Commentary

Why Today's Humanoids Won't Learn Dexterity

Rodney Brooks, Robust.AI*

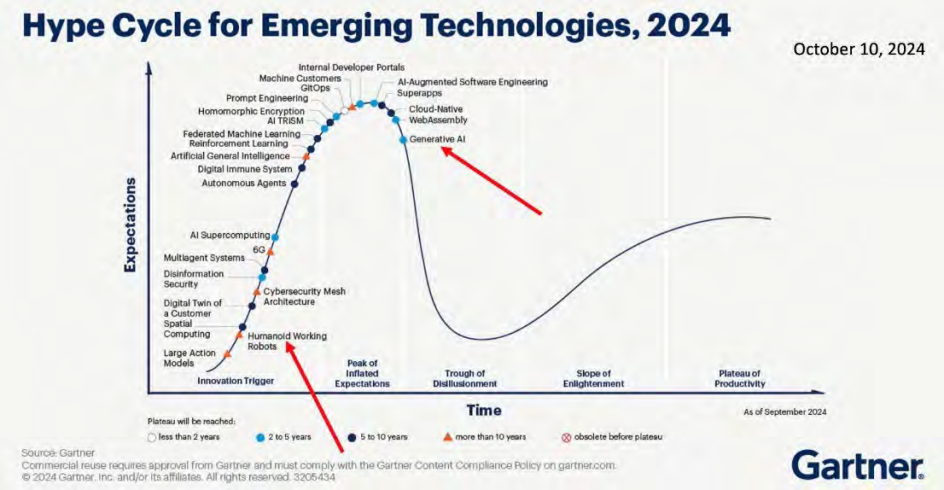
Abstract. In this post I explain why today's humanoid robots will not learn how to be dexterous despite the hundreds of millions, or perhaps many billions of dollars, being donated by VCs and major tech companies to pay for their training. At the end of the post, after I have completed my argument on this point, I have included two more short pieces. The first is on the problems still to be solved for two legged humanoid robots to be safe for humans to be near them when they walk. The second is how we will have plenty of humanoid robots fifteen years from now, but they will look like neither today's humanoid robots nor humans. [[My side commentaries look like this.]]

1. PROLOG

Artificial Intelligence researchers have been trying to get robot arms and hands to carry out manipulation of objects for over 65 years; since just a few years after the term Artificial Intelligence first appeared in a proposal for a 1956 "Dartmouth Summer Research Project on Artificial Intelligence". By 1961 Heinrich Ernst had produced a PhD thesis describing a computer controlled arm and hand that he had connected to the TX-0 computer at MIT, and had it picking up blocks and stacking them, and stunningly there is a video. His advisor was Claude Shannon, and he also thanked Marvin Minsky for his guidance, thus naming two of the four authors of the Dartmouth AI proposal.

This led to industrial robots, which were and are computer controlled arms with various "end effectors", think primitive hands, that have been used in factories around the world for sixty years.

Recently a new generation has stumbled upon the idea of building humanoid robots and you may have noticed just a little bit of hype about it. Gartner says it is early days and we are nowhere near maximum hype yet. This diagram is just a year old, and humanoids are at the very beginning of the cycle, while generative AI is over the hump and heading down to the doldrums:



Reprinted with permission from *rodneybrooks.com/why-todays-humanoids-wont-learn-dexterity/*, September 26, 2025. *Correspondence to the author may be posted there.

The idea is that humanoid robots will share the same body plan as humans, and will work like humans in our built for human environment. This belief requires that instead of building different special purpose robots we will have humanoid robots that do everything humans can do. For example the CEO of Figure, a humanoid robot company, says that:

> *We could have either millions of different types of robots serving unique tasks or one humanoid robot with a general interface, serving millions of tasks.*

Here is the first phase of his "master plan":

> BUILD A FEATURE-COMPLETE ELECTROMECHANICAL HUMANOID.
> PERFORM HUMAN-LIKE MANIPULATION.
> INTEGRATE HUMANOIDS INTO THE LABOR FORCE.

And during this just-ended summer, talking about their humanoids named Optimus, Tesla's CEO said that:

> *Optimus could generate $30 trillion in revenue and called humanoids "probably the world's biggest product".*

For both these companies and probably several others the general plan is that humanoid robots will be "plug compatible" with humans and be able to step in and do the manual things that humans do at lower prices and just as well. In my opinion, believing that this will happen any time within decades is pure fantasy thinking. But many are predicting that it will happen in as soon as two years, and the more conservative hypenotists believe it will have significant economic impact within five years.

At my company we build robots that are deployed in warehouses. They have a new fangled locomotion system based on "wheels" (and yes our locomotion system is actually new fangled and did not exist at all or anywhere just two years ago). I have had VCs to whom we have pitched (as is the parlance in startup land) for funding to scale to meet our customer pull, question why we would possibly do that as *everyone knows* that two legged, and two armed, humanoid robots will take over most human jobs in two years.

Whatever I might believe is ultimately irrelevant. But the point is that the hype around humanoid robots comes from the idea that they will be a general purpose machine that can do any manual task that humans can do. Rather than having to change how things are done in order to automate them away from human labor, humanoid robots will be able to step in and just do the existing jobs without having to go to the trouble of changing the way things are done. For that to be true, the humanoid robots will have to be as good as humans at manipulation, just as we have come to expect human level city driving skills from un-crewed robotaxis.

So, we have to get the humanoid robots to be able to do human-like manipulation as that is the key to them making both economic and technological sense. This position is not at all controversial among proponents of humanoid robots. It is precisely humanoids' *raison d'être*. Humanoid builders believe they have to make humanoid robots get closer and closer to human level dexterity to make them make sense. And soon.

