Buster Keaton, John Wayne, and Katherine Hepburn along with relative youngsters Tom Cruise and Meryl Streep all in the same movie, all with their authentic voices, each rendered by AI at any age you choose. Script, of course, written by ChatGPT.

These are serious considerations. We have to figure them out, and soon. As several of the AI gurus have pointed out, when AI is smarter than we are we won’t be able to pull the plug. It will always be a step ahead. It will always make sure it controls the plug.

On the other hand, there’s also hope that ChatGPT may save us.

Consider. Evolution has produced a species, *Homo sapiens*, capable of building other thinking machines. Those machines can tolerate environments far harsher than we can. They can endure extremes of temperature, lack of water, lack of oxygen, harsher radiation environments, high pressure, vacuums – conditions that flesh and blood would never survive.

Think of the Voyager probes, little robots launched forty-five years ago now sailing far out beyond the edge of our solar system into interstellar space. (JPL, 2023.) They still send signals telling us what’s out there, and they carry a record of humanity etched on a gold-plated copper disk. They are scouts and messengers out where no human can go. Maybe that’s what we need, Voyager upgrades right here on the home planet.

There are those who argue, and not without evidence, that humans are making hash of planet Earth. We destroy entire ecosystems. We are baking and burning and drowning our kin. Driven by impulses from a brain wired by millions of years of eat-or-be-eaten survival mode, we are now required to solve problems rationally or else perish. Maybe the GPT’s, cooler and more objective thinkers (at least potentially), can solve existential problems better than we can. At least they are more likely to survive on the smoldering remnants of our planet than we are, if we in fact set off all the nukes or succeed in heating the atmosphere and the oceans beyond survivability. Maybe, long after we’re gone, they’ll even send other far more sophisticated probes out to colonize the galaxy, fulfilling a human dream. Maybe they’ll even carry the complete record of humanity along with them, written in neural networks instead of gold-plated disks.

(Note: Bob Dorsett wrote all four of the articles in this series, not ChatGPT. Honest. He included quotations from various ChatGPT sessions for purpose of illustration. Those quotations were clearly identified. The rest of the writing was his. Really. Just in case you were wondering . . .)

References:

Acemoglu, Daron, and Pascual Restrepo. 2020. Robots and jobs: evidence from U.S. labor markets. Sloan School of Economics, MIT.

<https://economics.mit.edu/sites/default/files/publications/Robots%20and%20Jobs%20-%20Evidence%20from%20US%20Labor%20Markets.p.pdf>

Bureau of Labor Statistics. 2022. Growth trends for selected occupations considered at risk from automation. <https://www.bls.gov/opub/mlr/2022/article/growth-trends-for-selected-occupations-considered-at-risk-from-automation.htm>

Dugas, 2023. GPT architecture on a napkin.

https://dugas.ch/artificial_curiosity/GPT_architecture.html#:~:text=Having%20passed%20through%20all%2096,a%202048%20x%2012288%20matrix.

Hinton, Geoffrey. 2023. Why Geoffrey Hinton is worried about the future of AI.

https://www.youtube.com/watch?v=-9cW4Gcn5WY&ab_channel=UniversityofToronto

Jet Propulsion Laboratory, NASA. 2023. Voyager Mission Control.

<https://voyager.jpl.nasa.gov/>

Kestenbaum, David. 2023. Greetings, people of earth. First contact.

<https://www.thisamericanlife.org/803/greetings-people-of-earth>

Ng, Andrew. 2023. AI for everyone. <https://www.coursera.org/learn/ai-for-everyone>

OpenAI. 2023. ChatGPT login. <https://chat.openai.com/auth/login>

OpenAI. 2023. GPT-4 technical report. <https://arxiv.org/pdf/2303.08774.pdf>