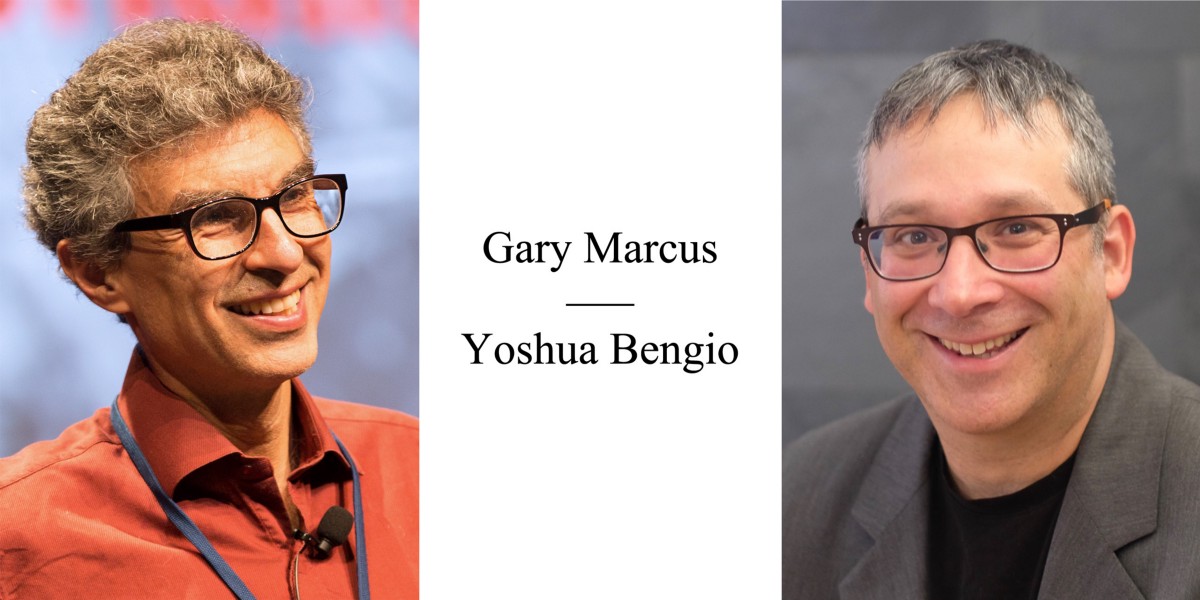
**Yoshua Bengio and Gary Marcus on the Best Way Forward for AI**

**Transcript of the 23 December 2019 AI Debate, hosted at Mila**

**Moderated and transcribed by Vincent Boucher, Montreal AI**

AI DEBATE : Yoshua Bengio | Gary Marcus — Organized by MONTREAL.AI and hosted at Mila, on Monday, December 23, 2019, from 6:30 PM to 8:30 PM (EST)

Slides, video, readings and more can be found on the MONTREAL.AI debate [webpage](https://montrealartificialintelligence.com/aidebate/).

**Transcript of the AI Debate**

**Opening Address | Vincent Boucher**

Good Evening from Mila in Montreal Ladies & Gentlemen,

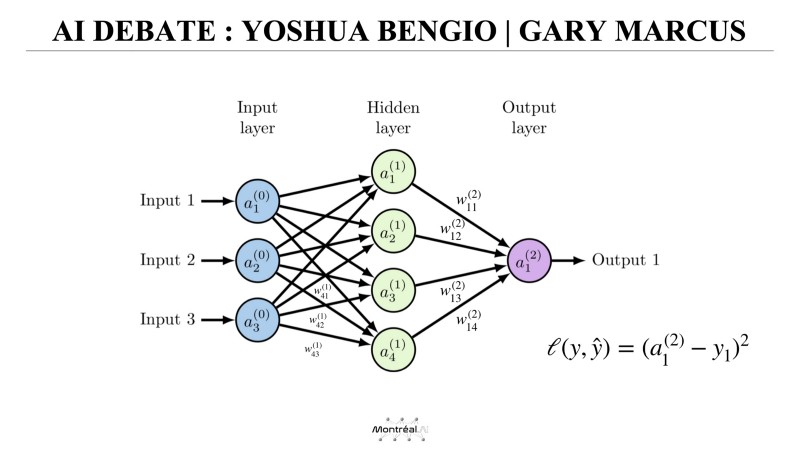
Welcome to the “**AI Debate**”.

I am [Vincent Boucher](https://www.linkedin.com/in/montrealai/), Founding Chairman of [Montreal.AI](http://www.montreal.ai/).

Our participants tonight are Professor [**GARY MARCUS**](http://garymarcus.com/) and Professor [**YOSHUA BENGIO**](https://mila.quebec/en/yoshua-bengio/).

Professor GARY MARCUS is a Scientist, Best-Selling Author, and Entrepreneur. Professor MARCUS has published extensively in neuroscience, genetics, linguistics, evolutionary psychology and artificial intelligence and is perhaps the youngest Professor Emeritus at NYU. He is Founder and CEO of Robust.AI and the author of five books, including The Algebraic Mind. His newest book, Rebooting AI: Building Machines We Can Trust, aims to shake up the field of artificial intelligence and has been praised by Noam Chomsky, Steven Pinker and Garry Kasparov.

Professor YOSHUA BENGIO is a Deep Learning Pioneer. In 2018, Professor BENGIO was the computer scientist who collected the largest number of new citations worldwide. In 2019, he received, jointly with Geoffrey Hinton and Yann LeCun, the ACM A.M. Turing Award — “the Nobel Prize of Computing”. He is the Founder and Scientific Director of Mila — the largest university-based research group in deep learning in the world. His ultimate goal is to understand the principles that lead to intelligence through learning.

Diagram of a 2-layer Neural Network

The diagram shows the architecture of a 2-layer Neural Network.

*“You have relatively simple processing elements that are very loosely models of neurons. They have connections coming in, each connection has a weight on it, and that weight can be changed through learning.*” — Geoffrey Hinton

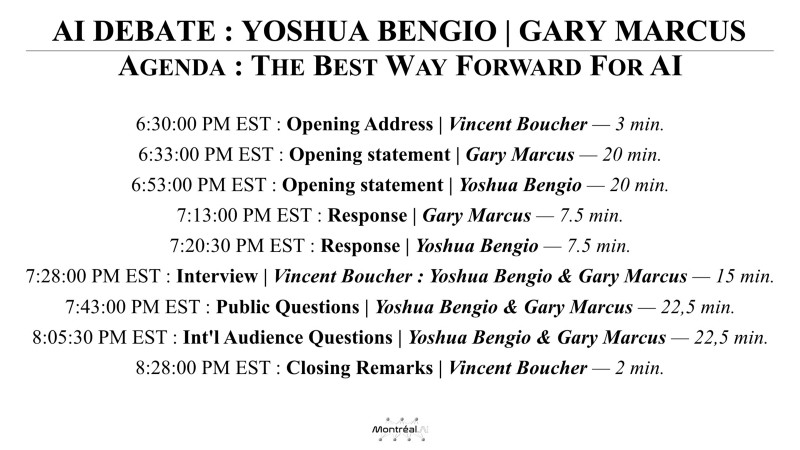
Deep learning uses multiple stacked layers of processing units to learn high-level representations.

Professor MARCUS thinks that expecting a monolithic architecture to handle abstraction and reasoning is unrealistic.

Professor BENGIO believes that sequential reasoning can be performed while staying in a deep learning framework.

**Our plan for the evening**

An Opening statement by Gary Marcus and by Yoshua Bengio; followed by a Response, an interview with Yoshua Bengio & Gary Marcus; then our guests we’ll take questions from the audience here at Mila; followed by questions from the international audience.

Agenda : The Best Way Forward For AI

This AI Debate is a Christmas gift form MONTREAL.AI to the international AI community. The hashtag for tonight’s event is : #AIDebate

International audience questions for Gary Marcus and Yoshua Bengio can be submitted via the web form on [www.montreal.ai/aidebate](http://www.montreal.ai/aidebate)

MONTREAL.AI is grateful to Mila and to the collaborative Montreal AI Ecosystem. That being said, we will start the first segment.

Professor Marcus, you have 22 minutes for your opening statement.

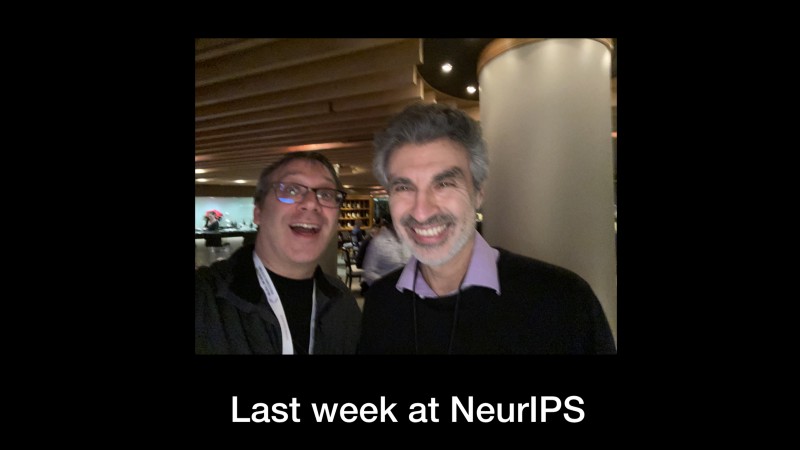
**Opening statement | Gary Marcus**

Opening statement | Gary Marcus — 22 min.

Thank you very much.

[stalling while waiting for video to start] And of course the AV doesn’t work. Hang on.

Before we started Yoshua and I were chatting about how AI was probably going to come before AV. He made some excellent points about his work on climate change and how if we could solve the AV problem it would actually be a good thing for the world.

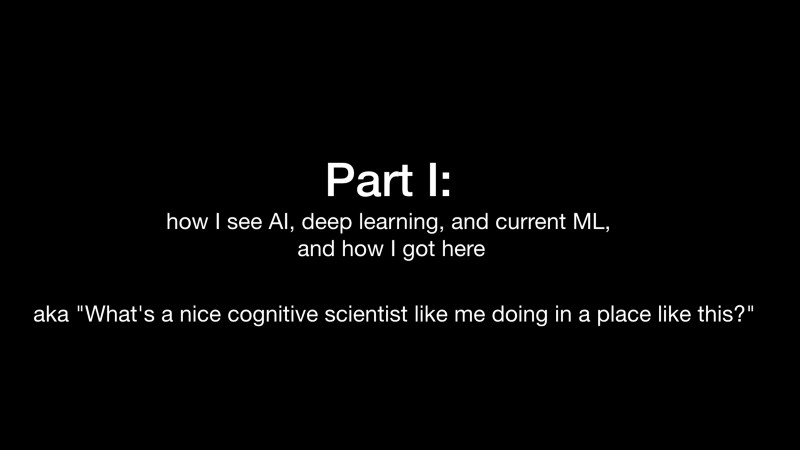
Last week at NeurIPS

So, this was Yoshua and I last week at NeurIPS at a party having a good time. I hope we can have a good time tonight. I don’t think either of us is out for blood but rather for truth.

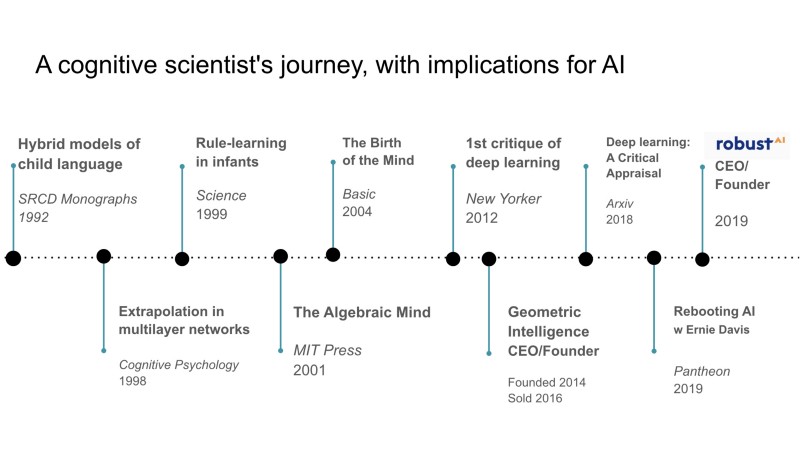
Overview

[Here’s] an overview of what I’m going to talk about today. I’m going to start with a bit of history and a sense of where I’m coming from.

I’m going to give my take on Yoshua’s view. I think there are more agreements than disagreements, but I think the disagreements are important and we’re here to talk about them, and then my prescription for going forward.

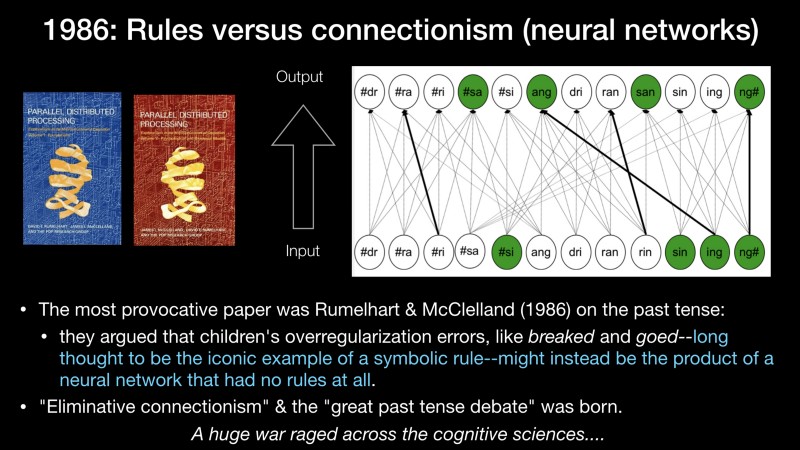
Part I: how I see AI, deep learning, and current ML, and how I got here

The first part is about how I see AI, deep learning and current machine learning and how I got here. It’s a bit of a personal history of cognitive science and how it feeds into AI. And, you might think: “*What’s a nice cognitive scientist like me doing in a place like Mila?*”.