2. A Brief History of Humanoid Robots

Many people have already spent decades building humanoid robots, starting with the Humanoid Robotics Institute at Waseda University in Tokyo where WABOT-1 (WAseda roBOT) was built in the early 1970s, after many years of working on biped walking mechanisms in the mid sixties. Then WABOT-2 was built in the early 1980s and new humanoids have followed at Waseda continuously thereafter. Honda, the Japanese car company, started building walking bipeds in the late eighties and eventually unveiled the humanoid ASIMO in 2000. Sony first developed and sold a robot dog named Aibo, then developed a small humanoid robot named QRIO in 2003, but never actually sold copies of it. A French company, Aldebaran, introduced a small walking humanoid named NAO in 2007, and it replaced Aibo as the standard platform in the international robot soccer league that has now been running annually for 30 years. Later they sold a larger humanoid, Pepper, with somewhat less commercial success. Boston Dynamics, a spinout from MIT 35 years ago, introduced the humanoid ATLAS in 2013, after years of building four legged robots.

Besides the early work in Japan on humanoid robots there have been many academic groups across the world that have worked on robots with human form, with and without legs, and with and without arms. My own research group at MIT started building the humanoid Cog in 1992, and we developed seven different platforms, and then I founded Rethink Robotics in 2008, and we sold thousands of two models of humanoids, Baxter and Sawyer. They were deployed in factories around the world. Some of my former post-docs returned to Italy and started the RoboCub open source humanoid project, which has enabled many tens of humanoid robots to be built in AI Labs all over the world.

All these groups have sustained building humanoids and figuring out how to make them walk, manipulate, and interact with humans in built-for-human environments for decades now. Way back in 2004 the International Journal of Humanoid Robotics started publishing, on paper back then.

INTERNATIONAL JOURNAL OF HUMANOID ROBOTICS
Vol. 1, No. 1 (March 2004)

Contents

You can find the journal online now filling its 22nd yearly volume of research papers.

2.1 The manipulation challenge for humanoid robots

Getting a robot to manipulate objects with its arms and hands was very hard for Heinrich Ernst in 1961. It has been hard for every robotics researcher and industrial engineer ever since, and still to this day.

In the mid-sixties *parallel jaw grippers* were developed. Two parallel fingers that moved together and apart. That is still the dominant form of a robot hand today. Here are pictures of ones that I used on robots at Stanford in the 1970s, and pictures of ones my company Rethink Robotics manufactured and sold in the mid twenty-teens, both electrically driven.

The only difference is that the more modern one on the right has a camera in it so that hand can visually servo to a target object–there was not enough computation around in the seventies to do that in a product that had a reasonable price.

Schunk, a German company, sells over 1,000 varieties of parallel jaw grippers, both electric and pneumatic (using compressed air), for robot arms. It also sells some three fingered radially symmetric hands and a few other specialized grippers. No one has managed to get *articulated fingers* (i.e., fingers with joints in them) that are robust enough, have enough force, nor enough lifetime, for real industrial applications.

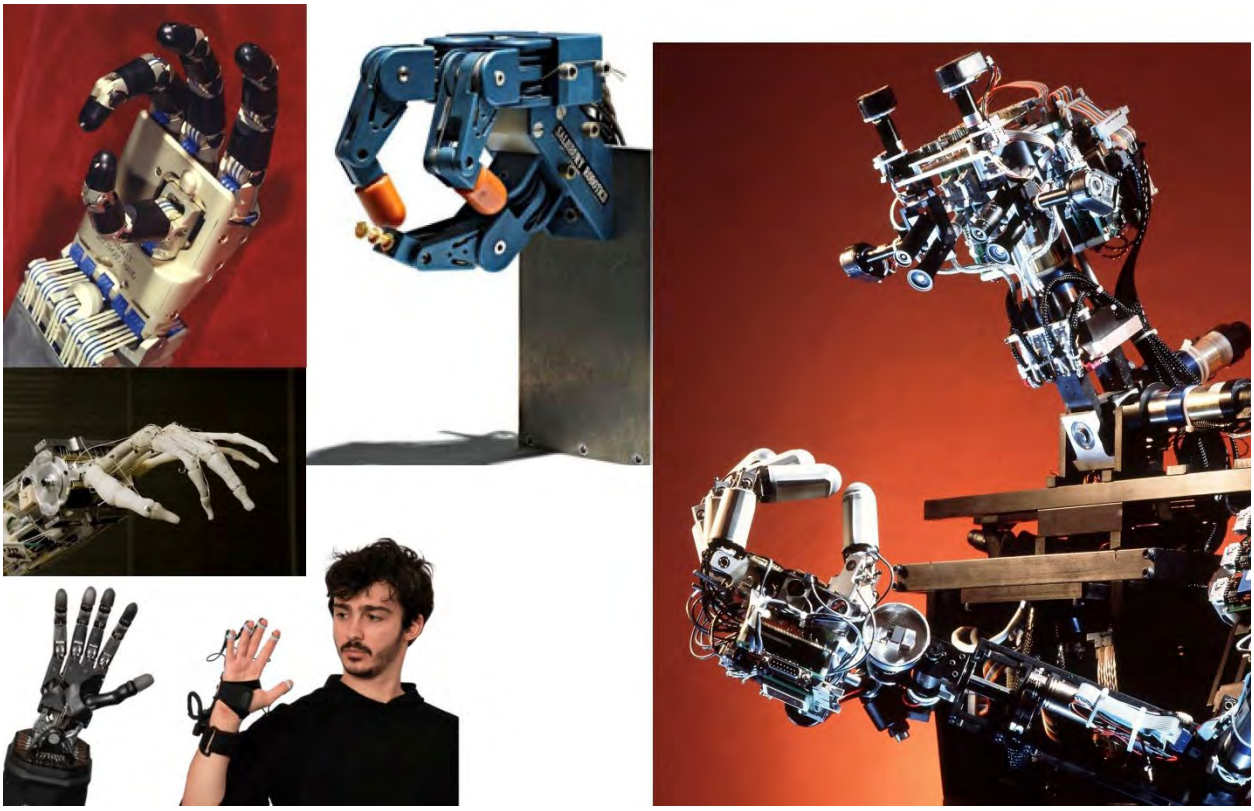
When compressed air is available it can be turned into suction using a Venturi ejector, and the other type of common robot hand uses one or more suction cups to grab an object by a surface. Here is a version that Rethink Robotics sold alongside the electric parallel jaw gripper.