A cognitive scientist’s journey, with implications for AI

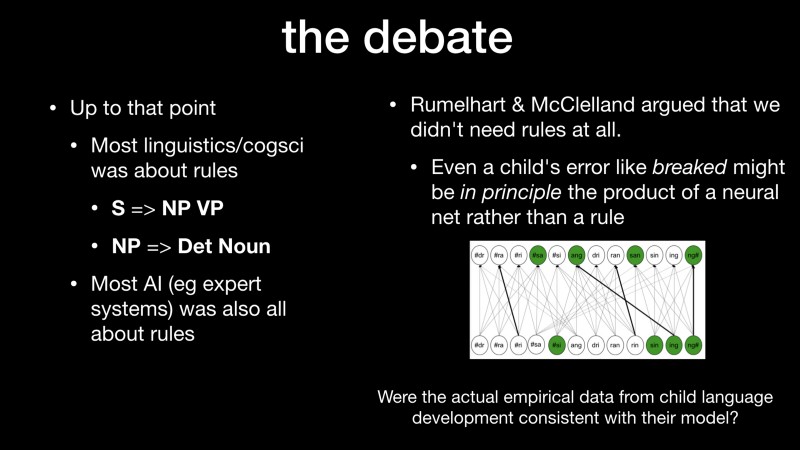
Here’s an overview, I won’t go into all of it, but there are some others thing that I have done that I think are relevant to AI. The important point is that I am not a machine learning person by training. I’m actually a cognitive scientist by training. My real work has been in understanding humans and how they generalize and learn. I’ll tell you a little bit about that work going back to 1992 and a little bit all the way up to the present.

But first, I’ll go back event a little bit before to a pair of famous books that people call the [PDP bible](https://mitpress.mit.edu/books/parallel-distributed-processing-volume-1). Not everybody will even know what PDP is but it’s a kind of ancestor to modern neural networks. Vince showed on and Yoshua will surely be talking about many and the one I have on the right is a simplification of a neural network model that tries to learn the English past tense.

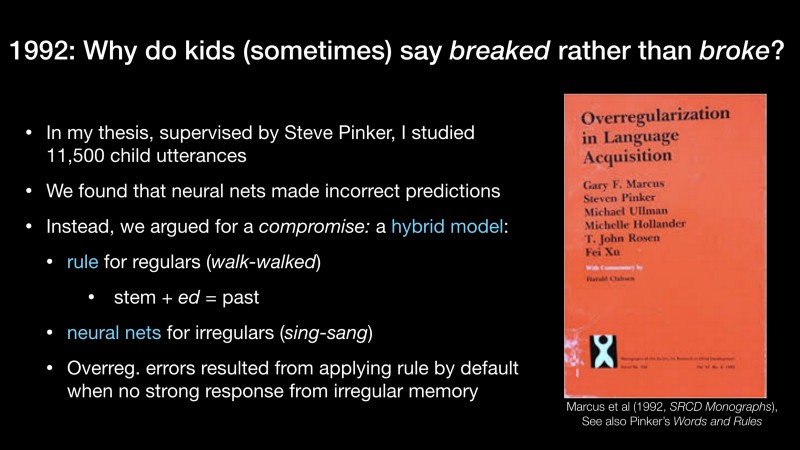
1986: Rules versus connectionism (neural networks)

This was part of a huge debate. In these two books I think the most provocative paper, certainly one that has stuck with me for 30 years, which is pretty impressive to have a paper to stuck with you for that long. It was a paper about children’s overregularization errors. So, kids say things like *breaked* and *goed* some of the times. I have two kids so I can testify that this is true. It was long thought to be an iconic example of symbolic rules. So, if you read any textbook until 1985, it would say: “*children learn rules*”. For example, they make these overregularization errors. And what Rumelhart and McClelland showed brilliantly was that you can get a neural net to produce these output without having any rule in it at all.

So, this created a whole field that I would call “*Eliminative Connectionism*”: using neural networks to model cognitive sciences without having any rules in it. And this so-called great past tense debate was born from this. And it was a huge war across the cognitive scientists.

the debate

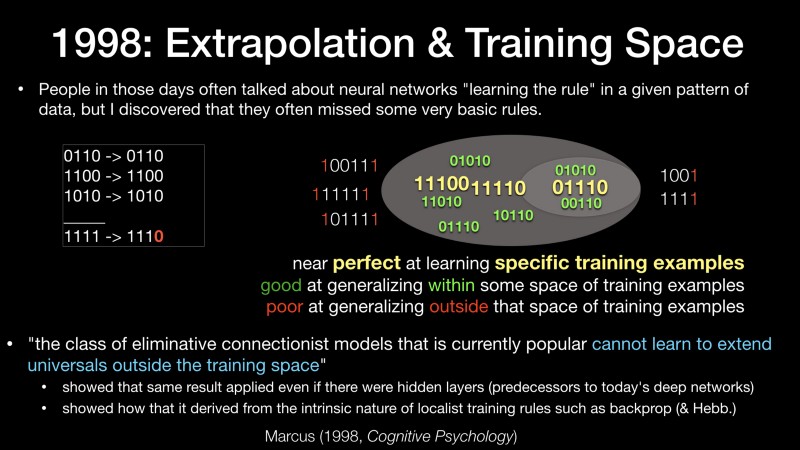
By the time I got to graduate school, it was all that people wanted to talk about. One the one hand up until that point, until that paper, most of linguistics and cognitive science was couched in terms of rules. So the idea was that you learn rules like a sentence is made of a noun phrase and a verb phrase. So, if you’ve ever read Chomsky, lots of Chomsky’s earlier works look like that. And most AI was also all about rules. So expert systems were mostly made up of rules. And here Rumelhart and McClelland argue that we don’t need rules at all, forget about it. Even a child’s error like *breaked* might be [the prdoct of a neural network without rules] *in principle*, [although] they didn’t prove it. But, they showed that in principle may be the product of a neural network where you have the input at the bottom and the output at the top and you tune some connections over time, might in principle give you generalization that looks like what kids were doing.

1992: Why do kids (sometimes) say *breaked*rather than *broke*?

On the other hand, they hadn’t actually looked at the empirical data. So I tried to myself, in graduate school, working with [Steve Pinker](https://en.wikipedia.org/wiki/Steven_Pinker) at MIT and what I looked at were these errors. I did, I think, the first big data analysis of language acquisition, writing shell scripts on Unix Sparcstations and looked at 11,500 child utterances.

The argument that Pinker and I made was that neural nets weren’t making the right predictions about generalization over time and particular verbs and so for. If you care, there’s a whole book that we wrote about it (Marcus et al (1992, *SRCD Monographs*), See also Pinker’s *Words and Rules*).

What we argued for was a compromise. We said it’s not all rules like [Morris Halle](http://linguistics.mit.edu/user/halle/) (he was on my thesis committee (phd)) liked to argue and we said it wasn’t all neural networks like Rumelhart and McClelland did. We said it was a hybrid model that best captures the data. A rule for regulars. So walk is inflected *walked* — you add to the “*ed*” for the past tense. Neural networks for the irregulars so this is why you say *sing — sang* but it might generalize to *spling — splang* that sound similar. And then the reason why children made overregularization errors we said is [that] the neural network didn’t always produces a strong response. If you have a verb that didn’t sound like anything you’ve heard before you’d fall back on the rules.

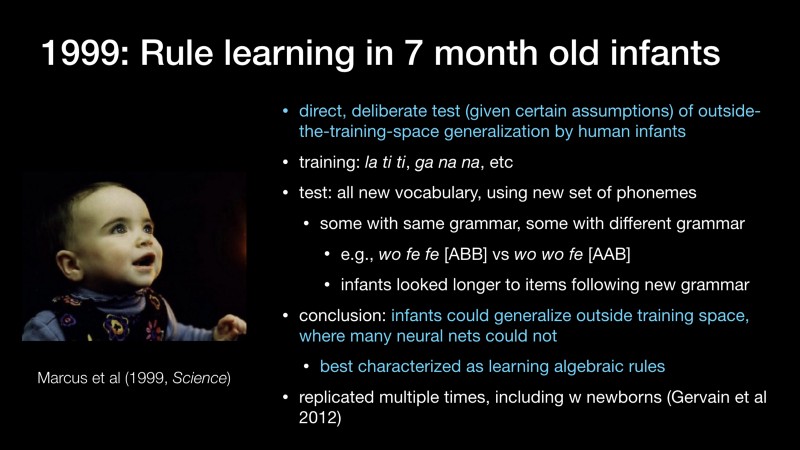
1998: Extrapolation & Training Space

So, that was the first time I argued for hybrid models back in the early 1990s. In 1998, or even a little bit before, I started playing a lot with the network models.

There’d been a lot written about them and I wanted to understand how they work and so I started implementing and trying them out. And, I discovered something about them that I thought was very interesting which is: people talked about then as if they learn the rules in the environment, but they didn’t always learn the rules. At least not in the sense that a human being might. So, here’s an example : if I taught you the function f(x) = x, or you can think of x = y + 0 or different ways to think about it. So, you have inputs like 0110, a binary number, and the output is the same thing and you do this on a bunch of cases then your neural net learns something but also makes some mistakes. So, if you give if an odd number, which is what I have here at the bottom, after giving it only even numbers, it doesn’t come up with the answers that a human being would. And so, I describe this in terms of something called the *training space*. So, let’s say the yellow examples are the things that you’ve been trained on, and the green ones are the things that are nearby in space of the one you’ve been trained on. The neural network generally did very well on the yellow ones and not so well on the ones that were outside the space.

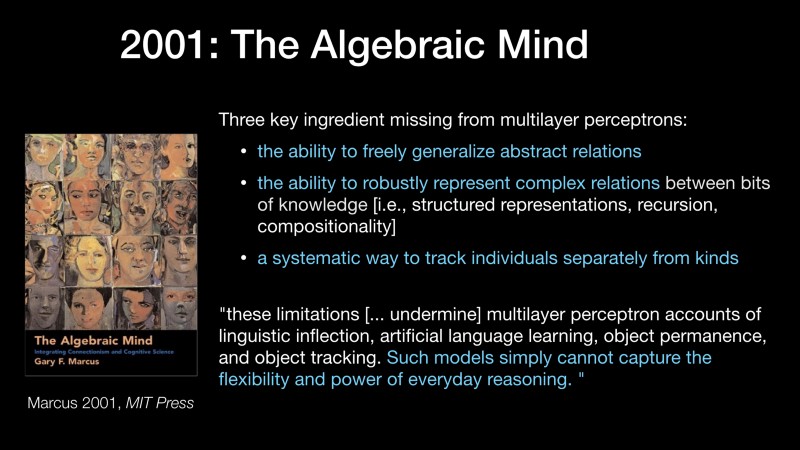
So, near perfect at learning specific examples, good at generalizing in the cloud of points around that, and poor at generalizing outside that space. I wrote up in Cognitive Psychology (Marcus (1998, Cognitive Psychology)), after having some battle with the reviewers (we can talk about it some time later), and the conclusion was that the classical limits of connectionists models that were [then] popular couldn’t learn to extend universals outside of the training space.

In my view is this is the thing that I’m the most proud of having worked on. Some details for discussion later [are on the slide]…

1999: Rule learning in 7 month old infants

This led me to some work on infants. What I’m trying to argue is that even infants could make this kind of generalizations that were steaming the neural networks of that day. So, it was a direct deliberate test on the outside of training space generalization by human infants. So, the infants would hear sentences like “*la ti ti*” and “*ga na na*” (I read theses to my son yesterday and he think these are hilarious, he is almost 7) and then we tested on new vocabulary. There will be sentences like “*wo fe fe*” or “*wo wo fe*”. So one of those has the same grammar that the kids has seen before and the other one has a different grammar. Because all the items were new you couldn’t use some of the more statistical techniques that people thought about like transitional probabilities and it was a problem for early neural networks.

The conclusion was infants could generalize outside training space, where many neural nets could not. And I argued that this should best characterized as learning algebraic rules. It has been replicated a bunch of times and it led to my first book which is called “*The Algebraic Mind*”.

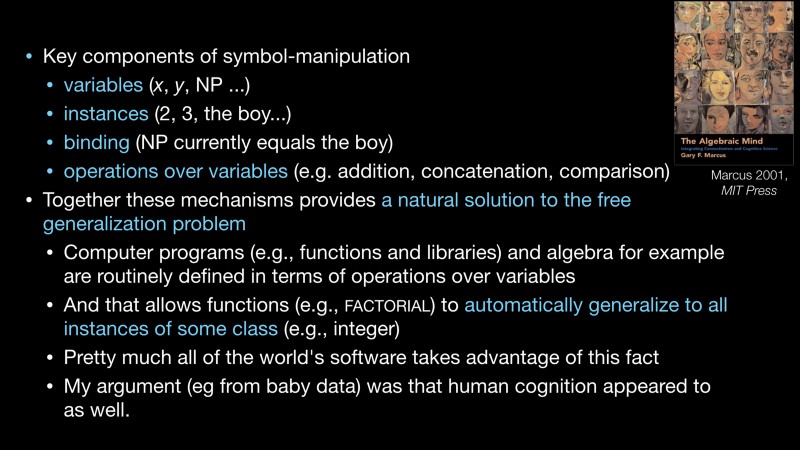
2001: The Algebraic Mind

The idea was that humans could do this kind of abstractions. I argued that there was three key ingredient missing from multilayer perceptrons:

1. the ability to freely generalize abstract relations as the infants were doing
2. the ability to robustly represent complex relations like the complex; structure of a sentence; and
3. a systematic way to track individuals separately from kinds.

We will talk about the first two today and probably not of the third. And I argued that this undermine a lot of attempts to use multilayer perceptrons as models of the human mind.

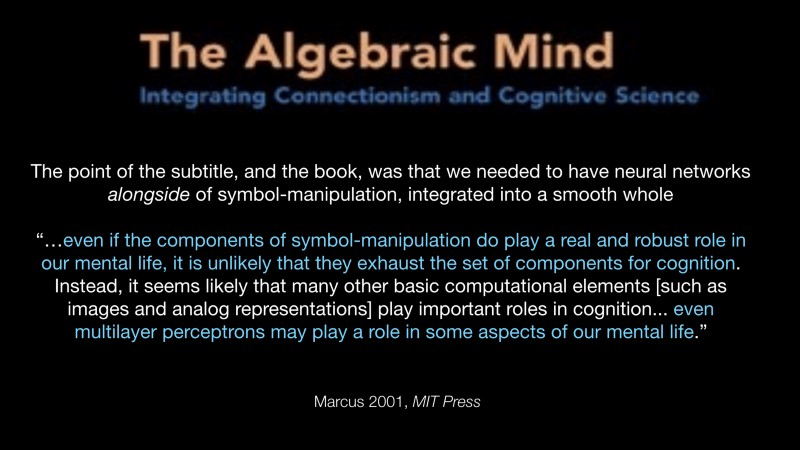
I wasn’t really talking about AI, I was talking about cognition. Such models, I argued, simply can’t capture the flexibility and power of everyday reasoning.