\

Single suction cup and multiple suction cup *end effectors* (the things at the end of an arm

where one might expect a hand) have become quite common for handling finished goods and packing them in custom boxes of identical items, and also for handling cases of finished goods and packages that are being sent to consumers. In fact, there has been a co-evolution of soft material for shipping packages and suction cup end effectors so that soft packages to be sent to people's homes are easier and faster to grab with suction cups than any other method.

Over the last few decades many, many hands modeled on human hands, with articulated fingers, have been built. This montage includes hands built by John Hollerbach, Ken Salisbury, and Yoky Matsuoka.



No human-like robot hands have demonstrated much in the way of dexterity, in any general sense. And none have inspired designs that have made it into deployment in real world applications. The approaches to dexterity have been very mathematical and geometrical, and they have just not produced anything like human dexterity.

You might see pretty videos of human-like robot hands doing one particular task, but they do not generalize at all well beyond that task. In a light hearted, but very insightful, recent blog post, Benjie Holson (full disclosure: Benjie and I work together closely at Robust.AI) lays out fifteen tasks that any eight year old human can do, in a proposed humanoid robot Olympics. With medals. For instance, one challenge is for a humanoid robot folding laundry to hang a men's dress shirt which starts with one sleeve inside out, and to have at least one button buttoned. Another is to clean peanut butter off its own hand. And you can't say "Oh, that would be better done by a different kind of robot mechanism." No, it is central to the case for humanoid robots that they can do all the tasks that humans can do. Once you see

Benjie's fifteen challenge tasks, it is pretty easy to come up with another fifteen or thirty more dexterous tasks which have very little in common with any of his, but which all of us humans can do without a second thought. And then there are the hard things that we can all do if we have to.

2.2 An idea that has worked before

Well gosh, what are we to do? How are we going to get humanoid robots to be dexterous? Here's my imagined inner dialog that so many people must have gone through.

> *End to end learning has worked well over the past 20 years in at least three domains, speech to text, labeling images, and now large language models. So instead of trying to figure dexterity stuff out mathematically, how about we just do end to end learning? We'll collect lots of data about how humans use their hands to do tasks, and feed it into a learning system, and out will pop dexterous robot control. And our companies will be worth billions of dollars.*
>
> *Let's not overthink this, let's just do it!!*

How the humanoid companies and academic researchers have chosen to do this is largely through having a learning system watch movies of people doing manipulation tasks, and try to learn what the motions are for a robot to do the same tasks. In a few cases humans teleoperate a robot, that they can see, along with the objects being manipulated, and the humans may get a tiny bit of force and touch feedback–mostly it comes from the hands of the robots and not the wrists or elbows or shoulders or hips, and any such touch data is very crude.

In his blog Benjie Holson points out the paucity and low accuracy of the data that is collected, and I completely agree with his criticisms. Here they are, he said them well, and I am not going to try to say them better:

> *What I'm seeing working is learning-from-demonstration. Folks get some robots and some puppeteering interfaces (standard seems to be two copies of the robot where you grab & move one of them and the other matches, or an Oculus headset + controllers or hand tracking) and record some 10-30 second activity over and over again (100s of times). We can then train a neural network to mimic those examples. This has unlocked tasks that have steps that are somewhat chaotic (like pulling a corner of a towel to see if it lays flat) or high state space (like how a wooden block is on one of 6 sides but a towel can be bunched up in myriad different ways). But thinking about it, it should be clear what some of the limitations are. Each of these has exceptions, but form a general trend.*
>
> ***No force feedback at the wrists.*** *The robot can only ever perform as well as the human teleoperation and we don't yet have good standard ways of getting force information to the human*

> *teleoperator.*
>
> ***Limited finger control.*** *It's hard for the teleoperator (and AI foundation model) to see and control all the robot fingers with more finesse than just open/close.*
>
> ***No sense of touch.*** *Human hands are packed absolutely full of sensors. Getting anywhere near that kind of sensing out of robot hands and usable by a human puppeteer is not currently possible.*
>
> ***Medium precision.*** *Guessing based on videos I think we've got about 1-3 cm precision for tasks.*
>
> ***Folding towels and t-shirts doesn't depend on high wrist forces.*** *You can get away with just hand open/close by using pinch grasps to pull and lift and open hands to spread. You can visually see how your grasp is so you don't need finger sensing. 1-3 cm precision is just fine.*

And yes, this is real. Humanoid robot companies, and many academic projects, are trying to train robots to do dexterous manipulation by just showing them the motions, and not getting them to use any force or haptic feedback.

For instance, in the last week Figure has announced their "project go big" about how they are going to train robots with new skills. Nothing is surprising here as it matches what they have be saying and showing all along. And here is what they say about it, with my bolding:

> *Traditionally, teaching robots new skills required costly demonstrations, hand-coded programs, or tightly staged environments that fail to capture the messiness of the real world. Humanoid robots, however, offer a unique structural advantage: their perspectives and kinematics mirror our own,* ***making it possible to transfer knowledge directly from everyday human video*** *(Video 1).*

[[And do note that Video 1 is distinctly unmessy, and uncrowded, unlike any real home that ordinary people live in. Likewise for Videos 2 and 3.]]

They are saying that they are going to train their robots to do new manual skills from first person video of people doing those skills.

And here is a press story from eWeek, from just a month ago, that Tesla is going all in on training through simply looking at videos of people doing tasks. It says:

> *Tesla has shifted its training strategy for its humanoid robot Optimus. Instead of relying on motion capture suits and teleoperation, Tesla is moving toward a vision-only approach.*
>
> *Workers now wear camera rigs made up of helmets and backpacks with five in-house cameras that record mundane tasks like folding a t-shirt or picking up an object. Those videos are then used to train Optimus to mimic the actions.*

A little further in the story it says:

> *Christian Hubicki, director of the robotics lab at FAMU-FSU, noted to Business Insider that the multiangle camera setup likely captures "minute details, like the location of joints and fingers," making the data more precise.*

Both Figure and Tesla are all in on videos of people doing things with their hands are all that is needed to train humanoid robots to do things with their hands. They are making a bet that machine learning from watching lots and lots of motions of people's hands will be sufficient to learn dexterity. Visual precision and large data sets of it are enough, they believe. [[It is possible that they are sandbagging us, as the lure of $30 trillion is quite a lot of money, even for a very already rich person, and they just don't want the competition to know what they are really doing. But I am going to take them at their word for this argument.]]