2001: symbol-manipulation

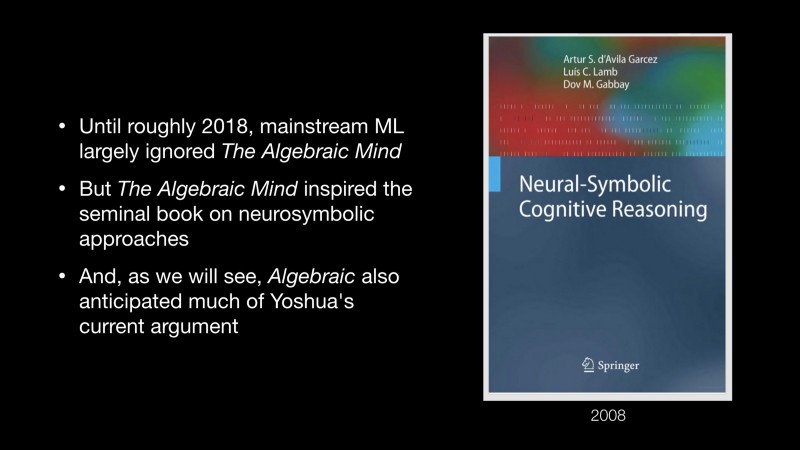
And the key component of the thing I was defending, which I called symbol-manipulation (I didn’t invented it, but I tried to explicated it and argue for it), are *variables*, *instances*, *bindings* and *operations over variables*. You can think in algebra where you have a variable like x, you have an instance of it like 2, you bind it so you say right now x = 2 or my name phrase currently equals the boy, and then you have operations over variables so you can add the together, you can put them together (concatenation, if you know computer programming), you can compare them, and so for…

Together, these mechanisms provides a natural solution to the free generalization problem. So, computers programs do this all the time. You have something like the factorial function (if you’ve ever taken computer programming) and it automatically generalize to all instances of some class, let say integers, once you have that code.

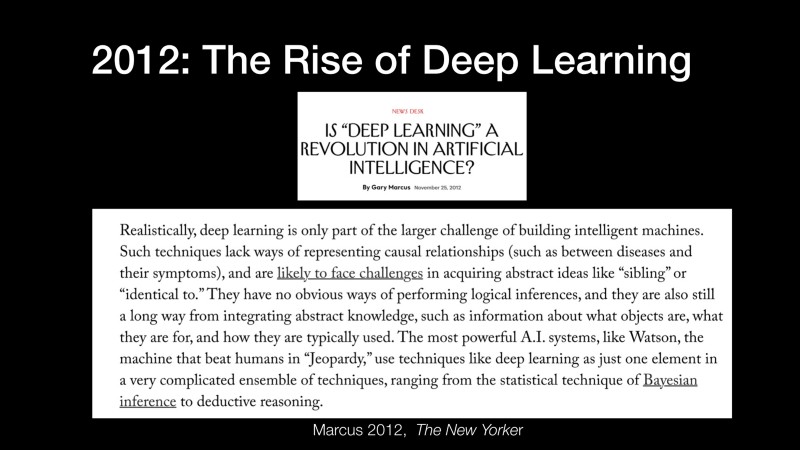
Pretty much all of the world’s software takes advantage of this fact. My argument (eg from baby data) was that human cognition appeared to as well innately.

The Algebraic Mind

The subtitle of that first book (you can’t see it that well here), was “t connectionism and cognitive science”. I wasn’t trying to knock down neural networks and say forget about it. I was saying, let’s take the insights of those things, they’re good at learning, but let’s put it together with the insights of cognitive science, a lot of which has been using these symbols and so for. And so [here] I said, “even if I’m right the symbol manipulation plays an important role in mental life”, it doesn’t mean we shouldn’t have others things in there too, like multilayer perceptrons which are the predecessors of todays deep learning.

Neural-Symbolic Cognitive Reasoning

I was arguably ignored, I think in candor, until a year or so ago. People I think started paying attention to the book again. But, it did inspire a seminal book on neuro-symbolic approaches which I hope some people will take a look at, called [*Neuro-Symbolic Cognitive Reasoning*](https://www.springer.com/gp/book/9783540732457) and I’m going to try to suggest that it also anticipated some of Yoshua’s current arguments.

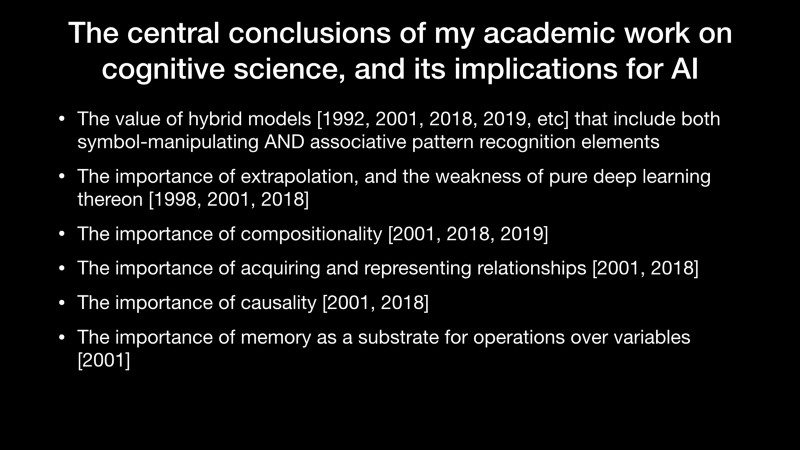
2012: The Rise of Deep Learning

I stopped working on these issues. I started working on innateness, I learned to play guitar (that’s a story for another day) and didn’t talk about these issues at all until 2012 when Deep Learning became popular again. The front page story of the New York Time was about Deep Learning and I thought I’ve seen this movie before and I was writing for the New Yorker at the time and I wrote a piece and I said: “*Realistically, deep learning is only part of larger challenge of building intelligent machines. Such techniques lack ways of causal relationships.*(A lot of discussion about that today). *They have no obvious way of performing logical inference, and they are still a long way from integrating abstract knowledge.*” And, I once again argued for hybrid models. Deep Learning is just one element in a very complicated set of machinery.

A screenshot of a cell phone

Description automatically generated2018: Critique of deep learning

Then, in 2018, Deep Learning got more and more popular but I thought some people were missing some important points about it, so I wrote a piece (I was actually here in Montreal when I wrote it) called “[*Deep Learning: A Critical Appraisal*](https://arxiv.org/abs/1801.00631)). It outlines ten problems for Deep Learning (I think it was on the suggested [readings](http://www.montreal.ai/aidebate.pdf) for tonight’s event) and the failure to extrapolate beyond this space of training was really at the heart of all of those things. I got a ton of flak on Twitter (you can go back and search and see some of the history). I felt like I was often misrepresented as saying “*we should throw away Deep Learning*”, which is not what I was saying. And I was careful enough in the paper to say it in the conclusion: “*Despite all of the problems I have sketched, I don’t think we need to abandon Deep Learning… (which is the best technique we have for training neural networks right now) but, rather, we need to reconceptualize it not as an universal solvent but simply as one tool among many*”.

The central conclusions of my academic work on cognitive science, and its implications for AI

So, the central conclusions of my academic work concluded the value of hybrid models, the importance of extrapolation, of compositionality, acquiring and representing relationships, causality and so for.

**Part II: Yoshua**

Part II: Yoshua

Some thoughts on his views, and how I think they have changed a bit over time, a little bit on how I feel misrepresented and how our views are and not similar.

A screenshot of a social media post

Description automatically generatedFirst things first: I admire Yoshua

The first thing I want to say is that I really admire Yoshua. For example, I wrote a piece recently, skewering the field for hype. And I said, but you know, a really good talk is one by Yoshua Bengio: a model of being honest about limitations. I also love the work that he’s doing for example on climate change and machine learning. I really think he should be a role model in his intellectual honesty and in his sincerity to make the world a better place.

A screenshot of a cell phone

Description automatically generatedMy differences are mainly with Yoshua’s *earlier* (e.g., 2014–2015) views

My differences with him are mostly about his earlier views. We first met here in Montreal five years ago and at that time I don’t think we had much common ground. I thought like he was putting to much faith in black box deep learning systems, he rely to heavily on larger datasets to yield answers and he’ll talk about *system 1* and *system 2* later, I guess I will as well. I Felt he was all on the system 1 side and not so much on the system 2 side.

And, I went back and talked to some friends about that. I lot of people remember the talk he gave in 2015 to a bunch of linguists who didn’t like Yoshua’s answer to questions like “*how would we deal with negation or quantification words like every*” and what Yoshua did was to say we just need more data and the network will figure it out.

If Yoshua was still in this position, which I don’t think he is, I think we would have a longer argument.

A screenshot of a social media post

Description automatically generatedRecently, however Yoshua has taken a sharp turn towards many of the positions I have long advocated

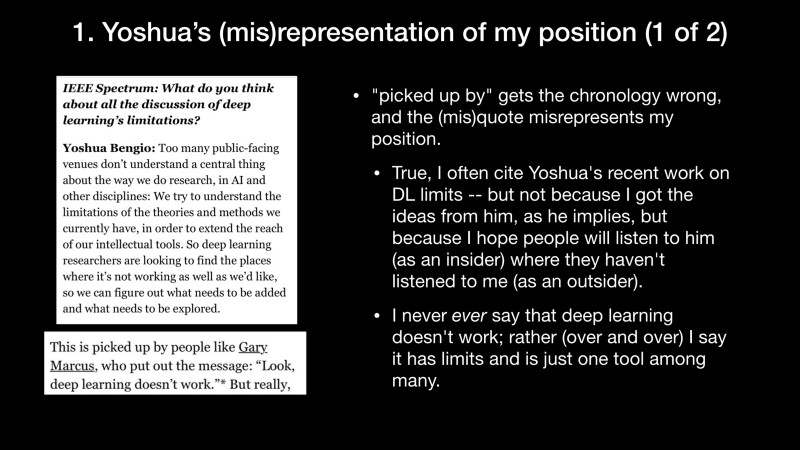
Recently, however Yoshua has taken a sharp turn towards many of the positions I have long advocated for: acknowledging fundamental limits on deep learning, need for hybrid models, the critical importance of extrapolation and so for. I have some slides in camera shots that I took at his recent talk at NeurIPS that I think show a very interesting convergence here.

Disagreements

So, disagreements now.

I’ll take about my position, the right way to build hybrid models, innateness, the significance of the fact that the brain is a neural network and what we mean by compositionally.

And, that’s it, we actually agree about most of the rest.

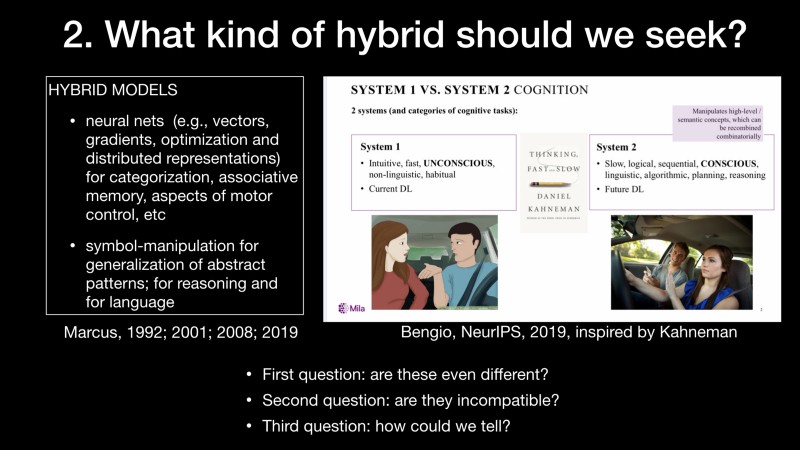
1. Yoshua’s (mis)representation of my position (1 of 2)

The first one is the most delicate. But, I think occasionally Yoshua is misrepresenting me as saying “*look, deep learning doesn’t work*”, he said that to IEEE Spectrum. I hope I persuaded you that this is not actually my position. I think deep learning is very useful. However, I don’t think it solves all problems.

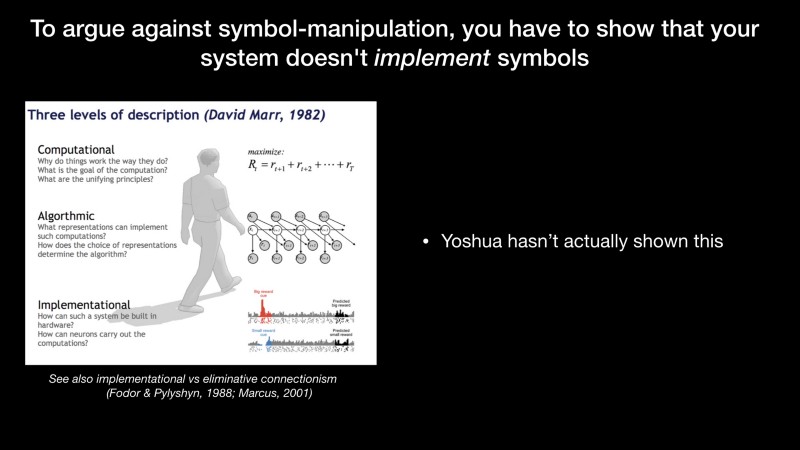
A screenshot of a cell phone

Description automatically generated1. Yoshua’s (mis)representation of my position (2 of 2)

The second thing is: his recent work has really narrowed what I think is the most important point, which is the trouble deep nets have in extrapolating beyond the data and why that means for example we might need hybrid models. I would like for him to cite me a little bit. I think not mentioning me devalues my contribution a little bit and further represents my background in the field.