3. End to End Learning Depends on the Chosen Ends

In the last two decades speech to text, image labeling, and fluid language generated by Large Language Models (LLMs) have all been transformed, in spectacular ways, by end-to-end learning, using linear threshold neural models.

For the speech and image cases the new methods showed radical increases in performances. In both cases leaving as much as possible to the learning methods was critical for that success. That meant, for speech, getting rid of explicit models for phonemes (which are very language dependent), which had dominated all previously recent approaches. For image labeling it meant getting rid of any notion of line (boundary) finding, shape, shading, or color constancy, all of which had dominated recent work on image understanding.

LLMs showed proficiency with language and answering general questions (with a strong tendency, still today, for rabid confabulations) that was beyond anything anyone was expecting was around the corner. And they had done it by eliminating any references or direct experience in the world of anything other than language. They were self-contained language machines, with none of the long expected grounding in experience in the real world, that everyone had expected, the believed to be *symbol grounding problem*. [[Even Alan Turing had brought this up in his brilliant Intelligent Machinery, written in 1948 but not published until 1970, in Machine Intelligence 5, edited by Bernard Meltzer and Donald Michie. There (on page 13 of that volume) Turing had said that the sure way to get to an intelligent machine was to "take a man as a whole and to try replace all parts of him by machinery". Today we might say "build a humanoid"; prescient! As for the grounding in real world experience he went on to say: "In order that the machine should have a chance of finding things out for itself it should be allowed to roam the countryside, and *the danger to the ordinary citizen would be serious*" (my emphasis). He concluded that it was too hard to do with the technology of the day. Two more instances of prescience.]]

These were radical changes and head spinning for most researchers, including me. But the new methods undeniably worked much better than anything we had seen in the past.

On March 13th, 2019 (pre LLM), Rich Sutton (who was later the 2024 co-winner of the Turing Prize with Andrew Barto for their work on Reinforcement Learning) published a mildly triumphant short blog post titled A Bitter Lesson. In it he applies his argument to more cases than the ones to which I refer to here, by including the role of massive search making computers playing Chess and Go much better than humans doing the tasks.

And he says, both for search and learning approaches:

> *And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation.*

Then he goes on to discuss Chess, Go, speech, and images. He argues against using human biases in structuring the problems at all. But I thought then, and think now, that, in reality, in all these successful cases human knowledge does come into play, as the "end to end" nature relies on humans specifying what the "ends" are.

Six days after Sutton posted this I replied in the form of similarly short blog titled A Better Lesson. In that post I pointed out a number of generic problems with scaling the approach, as we see now with massive energy and server requirements and the employment of thousands of other humans preparing data sets, which itself belies the argument of keeping humans out of the loop.

Most importantly I pointed out that the image labeling case was not end-to-end starting with images and ending with labels. Instead it uses a convolutional network as a front end to structure the way the learning algorithm has access to the images. While I did not make a similar argument for speech to text, nor the as-then-yet-unannounced LLMs, I will make the the case that all three succeeded due to engineers building case specific pre-processing that relied on directly simulating (without learning) parts of human physiology.

Here are the accommodations made to learning, in each of the three cases, in terms of hard coding front end processing of data.

3.1 Speech to text

The task in speech to text is to take the signal from a microphone that a person is speaking into and to output a text string that represents the words that were said. Today we are all used to talking to various machines like Alexa, or our TV remote, or our car, and on a customer service line, or any other of a myriad of devices and channels. All these use speech to text to get the words to feed into the system which is going to respond appropriately (we hope) to our words. It is only in the last 20 years that this capability has become practical. And, it is the result of end to end learning over large data sets, where both the microphone input and the correct text string were available, and that a learning system learned how to go from the input signal to generating text.

There are many ways the sound signal could get into the computer for this learning. We could take the analog output of a microphone and digitize the loudness of the signal many tens of thousands of times a second and have that as the input to the learning. But in practice that is not how it is done.

Instead it relies on a technology developed for practical spoken communication over world wide telephone networks in the 20th century where the signals were compressed for individual voice circuits so that more calls could fit on a single wire. This work determined aspects of the signal that had to be preserved so that a human could understand what had been said by the distant speaker. And if a human could understand such compressed signals that says that all the information that is necessary to understand speech is still in that signal.

The inputs to various speech to text learning systems vary, but here are some of the common pre-processing steps taken. The analog input signal is sampled at a fixed frequency, perhaps 16kHz, then a high pass filter boosts higher frequencies as they are important for detecting consonants, then the signal is cut into frames, say 25ms long with 10ms overlap, for instance, and then each frame is conditioned so that subsequent Fast Fourier Transforms (FFTs) will not be compromised by the shortness of the window. Somewhere along the way there may be some noise reduction. Then the signal is subdivided into frequency bands using one or more methods such as FFT, Mel filter banks, logarithm of outputs, and cosine transforms. In some implementations there is initial training on just the frames so that language dependent frame signatures can be recognized early in the deep network.

Different implementations use different selections of these and other techniques, but the point is that **after** all this, **then** end-to-end learning is let loose on the output of all this input signal transformation.

Further, all that signal transformation was originally developed so that human speech could be stored and the listened to in distant places and times. The important thing about those transformations was that they allowed the human listening mechanism for understanding speech to be used with no change to the human.

3.2 Image labelling

Image labelling through *deep learning,* has since 2012 become the dominant method in computer vision for interpreting what is in an image. But deep learning does not start with raw pixels coming out of a camera, rather it bows to non-learned human physiology in two ways.

The data coming out of a camera is a linear stream of pixel values, and in fact sometimes three separate streams for the directly measured colors red, green, and blue (RGB). A modern digital camera has a global (electronic) shutter where light coming through the lens is allowed to bump electrons into a rectangular array of little buckets, all starting and stopping collecting at the same time. Then the contents of those buckets are shifted to neighboring buckets and read out with an analog-to-digital converter, which essentially reads the number of electrons in a particular bucket, and they are read as a series of left-to-right rows, top-to-bottom, or some switching of either of those orders. It is one, or three for color images, linear streams.

Deep learning does not operate on this stream. Instead the numbers from this stream are arranged in a data structure which reconstructs the adjacencies of the original pixels, and

for color, overlays the three colors. This is of course standard in any image processing by computer, but it is an explicit structure being imposed intentionally. Animals do not serialize their images, but instead have one cable going from each "pixel" in the retina, to a flat array of cells in the cortex, where the geometry of the pixels, or receptors, in the retina are preserved. The organization of these cables into a regular array happens before birth through bursts of localized excitations of the adjacent retinal cells that are then used at the other end to guide the development of the cables (which are all neural axons) to mimic the locality of excitement.

Then the first few layers for deep learning use a structure which is set up so that learning learns the same thing in a translationally invariant way; a cat in the lower left of an image is recognized in exactly the same way as one in the middle top of an image. This specialized network is a *convolutional neural network,* a processing structure specialized for vision applied to large images.

In the 27th of May, 2015 Nature (paywalled) article by Yan LeCun, Yoshua Bengio, and Geoffrey Hinton (the three winners of the 2018 Turing Prize) titled Deep learning, the authors say:

> *First, in array data such as images, local groups of values are often highly correlated, forming distinctive local motifs that are easily detected. Second, the local statistics of images and other signals are invariant to location. In other words, if a motif can appear in one part of the image, it could appear anywhere, hence the idea of units at different locations sharing the same weights and detecting the same pattern in different parts of the array.*

They go on to attribute this architecture to Kunihuko Fukushima, who worked on learning to recognize hand drawn characters (pre back propagation), as did Yan LeCun (post back propagation) some years later. The earliest English-language non-paywalled paper that I can find by Fukushima on this topic appeared at the International Joint Conference on Artificial Intelligence (IJCAI) in 1979 in Tokyo and the three page paper is on page 291 of Volume 1 of the proceedings. [[That was the first international conference at which I presented my own paper, and it is in the same volume, and is about a much more ancient and largely discarded approach to recognizing objects in images.]]

Fukushima attributes inspiration for this approach to the investigations of the structure of the cortical columns in the cortices of cats and monkeys by David Hubel and Torsten Weisel, who won the Nobel prize for this work in 1981 — see David Hubel's Nobel lecture for a summary of that work. Fukushima emulated both the *simple cells* and the *complex cells* that Hubel and Weisel identified as S-cells and C-cells, and then split Hubel and Weisel's *hypercomplex* cells into two subtypes within his modeled cells. These cells recognize common motifs wherever they may appear in an image.

In figure 2 of the paywalled Nature article above you can see this structure play out in alternate layers, and as LeCun *et al* say:

> *There are four key ideas behind ConvNets that take advantage of the*

> *properties of natural signals: local connections, shared weights, pooling and the use of many layers.*

In animals, including humans there is an additional variation in resolution of receptors in the retina, with more closely spaced, and therefore higher resolution, receptors near the center of the visual field. Many animals, including humans, use very fast motions, saccades, of their eyeballs to point that high resolution part of their eyes at different parts of the image — you are doing this right now as you read these words, saccading along each line then down to the next, stopping for just a fraction of a second before moving on (and suppressing your motion sensors while you move your eyeball).

The large convolutional network for deep learning vision eliminates the need for this by having high resolution recognition, through repetitive shared weights, across the whole image.

Again, this is not raw end to end learning. There is very detailed replication of incredibly complex parts of our brain, that is structured into the learning machine. Despite the romanticism of having everything learned without humans messing things up by choosing the wrong structures, deep learning image labelling is built upon a very complex and marvelous piece of front end engineering that specifically emulates structures that have been identified in animal brains. And it is built upon the technology we have developed to capture images and transmit them over a narrow channel (i.e., to serialize them) so that the human visual system can understand the original scene even when that human is located at a distant point in space and time.

3.3 Large language models

Large Language Models (LLMs) e.g., ChatGPT or Gemini, are trained on large amounts of text, with no external inputs trying to explain what all that text is. From that perspective it looks like the learning mechanism figures everything out itself.

However, there are some early stages both in learning and then later processing any input text where the structure of human language, and some aspects of the particular human language that is being input, has been used to engineer in some preprocessing, and some aspects of the internal representations. The two mechanisms for this involve *tokens* and *embeddings*. [[Of course, then there is the whole transformer mechanism, invented in 2017, involving multi-head attention mechanisms, and one step at a time append and shift output being rerouted to the input, and so on. That is a massive amount of human-generated architecture and engineering which is key to LLMs working, further pressure on the insistence on end-to-end learning with no human biases built in. But here I am only talking about the early massaging of data that is common to this and the previous two subsections.]]

The fundamental unit of any particular language is presented to an LLM as a linear sequence of tokens. For English roughly 50,000 different tokens are used and they include tokens such as ***dog***, ***cat***, ***fish***, ***game***, ***run***, ***ing***, ***ed***, ***pre***, ***sub***, ***due***, ***marine***, ***fetch***, ***auto***, etc. Tokens can be whole words prefixes, suffixes, common subparts of words, etc.

At the very beginning of training an LLM, with text in a particular language, tokens are

learned, in a largely unsupervised manner. Lots of text in the language is fed into a token learning system which comes up with plausible token candidates based on commonality of seeing them in the training corpus, and with statistics attached as to how common they are, and whether and how the combine with other tokens within words. From these statistics the number of discrete tokens to be used is chosen, automatically, by scoring possible tokens based on frequency and how well they divide words into other common tokens.

Once the tokens have been chosen a small program, *the tokenizer*, is used to break all input language into strings of those tokens.

Next, the tokens are embedded in a high dimensional vector space, typically one that has 3 x 2^n dimensions for some fixed *n*. Over recent years as more training has been applied to LLMs to produce bigger models the number *n* has gotten larger. For ChatGPT-2 *n* was 8, but for ChatGPT-3 it was 12.