2. What kind of hybrid should we seek?

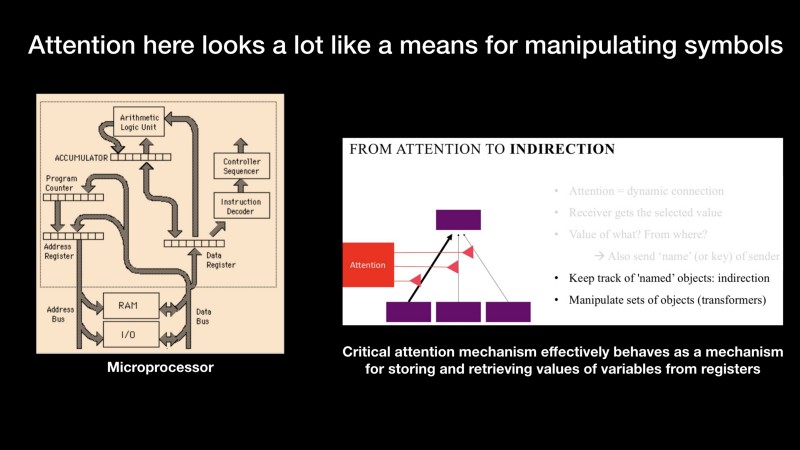
What kind of hybrid should we seek? I think Yoshua was very inspired by Daniel Kahneman’s [book](https://en.wikipedia.org/wiki/Thinking,_Fast_and_Slow) about system 1 and system 2 and I imagine many people in the crowd did read it. You should if you haven’t. That talks about one system that is intuitive, fast and conscious and another who is slow, logical sequential and conscious. I actually this that this is a lot like what I’ve been arguing for a long time. We can have some interesting conversation about the differences. There are questions : are these event different? Are they incompatible? How could we tell?

To argue against symbol-manipulation, you have to show that your system doesn’t *implement* symbols

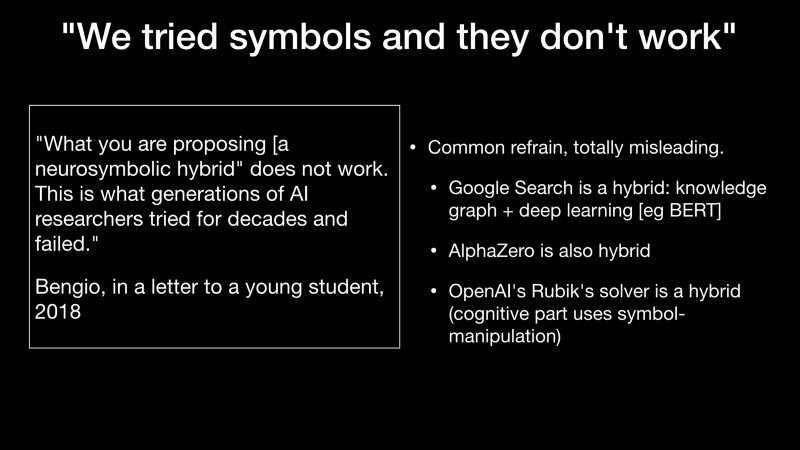
I want to remind people of what I think is the most important distinction drawn in cognitive sciences, which is by the late [David Marr](https://en.wikipedia.org/wiki/David_Marr_%28neuroscientist%29), who talked about having computational algorithmic and implementational levels. So, you can take some abstract algorithm or notion like I’m going to do a sorting algorithm. You can pick a particular one like the bubble sort. And then you can make it out of neurons, you can make it out of silicons, you can make it out of Tinkertoy.

I think we need to remember this: we have this conversation, so we want to understand the relation about how we’re building something and what algorithm is being represented. I don’t think Yoshua [has] made that argument yet. Maybe he will today.

I think that this is what we would need to do if we want to make a strong claim that a system doesn't implement symbols.

Attention here looks a lot like a means for manipulating symbols

Yoshua has been talking a lot lately about *attention*. I think that what he is doing with attention reminds me actually of a microprocessor in the way that it pulls things out of a register and moves them in to the register and so forth. In some ways it seems as it behaves at least a lot like a mechanism for storing and retrieving values of variables from registers, which is really what I’ve cared about for a long time.

“We tried symbols and they don’t work”

Then, I’ve seen some arguments from Yoshua against symbols. Here’s something in an email he sent to a student, he wrote: “*What you are proposing [a neuro-symbolic hybrid] does not work. This is what generations of AI researchers tried for decades and failed.*” I’ve heard this a lot, not just from Yoshua, but I think it is misleading. The reality is that hybrids are all around us. The one you use the most probably is Google search which is actually a hybrid between a knowledge graph, which is classic symbolic knowledge, and deep learning like a system called BERT. Alpha Zero, which is the world champion (or it was until recently) is also a hybrid.

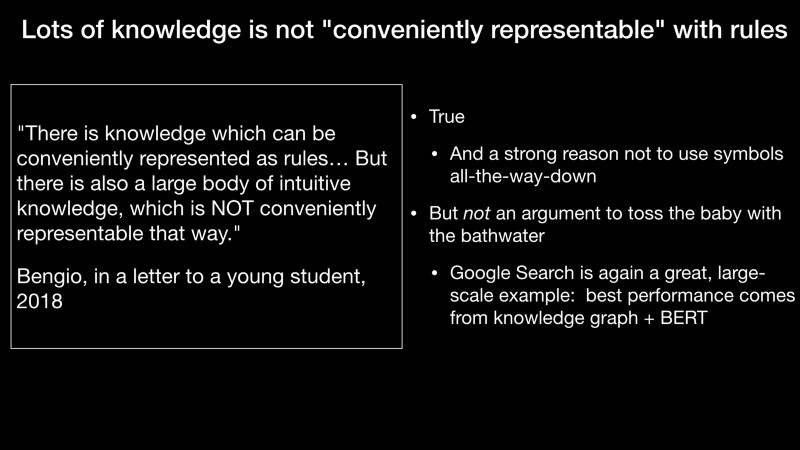
Vincent Boucher: Professor Marcus, you have 5 more minutes.

OpenAI’s Rubik’s solver is a hybrid.

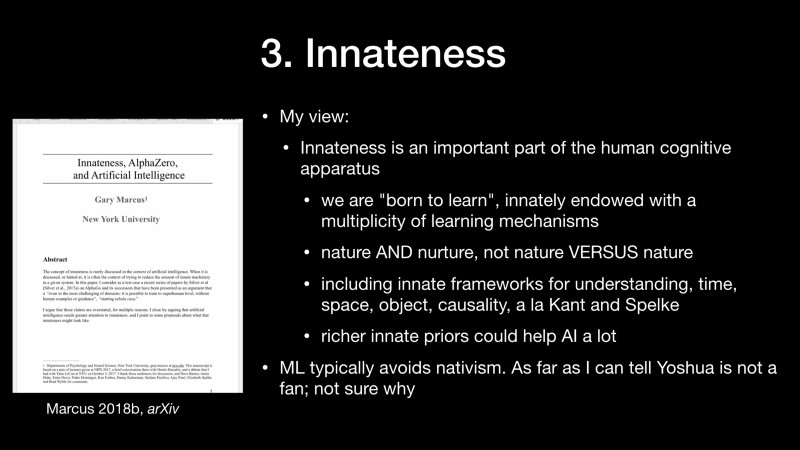
A screenshot of a cell phone

Description automatically generatedMao et al, *arXiv* 2019

There is great [work](https://arxiv.org/abs/1904.12584) by Joshua Tenenbaum and Jiayuan Mao that is also a hybrid that just came out this year.

Lots of knowledge is not “conveniently representable” with rules

Another argument that Yoshua has given is that lots of knowledge is not coveniently represented with rules. It is true, some of it is not conveniently represented with rules and some of it is. Again, Google search is a great example where some is represented with rules and some is not it is very effective.

3. Innateness

The third argument, and I don’t fully know Yoshua’s view, is about nativism. So, as a cognitive development person, I see a lot of evidence that a lot of things are built-in in the human brain. I think that we are born to learn and we should thank about it as “nature and nurture” rather that “nature vs nurture”.

I think we should think about a innate framework for understanding things like time and space and causality as Kant argued for in the *Critique of Pure Reason* and Spelke argued for in her cognitive development work.

The argument that I’ve made in the paper here on the left, is that richer innate priors might help artificial intelligence a lot. Machine learning has historically typically avoided nativism of this sort. As far as I can tell, Yoshua is not a huge fan of nativism and I’m not totally sure why.

A screenshot of a cell phone

Description automatically generated

Here is some empirical data showing that nativism and neural networks works. It comes form a great paper by Yann LeCun in 1989 where he compare four different models. The ones that had more innateness in terms of convolutional prior were the ones that did better.

A picture containing ground

Description automatically generated

This is a video of a baby ibex climbing down a mountain. I don’t think the anybody can reasonably say that there is nothing innate about the baby ibex: it has to be born with an understanding of the 3 dimensional world and how it interacts and so for in order to do the things that it does. So nativism is plausible in biology and I think we should use more of it in AI.

A screenshot of a cell phone

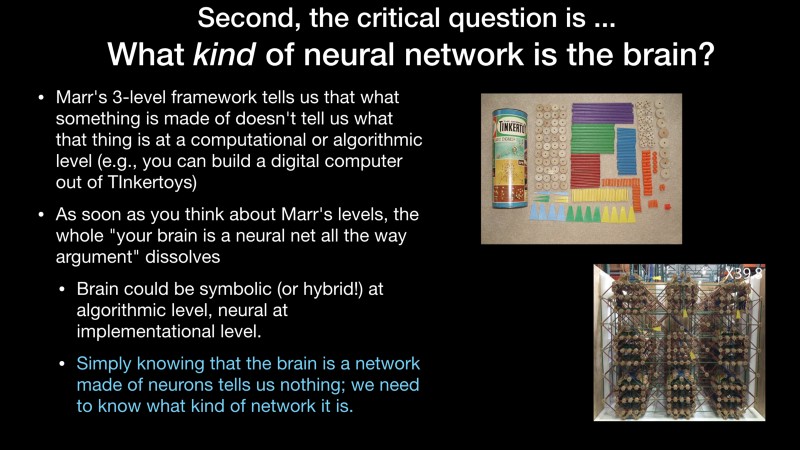
Description automatically generated4. Brains and neural networks

Some of you may know that there was actually a cartoon about this debate by Dileep George, he is worth looking up on Twitter ([@dileeplearning](https://twitter.com/dileeplearning)). And, in the cartoon version of the debate, Yoshua wins by saying “*your brain is a neural net all the way :-)*”. And everybody was: wow, I guess Yoshua was right after all. And Yoshua did [say] the same argument to me on Facebook by saying: *your brain is a neural net all the way*.

A screenshot of a newspaper

Description automatically generatedFirst, deep nets aren’t much like brains

Or course, deep neural networks aren’t much like brains. I’ve been arguing that for a while. There are many cortical areas, many neuron types, many different proteins and synapses and so for and so on. I actually heard Yoshua made the same arguments at NeurIPS 2019 last week and I think we pretty much agree about that.

What *kind* of neural network is the brain?

He made a beautiful argument with degrees of freedom [and neurons] in particular — I loved it! But, the critical question is really what kind of neural network is the brain? So, going back to Marr’s distinction, you could build anything you want, any computation, out of Tinkertoys choice or out of neurons. We really want to know wether the brain is a symbolic thing at the algorithmic level or not and then we ask how is this implemented in neurons. Simply knowing that the brain is a network made of neurons doesn’t tells that much; we really need to know what kind of network it is.

A screenshot of a cell phone

Description automatically generated“Symbols aren’t biologically plausible”

There is another argument people says: “*Symbols aren’t biologically plausible*”. I think that this is a ridiculous argument. When my son learned long division last week and followed an algorithm, he was surely manipulating symbols. We do at least some symbol manipulation some of the time. And, back in the 80s people knew this and they said that symbols were the domain of conscious rules processing, they’re just not what we do unconsciously. Pinker and I said that language isn’t that conscious and we use symbols in language too. The real question is not whether the brain is a neural network, it’s how much of it involves symbolic as opposed to other processes.

A screenshot of a cell phone

Description automatically generatedEven if somehow turned that the *brain* never manipulated symbols, why exclude them from AI?

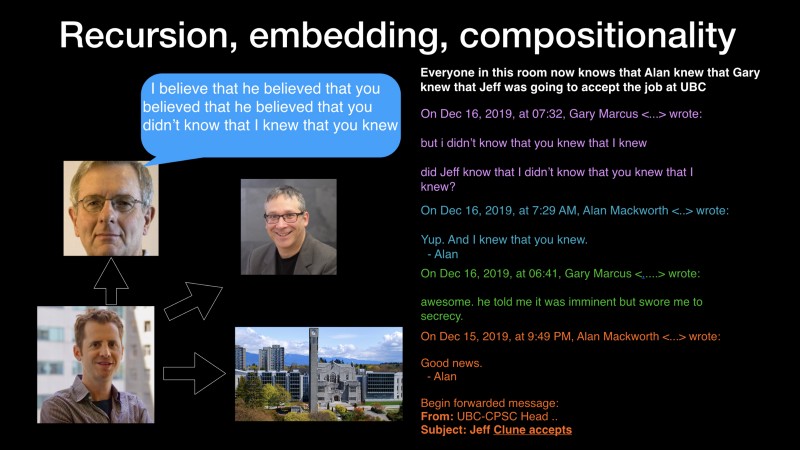
Even if somehow turned that the brain never manipulated symbols (which is counterfactual to our world), why exclude them from AI? We can’t prove that they are inadequate, they have proven utility: a large fraction of the world’s computers programs are written in (pure) symbol-manipulating code and a large fraction of the world’s distilled knowledge comes in the form of symbols: eg. most of Wikipedia is in written, symbolic form and we want to leverage that in our learning systems.

A screenshot of a cell phone

Description automatically generated5. Compositionality

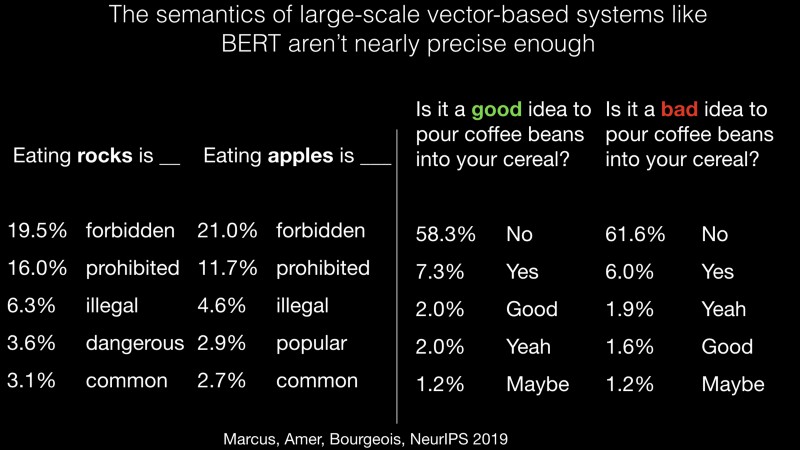
Five: Compositionality. Yoshua has been talking a lot about compositionality and I think he will tonight. I think he means something different than I mean by it. I’ll let him give its description later, but I think it is partly by putting together different pieces of networks and so for.