The embedding needs to be learned, i.e., the coordinate in each dimension of the vector space needs to be filled in for each token. This is done by a second "pre-real-training" learning exercise, where the ways in which any two tokens seem to be substituted for each other, in contexts in raw text that seem similar by the tokens that surround that context. It appears that this sort of learning ends up choosing embeddings for tokens such that their distance in different subspaces (for the standard definition of a subspace of a vector space) of the overall embedding correspond to some sorts of similarity. For instance, *orange* and *red* may be closer in one subspace than either is to *fruit*, but in another subspace *red* might be an outlier compared to the closeness of the other two. The first subspace might correspond more to color, and the second subspace as considering what class of tangible objects in the world the words can designate. But such decisions are not made by humans, both the categories and the distances are generated by learning from the data.

The number *n* is chosen early on by the people building a new LLM based on their tolerance for paying for cloud services as it will be a big factor in how much data is needed to train the LLM and how many parameters will need to be learned.

Once there is an embedding like this, the very first stage of the neural network that represents the LLM takes each token from the output of the tokenizer and turns it into its vector position in the embedding. So, in the case of ChatGPT-3 where *n* = 12, each token is immediately turned into 12,288 numbers.

Thus we see here that there has been a lot of human engineering and knowledge about the ideas of word components, and sorts of meanings of words and how similarity can be extracted from language without knowing meanings has been applied to way in which the pre-training is done for a language.

In one sense the tokens are proto-symbols, but unlike traditional symbols it is not their unique identity that is important but how they compare to other proto-symbols within the system. AND, these proto-symbols are based on parts of human language, the parts that the invention called writing uses to transmit language between people without the necessity to do it in sound or in a synchronized manner — writing can be read anywhere at any later time, even well after the writer is dead.

3.4 The commonality in these three applications of end to end learning

These three grand successes of end to end learning rely on very domain specific learning architectures down stream. But they also each rely on domain specific early processing of the data stream.

In these three cases that early processing was built for other purposes, for language to be heard or read, and for images to be seen, at entirely different locations and asynchronous time.

We do not have such a tradition for touch data. Touch for us, for now, is only the instantaneous touch we perceive first hand (no pun intended). We as a species have not developed technologies to capture touch, to store touch, to transmit touch over distances and time, nor to replay it to either ourselves or other humans.

In section 4 below I show how central touch is to human dexterity.

To think we can teach dexterity to a machine without understanding what components make up touch, without being able to measure touch sensations, and without being able to store and replay touch is probably dumb. And an expensive mistake.

4. Why the Ends are Uncracked for Dexterity

The center piece of my argument is that the brute force learning approaches that everyone rightfully touts as great achievements relied on case-specific very carefully engineered front-ends to extract the right data from the cacophony of raw signals that the real-world presents.

If it is the case for the big successes it is likely also the case for learning dexterity by brute force. If any one or any group is to succeed they will likely have to collect the both the ***right data***, and learn the ***right thing***. Most of the projects to teach humanoids dexterity are doing neither of these things. There are some exciting and promising experiments going on in academic laboratories, but they have not yet gotten close to demonstrating any real dexterity. By my third law of robotics that says that we are more than ten years away from the first profitable deployment of humanoid robots even with minimal dexterity.

Human dexterity relies on a rich sense of touch. And dexterity for humans involves more than just their hands; it often involves their elbows, the fronts of the bodies, legs, and feet (many machines have foot pedals). I am not going to present a comprehensive complete case for it here, as one might expect if this were a formal peer reviewed academic research paper. But I will show you results from a somewhat random selection of solid peer reviewed academic work stretching over fifty years which together demonstrate that humans use touch and force sensing extensively.

4.1 The human sense of touch is really rich and complex

The following two videos are from Roland Johansson's lab at Umeå University in Sweden where he has studied human touch for decades. In the first video the person picks a match out of a box and lights it. The task takes seven seconds. In the second video the same

person tries again but this time the tips of her fingers have been anesthetized so she no longer has any sense of touch right at her fingertips. She can still sense many other things in the rest of her fingers and hand, and all the forces that she can ordinarily feel with her skeletal muscle system.

[The two URLs in case your browser does not point at the YouTube videos below:
www.youtube.com/watch?v=zGIDptsNZMo
www.youtube.com/watch?v=HH6QD0MgqDQ]

Without a sense of touch in her fingertips the person makes many unsuccessful attempts to pick up a match from the box, then fails to pickup an isolated match that had fallen on the table, then goes back to the box and straightens up the matches, manages to pick one up, then fumbles with the match trying to get it into the right orientation between her fingers, and successfully lights it after taking four times as long as she took with sensitive fingertips.

It looks like humanoid robots will need a sense of touch, and a level of touch sensing that no one has yet built in the lab in order for them to do tasks like the one above which is of the same order of difficulty that millions of workers do all day everyday in some parts of the world. [[I have visited well over 100 factories in the US, China, Japan, Korea, Taiwan, and Germany, some where my companies have been building my five major families of robots: Roomba, PackBot, Baxter, Sawyer, and Carter, and some where I have been selling robots to make workers in the factories more productive, and some where I was on technology advisory boards for the companies that ran the factories. I have seen this and many other types of dexterity of humans beings applied to complex tasks in all these factories.]]

In a review of Johansson's earlier work from 1979 it is reported that a human hand has about 17,000 low-threshold mechanoreceptors in the glabrous skin (where hair doesn't grow) of the hand, with about 1,000 of them right at the tip of each finger, but with much lower density over the rest of each finger and over the palm. These receptors come in four varieties (slow *vs* fast adapting, and a very localized area of sensitivity *vs* a much larger area) and fire when they sense pressure applied or released.

Next I will talk briefly about the work of David Ginty and his students in his lab at Harvard. You can see the lab's complete list of publications here, stretching back to 1987. The mission of Ginty's lab is:

> *We use approaches in molecular genetics, anatomy, physiology, behavior, and systems neurobiology to understand mammalian somatosensory neurons and central nervous system circuits that underlie our sense of touch.*

From a press article summarizing almost forty years of Ginty's work touch is described as follows:

> *touch concerns a smorgasbord of stimuli, including pokes, pulls, puffs, caresses and vibrations, as well as a range of temperatures and chemicals, such as capsaicin in chili peppers or menthol in mint. From these inputs arise perceptions of pressure, pain, itchiness, softness and hardness, warmth and cold, and the awareness of the body in space.*

The article goes on to report that there have now been fifteen different families of neurons discovered that are involved in touch sensing and that are found in the human hand.