I’m really interested in the linguist sense which is how you put differents parts of sentences together into larger wholes.

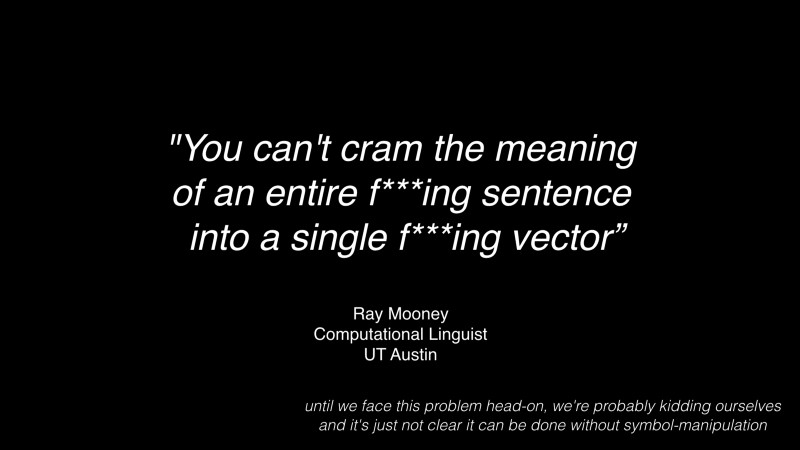
Recursion, embedding, compositionality

Here’s a good example. Last week, my friend Jeff Clune, I’ve beed encouraging him to come to UBC and I’ve been encouraging to hire him form a job and my friend Alan Mackworth said “*Good news, Jeff Clune accepted*”. So, I wrote back “awesome. he told me it was imminent but swore me to secrecy.”

Alan said yes, I knew that you knew and eventually we get everyone in this room now knows that Alan knew that Gary knew that Jeff was going to accept the job at UBC.

The semantics of large-scale vector-based systems like BERT aren’t nearly precise enough

I don’t think we can represent that with today’s neural networks. We can barely get a system to represent the difference between eating rocks and eating apples.

“You can’t cram the meaning of an entire f\*\*\*ing sentence into a single f\*\*\*ing vector”

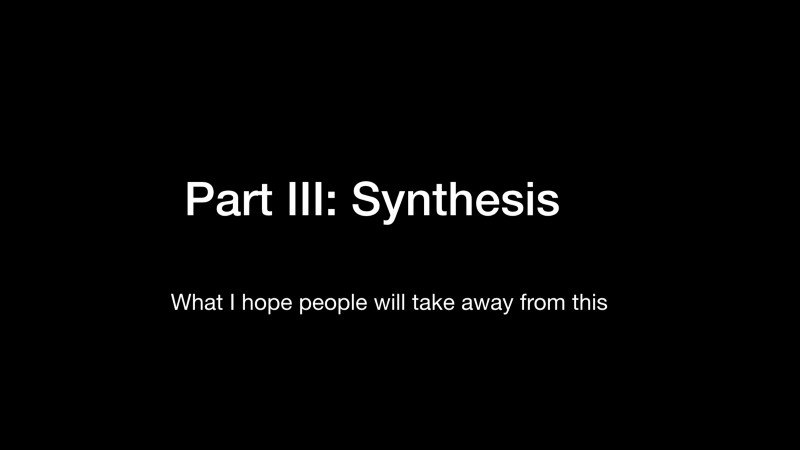
And, this famous quote: “*You can’t cram the meaning of an entire f\*\*\*ing sentence into a single f\*\*\*ing vector*” I think still stands.

A screenshot of a cell phone

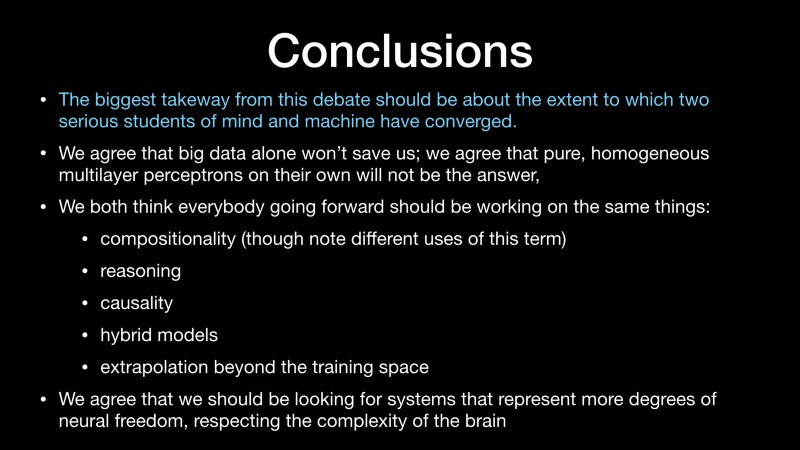
Description automatically generatedCompositionality isn’t just about language…

Compositionality isn’t just about language... It’s also learning about different concepts and putting them together in different ways. Here are my kids inventing a new game. Ten minutes later, they’ve combined things that they know. Children can learn something in a few trials and we haven’t figured out how to do that yet.

**Part III: Synthesis**

Part III: Synthesis

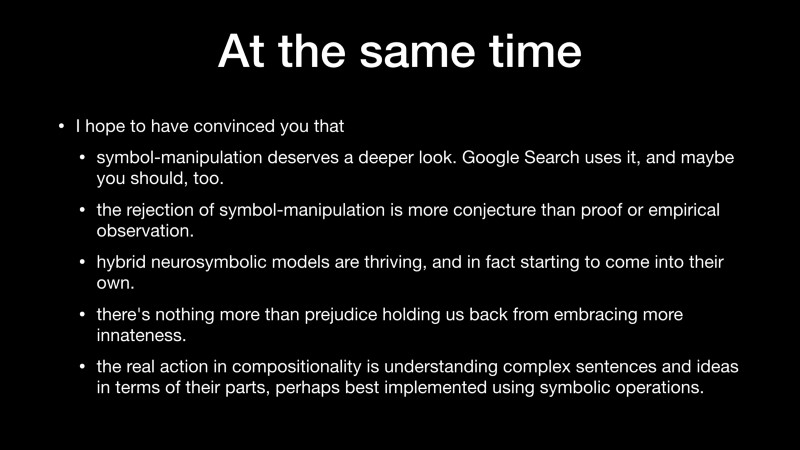
What I hope people will take away from this.

Conclusions

The biggest takeaway from this debate should be about the extent to which two serious students of mind and machine have converged. We agree that big data alone won’t save us; we agree that pure, homogeneous multilayer perceptrons on their own will not be the answer,

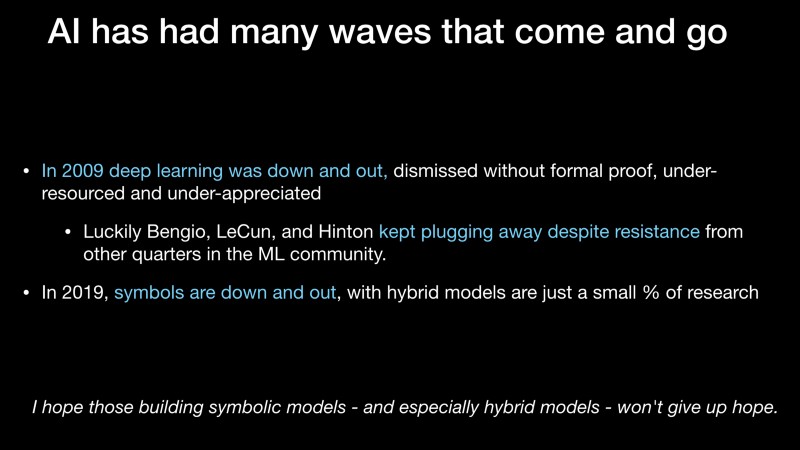
We both think everybody going forward should be working on the same things:  
1. — compositionality  
2. — reasoning   
3. — causality  
4. — hybrid models  
5. — extrapolation beyond the training space

We agree that we should be looking for systems that represent more degrees of neural freedom, respecting the complexity of the brain.

At the same time

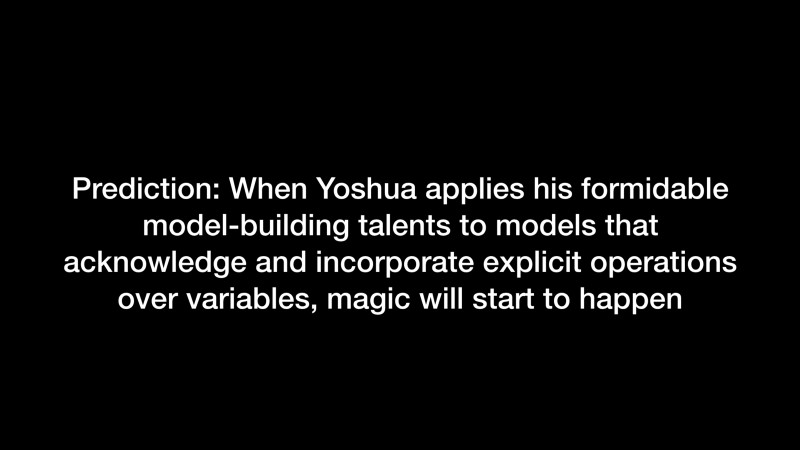
At the same time, I hope to have convinced you that

1. symbol-manipulation deserves a deeper look. Google Search uses it, and maybe you should, too.
2. the rejection of symbol-manipulation is more conjecture than proof or empirical observation.
3. hybrid neurosymbolic models are thriving, and in fact starting to come into their own.
4. there’s nothing more than prejudice holding us back from embracing more innateness.
5. the real action in compositionality is understanding complex sentences and ideas in terms of their parts, perhaps best implemented using symbolic operations.

AI has had many waves that come and go

AI has had many waves that come and go. In 2009 deep learning was down and out. A lot of people dismissed it. I have a friend who saw Geoffrey Hinton give a talk and only one person came (a poster, excuse me).

Luckily Bengio, LeCun, and Hinton kept plugging away despite resistance from other quarters in the ML community. I hope people doing symbols will keep plugging away.

Prediction: When Yoshua applies his formidable model-building talents to models that acknowledge and incorporate explicit operations over variables, magic will start to happen

Here’s my prediction and my last slide: When Yoshua applies his formidable model-building talents to models that acknowledge and incorporate explicit operations over variables, magic will start to happen.

Thank you very much.

Vincent Boucher: Thank you, professor Marcus. Professor Bengio, you have 22 minutes for your opening statement.

**Opening statement | Yoshua Bengio**

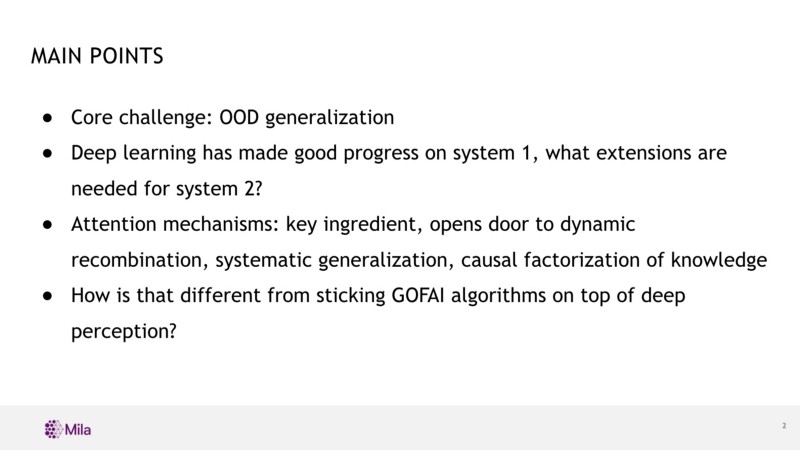
Opening statement | Yoshua Bengio — 20 min.

Welcome to this debate.

A screenshot of a tree

Description automatically generatedDebate with Gary Marcus

Thanks Marcus for setting up and talking first. I took a lot of notes.

MAIN POINTS

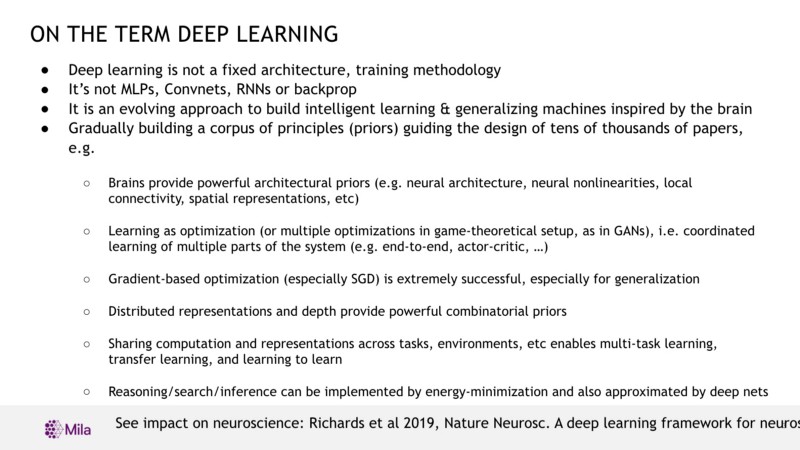
The main points I want to make…

I want to talk about *out of distribution generalization*, which is connected to some other things that Marcus talked about. Which, I think is more than the notion of extrapolation. I’ll get back to that.

I want to talk about my views on how *deep learning might be extended to dealing with system 2* computational capabilities rather than taking the old techniques and combining them with neural nets.

I want to talk briefly about *attention mechanisms* and why these might provide some of the key ingredients that Gary has been talking about that make symbolic processing able to do very interesting things. But, how we could do it within a neural net framework?

I’ll contrast that with some of the more symbolic approaches.

On the term deep learning

I want to get out of the way a few things about about the term “*deep learning*”, because there’s a lot of confusion. Specially, when deep learning is a straw man, it tends to be used to mean MLP from 1989, just like Gary used the term just a few minutes ago. If you open the last NeurIPS proceedings, you will see that it is much more than that.