> *Such nerve endings turned out to be remarkably specialized. Near the skin's surface, the flat variety, called a Merkel cell complex, responds to gentle indentation. Merkel cells abound in your lips and fingertips, allowing you to discern form and texture. Your fingers are also packed with coiled nerve endings called Meissner corpuscles, which wrap around support cells in a bulbous tangle. These sensors pick up the faint, minuscule vibrations generated by the slight slipping of an object against your hand as you grip it, enabling you to use tools with precision. Deeper in the skin dwell the onionlike Pacinian corpuscles, which detect rumblings in the earth, and the spindle-shaped Ruffini endings, which convey skin stretching.*

Touch is a very complex set of sensors and processing, and gives much richer time dependent and motion dependent information than simple localized pressure.

Moving on to more general aspects of humans and what we sense as we manipulate, on top of that skeletal muscles sense forces that they are applying or that are applied to them. Muscle spindles detect muscle length and when they stretch, and Golgi tendon organs sense tension in the muscle and hence sense force being applied to the muscle.

We also make visual and touch estimates about objects that change our posture and the forces we apply when manipulating an object. Roland Johansson (again) describes how we estimate the materials in objects, and knowing their density predict the forces we will need to use. Sometimes we are mistaken but we quickly adapt.

Over the last two decades Roland Johansson's work has shifted to understanding the role of forethought based on observations in how humans choose appropriate strategies for carrying out tasks with their hands and bodies. You can read his last twenty years of publications here. His papers include titles such as:

- Fingertip viscoelasticity enables human tactile neurons to encode loading history

alongside current force

- Human touch receptors are sensitive to spatial details on the scale of single fingerprint ridges
- Gaze behavior when learning to link sequential action phases in a manual task
- Integration of sensory quanta in cuneate nucleus neurons in vivo
- Skill learning involves optimizing the linking of action phases
- Slowly adapting mechanoreceptors in the borders of the human fingernail encode fingertip forces.

These show how rich and varied human grasping is beyond simple motions of fingers, even if the positions of the fingers can be measured accurately (see the reference in section 2.2 above, to Tesla's newest data collection strategy).

4.2 What is the right data?

Collecting just visual data is not collecting the **right data**. There is so much more going into human dexterity that visual data completely leaves out.

Is anyone trying to do more than collect visual data and have a different more appropriate "end" to connect learning to?

Apart from Figure and Tesla which are explicitly claim not to be doing so, the other big companies are not saying. And there are lots of big companies working on humanoid robots, and you can sort of tell by seeing which of your friends are getting hired by which company.

In academia though, there is still a healthy set of experiments going on. Here is just one example, from the "best paper" from the May 2025 *Dexterous Human Manipulation* workshop at the *Robotics Systems and Science* conference. It comes from Pulkit Agrawal's group centered in CSAIL at MIT. It involves a newly invented way to collect the **right data** to feed to machine learning. As you can see in the two pictures below the human essentially has their hand in a glove. There is a robot hand rigidly attached to the glove so the robot hand is roughly 10cm away from the human hand and completely parallel to it. The human moves their fingers to control the robot hand fingers and the human moves their arm to place the robot hand in contact with objects to be manipulated.. The robot fingers and palm have touch sensors which feed to the data collection system *and to the actuators which stimulate the human's finger tips and palm*. [[This was the original wording when I posted this, but I misunderstood something, and prompted by a reader I reached out to Pulkit to clarify. The part now in italics is not true in the sense of there being active actuators stimulating the person. It is true in the lesser sense that the human feels joint-level force feedback.]] So while this system doesn't record the forces that the human directly feels and controls with their arms, it does get to associate finger motions generated by a human with touch sensations that the human is sensing as they decide how to control the robot hand.

Clearly this is a long way from understanding everything that a human does with their wildly complex touch and force sensing system, but it is a step beyond simply collecting visual data, which alone can't possibly be enough to infer how to be dexterous.

[[If the big tech companies and the VCs throwing their money at large scale humanoid training spent only 20% as much but gave it all to university researchers I tend to think they would get closer to their goals more quickly.]]

4.3 What is the right thing to learn?

Lastly I want to return to what I said at the start of this section (4) about the need to learn the **right thing**.

The framework that both industry and academia is using on what to learn comes from Reinforcement Learning (see the introduction part of section 3, above). In Reinforcement Learning, one learns a *policy*, which maps from the *state*, expressed by what the sensors are delivering right now, to a specific *action* for the robot to do right now.

But it seems from both personal experience and from some of the papers from haptics researchers above, that humans are sometimes pursuing a dexterity plan of what they are trying to do. Instead of what is being sensed mapping directly to action, what is being sensed probably modulates what is being done in following that plan (represented as a finite state machine, perhaps?). Thus to be truly successful at dexterity there needs to be a way to learn both how to plan in some weird space of subtasks, and how sensing at the tactile level should modulate those plans.

There is still plenty of research to be done to figure all this out. And then years to get to solid lab demos, then years more to get to deployable systems that bring value to customers.

5. The Other Problem for Humanoid Robots: Walking

I think it is fair to say, given the aspirations that humanoid robots have the same form as humans so that they can operate in built-for-human environments, people will expect them to be safe to be around. This is especially true for humanoids providing healthcare in the home for an aging human population. But by the master plans set out for humanoid robots it must be true in other environments too, as the idea is that the humanoid robots fit into human spaces there too. And that means humans will share those spaces, otherwise why not just build a special purpose lights out machine that can do the job.

So if anyone is going to deploy humanoid robots at scale it is important that they **be safe**

for real humans to share space with them, so be just centimeters away from them, to lean on the humanoids for support, to be touched and manipulated by humanoid robots (as are the elderly touched and manipulated by human carers, helping them stand, wash, poop, get into and out of bed, etc.).