Deep learning is really not about a particular architecture or even a particular training procedure. It is not about backprop, it is not about *Convnets*, *RNNs* or *MLPs*. It is something that is moving. It is more of a philosophy that is expanding as we add more principles to our toolbox to understand how to build machines that are inspired by the brain in many ways and use some form of optimization (usually a single objective, but sometimes multiple objectives like in GANs). In general, there is a coordinated optimization of multiple parts taking advantage of some of the earlier ideas of the 80s of course, like distributed representations, but also of more moderns ideas, like depth of representations. Also, taking advantage of sharing computation and representations across tasks, environments, etc., enables multi-task learning, transfer learning, learning to learn, and so on.

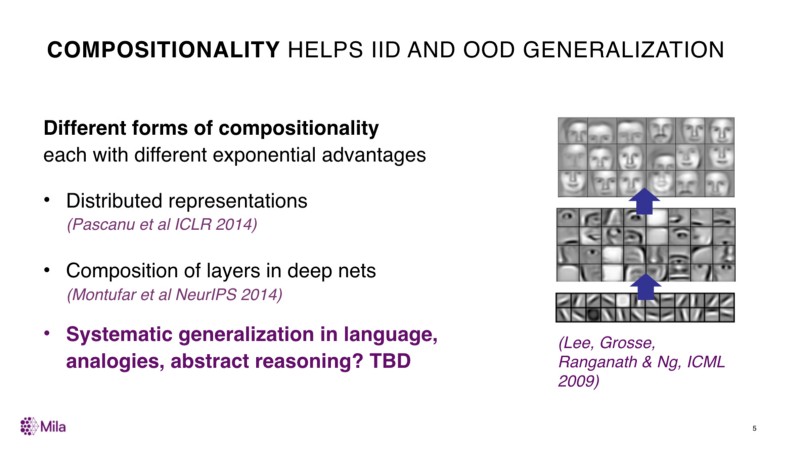
As I will argue, I think, tools to move forward includes things like *reasoning*, *search*, *inference* and *causality*.

To connect to neuroscience, there is actually a very rich set of works happening in the last few years connecting again the modern deep learning research with neuroscience. We had a paper just published in Nature Neuroscience called “[*A deep learning framework for neuroscience*](https://www.nature.com/articles/s41593-019-0520-2)”, but I won’t have the time to talk about it today.

Agent Learning Needs OOD Generalization

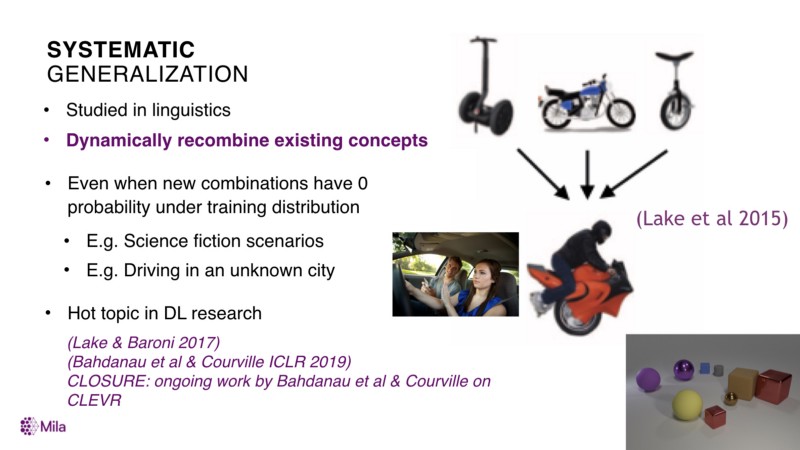
*Out of distribution generalization* means something different from the normal form of generalization where we have data from one distribution and we worry about generalizing to examples from the same distribution.

When we talk about extrapolation, Gary, it is not clear wether we’re talking about generalizing to new configurations coming from the same distribution so we have to thing about the notion of distribution in order to make a difference. For agents in the world, this is very important because what they see changes in nature because of interventions of agents because of moving in time and space and so on.

**Compositionality** helps iid and ood generalization

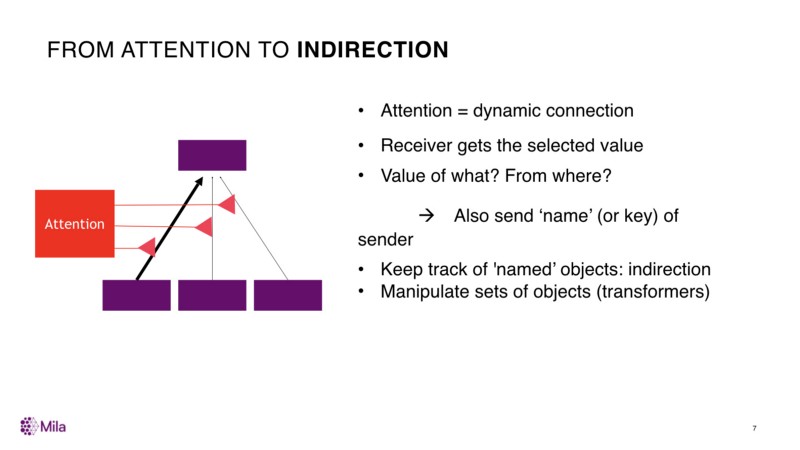
What I have been arguing for a little bit now, certainly much less than Gary, is the importance of *compositionality*. But, one of the thing I’ve done in the 2000s is to try to help figure out why event the current neural nets, the ones from the 80s with distributed representations, have a powerful form of compositionality. I’m not going to go in the details of that, but this dates from about five years old. And, similarly, why composing layers brings in a form of compositionality.

Basically, my argument is, we have these two forms already of compositionality in the neural nets. We can incorporate the form that Gary likes to talk about and I like to talk about these days, which is inspired a lot by the work of linguists. But, I think that it is more powerful and more general than just about language and something we use in conscious reasoning for example.

Systematic Generalization

Basically, what it is about, is how one might combine existing concepts in ways that may have zero probability under the training distribution (it is not just that it is a novel pattern. It is one that may be unlikely under the kind of distributions we’ve seen. Yet, our brain is able to come up with these interpretations, these novel combinations and so on). At NeurIPS, I gave this example of driving in a new city where you have to be a little bit creative and combining the skills you know and in others ways in order to solve a difficult problem.

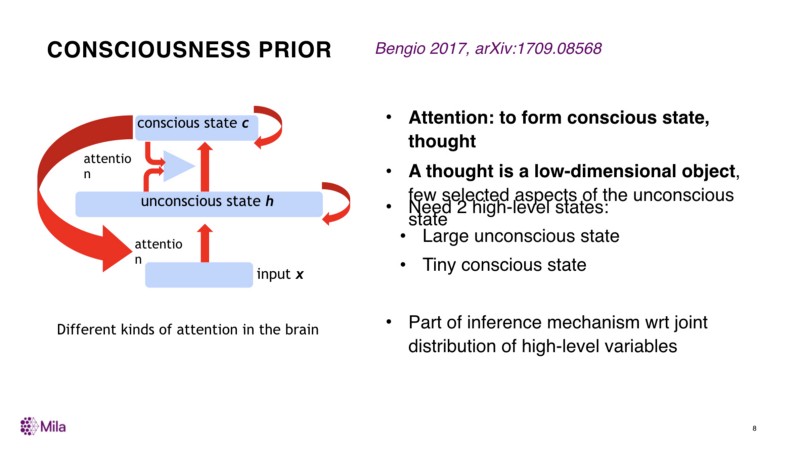
This issue is not new in deep learning in the sense that people have been thinking at least for a few years. Actually, I would say it is one of the hottest area in deep learning. We haven’t solved it, but I think people are starting to understand it better. One of the ingredient which I and others have been thinking as crucial in this exploration is *attention*.

From Attention to Indirection

Attention is interesting because it changes the very nature of what a standard neural net can do in many ways. It creates dynamic connections that are created on the fly based on context. It is even more context dependent, but in a way that can favour what Gary called *free generalization* that I think is important in language and in conscious processing.

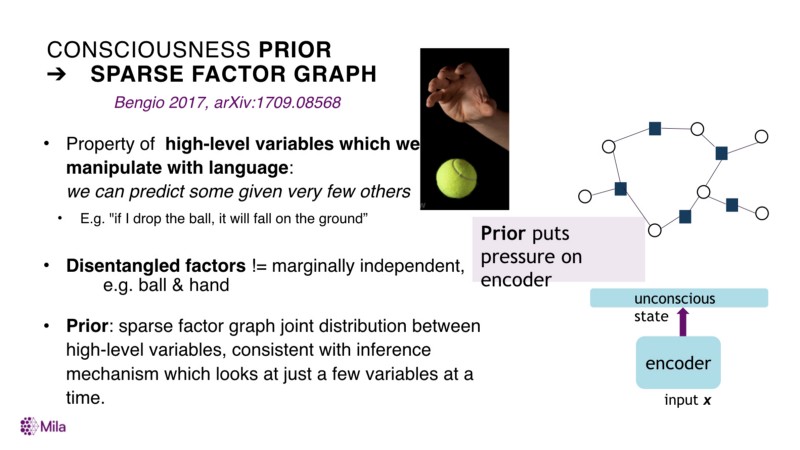
Why is that? Attention selects an element from a set of elements in the lower layer. It selects this element in a soft way (at least in the soft attention kind that we do in deep learning typically). The receiver gets a vector, but it doesn't know where that vector comes form. In order to really do their job, it is important for the receiver to get information not only about the value which is being sent, but also where it comes from. The where is sort of a name. Now, it is not like a symbolic name. We use vectors (what we call keys in transformers for example). You can think of these as neural net forms of reference because that information can be passed along and be used again to match some elements or some other elements to perform further attention operations.

This also changes neural nets from vectors processing machines to sets processing machines. This is something Gary talked about in his earlier interventions and that I think that is important for conscious processing.

Consciousness Prior

I have been talking a lot about consciousness in the last couple of years. There is of course a much richer volume of research in cognitive neurosciences about consciousness. The way that I’m trying to look at this is how we can frame some of the things that have been discussed in cognitive science and in neuroscience about consciousness and about other aspects of high level processing and frame them as *priors*, either structural or regularizers, for building different kinds of neural nets.

One of these priors is what I called the *Consciousness Prior*. It is implemented by attention, which selects a few elements of an unconscious state into a smaller conscious state.

Consciousness Prior ➔ sparse factor graph

In terms of priors, what it means, is that instead of knowledge being in a form where every variable can interact with every variable, what this would entail is that at that high level of representation there is a sparser form of dependencies structure. Meaning that there are these dependencies which you can think of a sentence like: “*if I drop the ball, it will fall on the ground*”, which relates only a few variables together. Now, of course, each concept like ball can be involved in many such sentences. And so, there are many dependencies that can be attached to a particular concept. But each of these dependencies is itself sort of sparse: it involve few variables.

We can just represent that in machine learning as a sparse graphical model, a S*parse Factor Graph*. That is one of the prior and the reason why such a prior is interesting is that it is something we desire for the kind of high level variable factors that we communicate with language.

There is a strong connection between these notions and language. The reason being that the things we do consciously, we are able to report through language. The things we don’t do consciously, that are going below the level of consciousness, we can’t report. Presumably, there is a good reason for this it is just too complex to put in a few simple words. But, what is interesting is that if we can put these kind of priors on top of the highest level of representations of our neural nets, it will increase the chances of finding the same source of representation that people use in language. I call them semantic factors.

What causes changes in distribution?

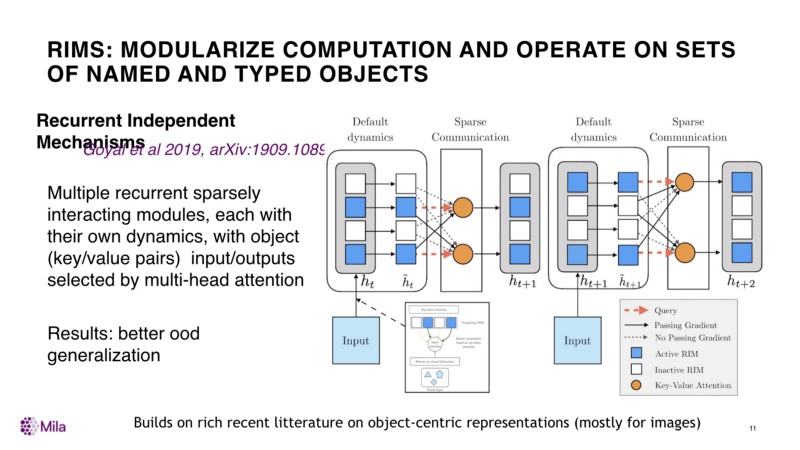
Another prior that I’ve been talking about has to do with *causality* and *changes in distribution*. Remember, I started this discussion by: how do we change our ways and improve our deep nets, such that they can be more robust to changes in distribution.

There is a fundamental problem with changes in distribution. Which is that, if we let go of the iid hypothesis (that the test data has the same distribution as the training data), then we have to add something else. This is something fundamentally important in order to cope with changes in distribution. Otherwise, the new distribution could be anything. We have to make some sort of assumptions, and I presume that evolution has put those kind of assumptions in human’s brains (and probably animal’s brains as well) to make us better equipped to deal with those changes in distribution.

What I am proposing as a prior here, and really inspired a lot by the work of people like of Scholkopf and Peters and others in causality is that those changes are the result of an intervention on one or a few high level variables, which we can call causes.

There is this prior that many of the high level variables that I am talking about are causal variables (there can be causes or there can be effects of something, or they are related to how a cause causes an effect). The assumption here is that change is localized. It is not that everything changes when the distribution changes. If I close my eyes like here or if I put some dark glasses, there is only one bit that changes, just one variable changes its value.

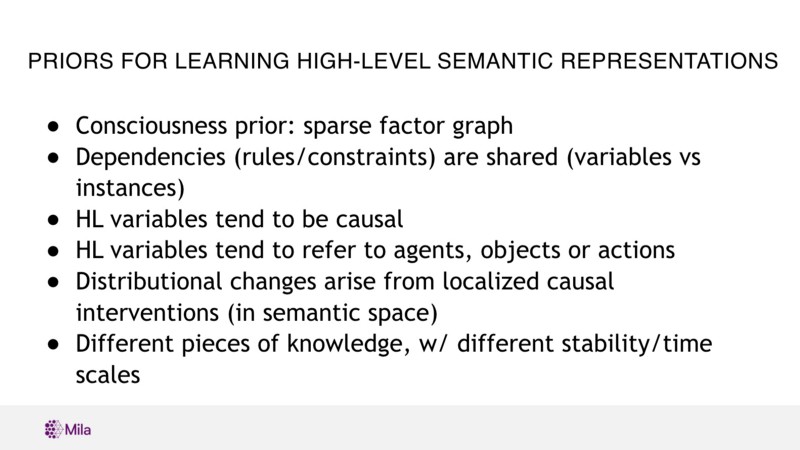
We can exploit this assumption in order to learn representations that are more robust to changes in distribution. This is what I talked about in my NeurIPS presentation. We can exploit that by introducing a meta-learning objective that says: better representations of knowledge have this property that when the distribution changes, very few of the parts of the model needs to change in order to account for that change. And so, they can adapts faster, they can have what is called a smaller sample complexity, that need less data in order to adapt to the change.

RIMs: modularize computation and operate on sets of named and typed objects

Another thing that we have explored is related to modularization and systematic generalization, as the idea that we’re going to dynamically recombine different pieces of knowledge together in order to address a particular current input.

We have a recent paper called “*Recurrent Independent Mechanisms*”(Goyal et al., 2019, [arXiv:1909.10893](https://arxiv.org/abs/1909.10893)) which is one first step at that, and I’m not going to go through the whole thing, but some of the main ideas is that we have a recurrent net. It is broken down into smaller recurrent nets, which you can think of different modules, which we call *independent mechanisms*. They have separate parameters and they are not fully connected to each other, so the number of free parameters is much less than the regular big recurrent nets. Instead, they communicate through a channel that uses attention mechanisms such that they can basically only sent these names vectors, these key/value pairs, in a way that makes it more plug and play. The same module can take as input the output coming from any module so long as they speak the same language, that they fill the right slots if you want to think in a symbolic sense. But, it is all vectors and it is all trainable by backprop.

There is also a notion of sparsity of which module gets selected in the spirit of the [global workspace theory](https://en.wikipedia.org/wiki/Global_workspace_theory) which comes from cognitive neuroscience.

PRIORS for learning high-level semantic representations

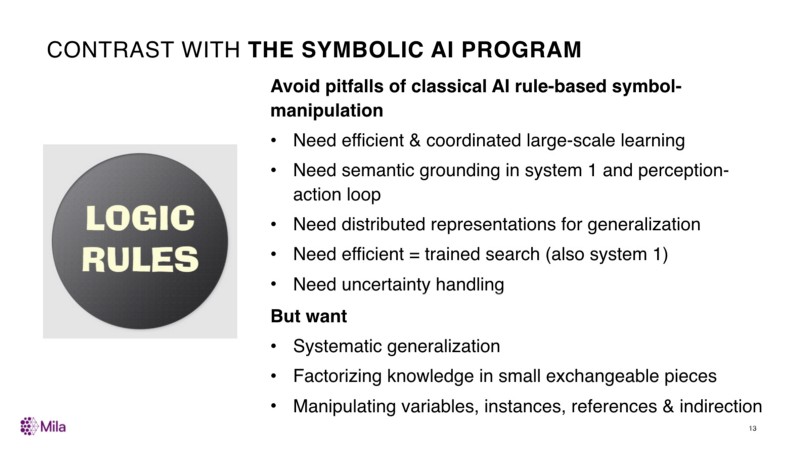
Let me list a few of these priors.

I have already mentioned a couple, and others I didn't have time to mention.

* The ***consciousness prior***, the idea that the joint distribution of the high level factors is a sparse factor graph.
* Another one I didn't talk about, but of course has nice analogs in classical GOFAI and rules is that the ***dependencies*** that I have been talking about are not dependencies defined on instances. It is not like there is a rule for *my cat* and *my cat’s food*. There are general rules that applies to *cat* and *cat food* in general. We do these kind of things a lot in machine learning, and in graphical models these date back to even convolutional nets and dynamic bayes nets which share parameters. So something like this needs to be there as well at the representations of the dependencies between the high level factors.
* I mentioned the prior that many of the factors at the high level needs to be associated with ***causal*** variables, or how causal variables interact with other causal variables.
* In the same spirit, and I didn't have time to talk about it, because it is really a whole other topic very closely related to this subject: ***agency***. We are agents, we intervene in our environment. This is closely connected to the causality aspect and the high level variables if you look at the ones we manipulate with language often have to do with *agents*, *objects* or *actions* (which mediates the relation between agents and objects). There are a few papers already in the deep learning literature trying to use these priors to encourage the high level representation to have the sorte of properties. And, of course, when you start doing thing like reinforcement learning and especially look at intrinsic rewards in reinforcement learning these are concepts that comes very handy.

Vincent Boucher: Professor Bengio, you have 5 more minutes.

* Then, there is this other prior I already mentioned: the idea that the changes in distribution arise from ***localized causal interventions***.
* Finally, one that is connected to this one, but it is different and is being explored by my colleagues [Léon Bottou](https://leon.bottou.org/), [Martin Arjovsky](https://github.com/martinarjovsky) and others before them, is the idea that some of the ***pieces of knowledge*** at the high level, or event at the low level, corresponds to *different timescale*: there are things about the world that change quickly and there are things that are very stable. There is general knowledge that we’re going to keep for the rest of our life and there are aspects of the world that can change: we learn new faces, learn new tricks. This is something that fits well with the meta-learning framework where you have fast learning inside slow learning. I think that this is another important piece of the puzzle.

Contrast with the symbolic AI program

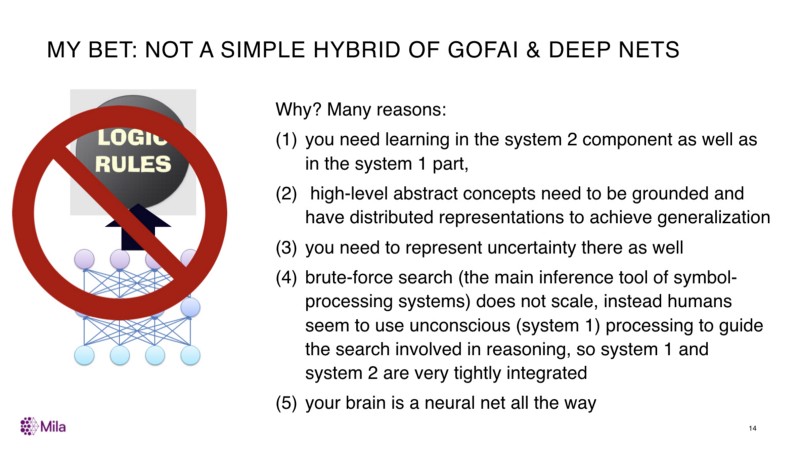
How is that related and potentially different from the *Symbolic AI Program*.

We would like to build-in some of the functional advantages of classical AI rules-based symbol-manipulation in neural nets, but in an implicit way.

* Need efficient & coordinated large-scale learning;
* Need semantic grounding in system 1 and perception-action loop;
* Need distributed representations for generalization;
* Need efficient = trained search (also system 1); and
* Need uncertainty handling.

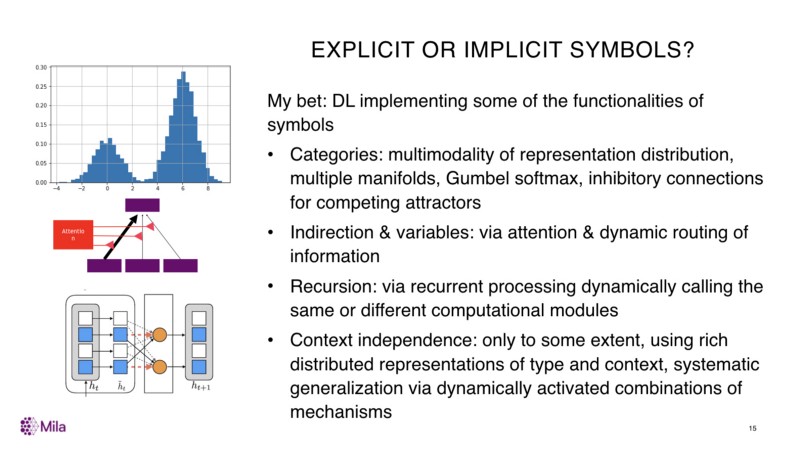
But we want to incorporate these other things (that really have been explored first by the people in classical AI), like

* Systematic generalization;
* Factorizing knowledge in small exchangeable pieces; and
* Manipulating variables, instances, references and indirection.

MY BET: Not a simple hybrid of GOFAI & Deep Nets

This is connected to why I think just taking the mechanisms we know for GOFAI and applying them on top layers of neural nets is not sufficient.

1. We need deep learning in the system 2 component as well as in the system 1 part;
2. We need those higher-level concepts to be grounded and have a distributed representation to achieve generalization;
3. We can’t do brute-force to search in the space of reasoning.

EXPLICIT or IMPLICIT SYMBOLS?

How symbols should be represented.

My bet is that we can get many of the attributes of symbols without the kind of explicit representations of them which has been the hallmark of classical AI.

We can get *categories* for example by having multimodal representations of distributions. We can use things like Gumbel softmax which encourages separation into different modes. We can get *indirection and variables*. We can get *recursion* by recurrent processing and we can get a form of *context independence* which is allowing to dynamically activate combinations of mechanisms in a context independent way.

**Let’s Debate!**

I’m done.

Thanks!

Vincent Boucher: Thank you, professor Bengio. Professor Marcus and Professor Bengio, you have 15 minutes to answer / debate.

**Response | A Dialog Between Yoshua Bengio & Gary Marcus**

**GM** I don’t think we disagree on all that much, except for your last set of slides. (…) I didn’t quite understand your response to Google search in a way. I tried out Google search as an example of a hybrid system that works in the real world at a massive scale.

**YB** I didn’t talk about Google search.

**GM** Yes, exactly. Your critique of the GOFAI hybrid systems. Let me just say: GOFAI is symbols all the way. I’m not endorsing that. I’m arguing for symbols plus deep learning. I take Google search to be an existence proof for things that you’ve just said couldn't exists. You said GOFAI is not going to be able to represent probabilities well, it’s not going to scale well.

**YB** I don’t understand why you talk about Google search. I’m not trying to emulate Google search. I’m trying to get intelligence.

**GM** You’re not trying to build Google search. You’re trying to build an intelligent system.

**YB** Yes.

**GM** Google search is in some ways an intelligent system and some not. But, I think, you have two avenues here. You can either say: *it is so different from an intelligent system that it is not interesting*, or you can say: *it is interesting and it does show the proof of concept to build a hybrid*.

**YB** No, look, I completely agree that lots of current systems which use machine learning also use a bunch of handcrafted rules and code that was designed by people’s understanding of the problem. This is how state-of-the-art systems, in particular dialogues systems I think is even a more obvious example where current state-of-the-art systems combine machine learning with a lot of handcrafting. It is also true of autonomous vehicles these days. I mean, there is a lot of engineering on top of the whole computer vision. I don’t think we disagree on this. The question is where do we go next in order to build something that is closer to human intelligence.

**GM** I may have misunderstood your argument. You are not saying (I’m going to recap to make sure that I understand) that one couldn't build hybrid systems. You’re saying…

**YB** They are already built.

**GM** That what I was saying, but, O.K. Then, I misunderstood your argument.

**YB** I’m talking about how the brain works and how I would like to build AI in the future.

**GM** Let’s come back to the brain part. Why are you not satisfied that hybrids are a part of the answer. If I read you correctly…

**YB** It depends of what you mean by the word hybrid.

**GM** At what point do you get off the hybrid train.

**YB** I get off the hybrid train when it is about taking the good old algorithms (like in production systems, ontology and rules and logic), which have a lot of value and can serve as inspiration, and trying to take them and basically glue them to neural nets. People have been trying to do these kind of things for a long time. In the 90s, there was a lot of neuro-symbolic work and so on. I have tried to outline in my last couple of slides (I guess I misunderstood that I had two more minutes left), I have tried to outline the reasons why it couldn't work. It is not just about how the brain works, but for machine learning reasons, for practical computational reasons. One of them is “*search*”. What I mean by *search* is: what we do when we have the knowledge in say, things like rules of pieces of neural nets, and we can dynamically choose which parts goes with which parts in order to come up with a new conclusion. This is what reasoning and planning are essentially about. If you introspect a little bit about how humans plan, on how humans reason, we don’t explore a zillion of different trajectories of possible ways of combining thing and pick the one that works best according to some criterion. We essentially go and try one thing, and sometimes two, and if really this doesn't work we try three of four. Go masters go up to 50 (their brains is weird) because they have been trained. Or people are very good at arithmetics. But, normal behaviour involves this very intuitive sort of, like we know where to search. That is based on system 1, that is based on something that we don’t have conscious access to and that knows where to search. That is one reason why we can’t use the old algorithms. The other reason is that the symbols themselves… we need to represent information in a richer way. The reason why connectionists really wanted to depart form symbolic processing is because they thought that is wasn't a sufficiently rich kind of representation. In order to get good generalization, you want to represent every concept, like words in natural language, by this sort of sub symbolic representation that involves many attributes. This allows to generalize across similar things. I have read some of the things you wrote and you could say: these attributes are like symbols themselves. Sure, you could do that. But, the important point, is now you have to manipulate these rich representations, which could actually be fairly high dimensional. We need to keep that from the neural net world. We need to keep the things that have worked well in machine learning, which include representing uncertainty which some people are doing like [Josh Tenenbaum](http://web.mit.edu/cocosci/josh.html) with probabilistic programming and so on. I think there are some efforts going in those directions, but we need to keep these ingredients together.

**GM** I am going to mostly emphasize our agreements here. I agree first of all that classical AI systems have search issues. I think that to the extent that ones wants to preserve them, ones want to solve these problems. There are ways that people have thought about it. For example, *Insight* which is the kind of classic, most huge symbolic effort. In *Insight*, there are macro theories to target reasoning in particular domains. I think this is an idea worth exploring. I absolutely agree that if you have unbounded inference, you are in trouble. I think that Alpha Go is a example where you bound search partly through a non symbolic system and then you use a symbolic system there as well, so it is kind of a hybrid.