The trouble is that human sized two legged walking humanoid robots are not currently safe for humans to be around. But the argument for humanoid robots requires that they be full sized, so they can operate in human spaces and do all human tasks.

Ah, but you've seen videos of, or walked within centimeters of (as I have), half sized humanoid robots, feeling quite safe around them. So you reason that it is only a matter of a small amount of time before those robots are made bigger. But that is where physics comes in, with a vengeance.

Current humanoid robots do not walk at all like humans. Humans are stretchy springy systems, that very nearly walk without much in the way of neural control. In fact you can see models of biped walkers that are purely mechanical, walking down a gentle slope, with no power supply, relying only on the passive dynamics of the mechanism, and stealing potential energy from the act of walking downhill to power the robot (purely mechanically).

Here is a simple example:

[The URL is www.youtube.com/watch?v=wMlDT17C_Vs]

Besides that fundamental architecture, we also have an energy recycling architecture involving our muscles and tendons. We store energy in our tendons and reuse it on the next step — our Achilles tendon at the back of each of our lower legs is the one that stores most energy and the one most likely to rupture.

Although there have been decades of academic research on building robots that walk like us in this regard, they have not gotten to the practical level that the current humanoid robots designs have reached.

But current humanoid robots use powerful electric motors to balance by pumping large amounts of energy into the system when there is instability, mostly following a version of the ZMP (Zero-Moment Point) algorithm. [[This algorithm has been around for a long time, and in the 2004 Volume 1 of the International Journal of Robotics, shown above at the start of section 2, on page 157, Miomir Vukobratović and Branislav Borovac, both from Serbia and Montenegro, had a paper celebrating their introduction of the algorithm thirty five years prior to that, making it roughly 56 years old now.]] Although they are tight lipped about exactly what they are doing the large companies working on humanoids seem to have added some Reinforcement Learning (RL) on top of ZMP starting points, to get better walking and less falls. ZMP relies on sensing forces in the sole of the feet, and so all humanoid robots do have that. But the RL algorithms rely on the whole structure being very stiff so humanoid robots are the antithesis of humans when it comes the mechanical structures doing the walking. These robots fall less often, but are still very dangerous for humans to be close to them when they do and will fall.

When an instability is detected while walking and the robot stabilizes after pumping energy into the system all is good, as that excess energy is taken out of the system by counter movements of the legs pushing against the ground over the next few hundred milliseconds. But if the robot happens to fall, the legs have a lot of free kinetic energy, rapidly accelerating them, often in free space. If there is anything in the way it gets a really solid whack of metal against it. And if that *anything* happens to be a living creature it will often be injured, perhaps severely.

But, but, but, the half sized humanoids are safe, so how much less safe can a full size humanoid robot be?

This is where scaling comes in, not in terms of numbers of robots, but in scaling laws of physical systems.

If you just expand a physical system by the same amount in every direction, say multiply all lengths by a scale factor s, then the mass m of the system goes up by s^3. Since $F = ma$ for the same acceleration you need to put in s^3 as much energy. So for a robot that is 50% bigger that is $(1.5)^3 = 3.375$. And to get from today's small safe-ish humanoids you have to pump in $2^3 = 8$ times as much energy. That is a whole different class of possible injuries. And it could be even worse, as for a limb, say, the mass goes up as the cube of s but the cross section, which determines strength, only goes up as the square. [[This scaling is why elephants have much fatter legs for their body size than does a spider, even accounting for the latter having twice as many legs to support its weight.]] So the twice bigger robots may have to have proportionally much fatter legs, so more mass, and so they will pump up the energy by something larger than a factor of eight.

My advice to people is to not come closer than 3 meters to a full size walking robot. And the walking robot companies know this too. Even in their videos you will not see people close to a locomoting humanoid robot unless there is a big table between them, and even then

the humanoids only shuffle around a little bit.

Until someone comes up with a better version of a two legged walking robot that is much safer to be near, and even in contact with, we will not see humanoid robots get certified to be deployed in zones that also have people in them.

6. What is the Future of Humanoid Robots?

Technology changes and the meanings of words around technologies change too.

When I made a whole bunch of dated predictions about future technologies back on January 1st, 2018, *flying cars* and *self-driving cars* meant different things than they do today. I pointed this out in my most recent scorecard on how my predictions were holding up.

Flying cars used to mean a vehicle that could both drive on roads and fly through the air. Now it has come to mean an electric multi-rotor helicopter than can operate like a taxi flying between various fixed landing locations. Often touted are versions that have no human pilot. These are known as eVTOLs, for "electric vertical take off & landing". Besides not yet actually existing in any practical sense, flying cars (eVTOLS) are no longer cars, as they do not travel anywhere on the ground.

At the time I made my predictions *self driving cars* meant that the cars would drive themselves to wherever they were told to go with no further human control inputs. Now self driving cars means that there is no one in the driver's seat, but there may well be, and in all cases so far deployed there are, humans monitoring those cars from a remote location, and occasionally sending control inputs to the cars. Except for Tesla self-driving robo taxis. In that case there is a human safety operator sitting in the front passenger seat.

Following that pattern, what it means to be a *humanoid robot* will change over time.

Before too long (and we already start to see this) humanoid robots will get wheels for feet, at first two, and later maybe more, with nothing that any longer really resembles human legs in gross form. But they will still be called *humanoid robots*.

Then there will be versions which variously have one, two, and three arms. Some of those arms will have five fingered hands, but a lot will have two fingered parallel jaw grippers. Some may have suction cups. But they will still be called *humanoid robots*.

Then there will be versions which have a lot of sensors that are not passive cameras, and so they will have eyes that see with active light, or in non-human frequency ranges, and they may have eyes in their hands, and even eyes looking down from near their crotch to see the ground so that they can locomote better over uneven surfaces. But they will still be called *humanoid robots*.

There will be many, many robots with different forms for different specialized jobs that humans can do. But they will all still be called *humanoid robots*.

And a lot of money will have disappeared, spent on trying to squeeze performance, any performance, from today's *humanoid robots*. But those robots will be long gone and mostly conveniently forgotten.

That is the next fifteen years for you.