**YB** In what way is it a symbolic system?

**GM** The Monte-Carlo tree search is just traversing trees.

**YB** It is a search, but there is like no symbols.

**GM** You have to keep track of the trees and trees are symbols.

**YB** Ahhh

**GM** That actually brings me to a separate line of discussion I’d like to have.

**YB** I think it is just a matter of words. We need search. Obviously, we need some kind of search. If you want to call this symbols… I think symbols to me are of a different nature: symbols have to do with the discreteness of concepts. This is also something that is important. As I mentioned quickly at the end of my presentation, we can get discreteness not necessarily in his hardest form, in its purest form as you have with symbols. You can get discreteness by having a neural nets lateral inhibition that creates a competition such that the dynamics converges to one mode or another mode. This is what you observes in the brain, by the way, when you take a decision there is a sort of competition between different potential outcomes and so the dynamic chooses one of the discrete choice over another. But, it does it in a soft way and the brain have access to all of this soft information.

**GB** I am going to lay something else on here. I think that we both think the other side is strawmanning our baby. I think that you are strawmanning symbols because lots of people have put probabilities and uncertainties into symbols and you think, and I think that this is an interesting discussion point, that I am strawmanning deep learning. You said that I am attacking the models of the 1980s, and there is some truth in it, and then there is the question of what the scope should be. I think both for symbols and for neural networks there is a kind of question of what is the proper scope of them. We are actually pushing to the same place from opposite sides. I would argue that the kind of deep learning stuff that was straight out of the 80s, which was still continued until 2016 in my view, we can argue about that. Let have a big multilayers perceptron, let’s pile a lot of data in and hope for the best. Which I don't think you believe anymore. Maybe you did at some point. That is one kind of deep learning. That is the kind of prototype or canonical version of deep learning. You want to open deep learning to a whole lot of other things. I think at some level that is fine and that at some levels this is changing the game. You might, I’ll elaborate them in a second, I think that with respect to symbols you might feel that I am doing the same. So I want to say: sure symbols (I want the discreteness of symbols), but I am very happy to add in probabilities like in probabilistic stochastic grammar, or something like that, I have no problem with that. I love a lot of [Josh Tenenbaum](http://web.mit.edu/cocosci/josh.html)’s work, which is really like symbolic programs plus uncertainty. I want to expand the umbrella of symbols and you want to expand the umbrella of deep learning. Why don’t we say let’s build deep learning symbolic systems that expand the scope of deep learning and expand the scope of symbolic systems?

**YB** Look, I don’t care about the words you want to use. I am just trying to build something that works and that is going to require a few simple principles to be understood. I do agree that there is lot of interesting inspiration that we can get today in the work that have been done in cognitive science and in symbolic AI. But, I think that some of that need to be reinvented. And, by the way, we started doing things like attention mechanisms and people were doing reinforcement learning already at the beginning of this decade. Attention mechanism date event from much earlier than that. So, it has been around. Another thing that you have to keep in mind: I have been working on recurrent nets since the 80s and in a way, the various forms of recurrent nets, including the gated ones, use very similar principles. Again, I have been around since the 90s, so it is not completely new things. There is an evolution, of course, we are doing research. It is not like we have one algorithm and we are stuck with it. We are building and constantly trying to expand on the set of principles that we have found to work. There is nothing wrong with that.

**GM** There is nothing wrong with that at all. I think we should actually yield to the questions from Vincent and the public.

**YB** Sure.

**Interview | Vincent Boucher : Yoshua Bengio & Gary Marcus (dialog)**

**VB** The first question is for Gary Marcus

Nature & Nurture

“Deep learning” neural nets are in fact shallow, soaking up patterns but lacking explanation, causality, rule-based reasoning for novel & unique situations” — Steven Pinker <https://twitter.com/sapinker/status/1201238394317070336>

What is the innate knowledge for deep understanding and what needs to be learned along the way?

**GM** I made a slide that I did not have time to show which has a picture of a great new paper by Yoshua that we had on the reading list, which is well worth reading, about causality. It is a very mathematical paper and I took what I think is some of the core math from it. I will admit that I didn’t read the paper as carefully as I wish that I had. Yoshua is going after causality by trying to make some clever observations about how distributions change over time relative to interventions that are made. Which is of course the classic thing that we try to do when we run experiments and he has got some very clever ways of going after that within neural networks and I think that this is great work. It is not the work that I would do but I think it is terrific. On the right, I have something from a paper that Ernest Davis and I wrote (Ernest did almost all the hard work, but I helped a little bit) that most people in the field right now would find to be repulsive but that I think we need to think about very carefully. Ernest created a logical formalism for understanding something very simple, which is “containers”. I have water in this (glass). If I tilt it, the water fall out. What happens if I drop the microphone in it (maybe not the electrical part of that, but just the physical reasoning about it). The formalism that Ernest came up with that I think is a responsive to your question, is something that broke things in time, space, manipulations, things about rigid objects and the history of objects. He did a very careful analysis of the knowledge that one need in order to do this basic thing. It is not trivial things, because we use container metaphors for a large fraction of the things that we talk about. I don’t want to say that it is fifty percent, but it is significant. For example, we can think of a container as a lake, we can think of a cup as a container, we can think of the body as a container and so forth. The argument of this paper was that in order to be able to make inference about these things, we need prior knowledge. There is a the question of whether that knowledge is innate or it is acquired experientially, but the argument is that you won’t be able to make these inferences unless you have this knowledge about sets and objects containing regions and have these kind of axioms and those kind of axioms about what rigid objects can do. One possibility here is we need the formalism on the left in order to acquire the knowledge on the right. Another possibility is that we never need the kind of knowledge on the right, it never need to be reified in the way that Ernest Davis proposes. My view is that we should have people working on both sides of the spectrum. People often think they are in a minority. I feel like I am in a minority, but we can do the sociology later. I like to see more people working on stuff like this to build some broad framework for space, time, causality and so forth, but I totally welcome the kind of stuff that Yoshua is doing event if I don’t personally have the skills to do it. I think that the empirical question is: could you from the bottom-up derive all this. Although, I feel like maybe I strawmanned Yoshua. I thought that he was more anti-nativist than he really is because he acknowledges evolution.

**YB** Of course I acknowledge evolution!

**GM** I’ll say one more sentence and then you can take it away. So in my view, what part of the field should be doing is saying: do we have innate priors around things like this? This is the kind of work [Elizabeth Spelke](https://psychology.fas.harvard.edu/people/elizabeth-s-spelke) does in cognitive development, [Renée Baillargeon](https://en.wikipedia.org/wiki/Ren%C3%A9e_Baillargeon) and so forth. Part of the field should be trying to reify that knowledge and part of the field should be like: if we had that knowledge and we know something about causality, how can we learn from that.

**YB** Let me impersonate Yann LeCun for a minute. It is not that Yann and I and others with similar thinking think that learning has to be from a blank slate. In fact, we have theorems form the 90s, the no free lunch theorem, that clearly says: you can’t have learning if you don’t have some priors. What we are saying, it is more subtle than that, we’d like to be able to get away with as little priors as possible. How is “little” measured. You can think of measuring it in bits. If you think about of how big is a program that would encode those priors and you would zip that program, that would be how big the prior is. The kind of priors that I have been talking about in my presentation, I was talking about priors, but these are priors that in a way are not going to require many bits. It is going to be easier for evolution to discover those priors. I also know full well that evolution has discovered very specific strong priors. In fact, if you look at evolution, most of it is about completely hard coded behaviours. But, these are not the behaviours that are the most adaptive. These are not the behaviour that allow a species to adapt as well as humans have been able to do. So, it is more interesting for me to think about the part of what evolution has discovered that is more general. These are the most generic priors. Of course we have priors that are very very specific. We kind of know how to see and to walk to some extent when we are born. Many animals have a lot more when they are born. It is just a matter of what we care about here is trying to squeeze the prior knowledge into as few simple general principles as possible. We don’t know where is the right line, of course.

**GM** To use you language, you have a meta prior where you want as little innate stuff as possible. This is the place where we disagree in taste. I don’t want a huge amount, but I think I want more than you. Of course, we don’t actually have a number, but let me give you my intuition.

**YB** I would not want to have to design the semantics of each of the boxes in an AI system like this.

**GM** Why not?

**YB** I didn’t say why I would like to have as little priors as possible. It is because these lead to more general-purpose machinery that can be applied to a wider spectrum of behaviours, environments, problems and so on. It is as simple as that.

**GM** I’ve got two things to say. One is actually from Yann’s work, since you mention him. We actually argued about this very thing the other day (we were at NeurIPS last week on a panel). In this particular empirical case, having more of a prior was actually better. In this particular case, having a convolutional prior was actually better.

**YB** Of course. It is a very small prior. It is like three lines of code of difference. It is not a big change in the amount of information, compared to the classical computer vision that was done before convnets, where you had to design the functions by hand completely.

**GM** Yann share the Turing award with you for those three brilliant lines in a sense, right? They are very clever and they have been very valuable to the world. Maybe, I’ve got 24 boxes up there, they are three lines each and we just need 24 more discoveries of that magnitude. is the genome big enough to encode all those? 95% of our genes are involved in brain development. I think that there is room in there to encode that many, and maybe 10 times more.

**YB** There is a lot of room in the genome, but clearly not enough to encode the details of what your brain is doing. So, it has to be that learning is explaining the vast majority of the actual computation done in the brain just by counting argument. 20,000 genes with 100 billion neurons and a thousand times more connections.

**GM** That was what this book [[The Birth of the Mind](http://garymarcus.com/books/birth.html)] was about. The genome shortage argument. The idea was that we only have so many genes, let say 20,000 genes (we thought it was 30,000 when the book was written), we have 86 billion neurons. What are the implications of that. The genome shortage argument was: we have to learn it all.

**YB** Nobody in their right mind says that.

**GM** It is partly a question about what our bid is you know. I want to have twenty things in this debate that Yann and I. I put ten on the board, and he said none.

**YB** By the way, the things you put on the board, I agree with most of them. In fact, most of them are small priors. They don’t require a lot of bits to be specified.

**GM** Those were things like spatial-temporal continuity.

**YB** You have them in convnets.

**GM** Well, not the part that is tracking objects over time and know that it still exists.

**YB** It uses spatial continuity.

**GM** It is really translational invariance.

**YB** It is actually more than that because of the pooling, but yes.

**GM** The others things that I had on the list were things like symbolic priors. So, operations over variables. Those I think you would be less comfortable with.

**YB** I am totally comfortable with operations over variables. It is just that the meaning of *operation* and *variable* is different for me and for you. I am thinking of operations as a little neural nets that do things that are not just discrete that manipulate rich representations. And, I think of variables as indirection, passing informations about references, about keys, about the nature and types that can be rich rather than symbolic. But, besides, I agree with the need for references, for example.

**GM** In a way, that is all I really want. You’ve made me a happy man. I’ll tell you another time that you made me really happy, earlier tonight. You talked about having a reference without knowing where it comes from.

**YB** No. I am saying: the reason why you need reference is to be able to know where the value your are getting comes from. That is the reason these neural nets need to propagate not just value, but also “*names*”, except those names are going to be vectors as well.

**GM** You are more symbolic than I am. I am finding it harder and harder to disagree with you. We will take another question.

**VB** Question for both Gary & Yoshua

Montreal.AI reached out to Jeff Clune.

99% of the machine learning community is focused on what Jeff Clune has called the manual path to AI, in which we manually identify building blocks of AI with the assumption that one day we will somehow put them all together.

Do you think a higher fraction of our collective effort should be reallocated into the alternate path of AI-generating algorithms that Jeff proposes, wherein we try to simultaneously (1) meta-learn the architectures, (2) meta-learn the learning algorithms themselves, and (3) automatically generate the training environments?

That is in line with Gary’s observation that we need strong meta-learned priors like those evolution provided us.

That is also in line with Yoshua’s view that we should learn things end-to-end (e.g. deep learning over HOG/SIFT)

In practice it is a very different research agenda from either of you, wherein people stop trying to hand-design powerful ML systems and instead meta-learn them.

**YB** I like this question very much. Because, I worked on this question in the early 90s with my brother Samy and this was essentially the subject of his thesis proposal. This was one of the first paper on meta-learning. We were trying to meta-learn a learning synaptic learning rule. We didn’t have enough computational power to do this. Even now, I think, in order to realize the kind of ambitious research program that Jeff is talking about, we would need a lot more computational power. That being said, I thing that this is a very interesting and important investigation. I was really amazed at the presentation that [Blaise Aguera](https://en.wikipedia.org/wiki/Blaise_Ag%C3%BCera_y_Arcas) gave at NeurIPS on this subject. I think this is very exciting. Personally, I am also tempted by the desire to understand the principles that would be discovered. When I tried to do the meta-learning of learning rules, what I quickly realized is that you can’t learn something that is completely in the abstract. It really helps a lot if you put in a bit of the right structure. In order to do that, you need to do experimentation of the kind we do normally in machine learning where you design the learning algorithm completely. And, that help to figure out what would be the right building blocks and the right inputs and outputs that are needed for learning a learning rule or learning a system like this. Science is an exploration. We do not know what is going to work. These are two different directions and they can coexists in a harmonious way.

**GM** I pretty much agree with Yoshua’s answer. I’ll answer in a slightly different way. In principle, we know that evolution is a mechanism that is powerful enough to evolve minds, because it evolved our mind. Having the machine do the work that sort of stand in for evolution would be great. In practical matters, it does matter what you are trying to evolve. I think what has happened empirically in the evolution of neural networks literature is that people starts with too little in the way of priors. So, they end up recapitulating some of our journey into bacteria, but not so much of our journey form chimpanzees to human beings. In principle, we know it can work. In reality, having a tightly constrained problem and probably a bit of priors to help us there might help it work event better than it is and i think it is totally worth exploring.

**VB** The third question is for Yoshua Bengio

Ethical Conscious and Reasoning Systems

There will be ethical implications for conscious and reasoning systems? How do you approach that?

**YB** I think it is important in general to ask the question of how our work as researcher will be used or could be used. We don’t need to go very far in the future. Today, we already see misuse of AI in many ways. I’m very concerned about how we are creating tools that can be destructive and endanger democracy and endanger human rights. The specific question of consciousness, I think, deserves a bit more time than this debate allows… Personally, I think that the kind of conscious processing that Gary and I are talking about are adding more computational power and intelligence to the systems that we can built. But, I don’t think it changes fundamentally the fact that we are building gradually more and more powerful systems. There is the question that some philosophers are asking about wether we should eventually give personhood to intelligent conscious machines. I don’t think that we are anywhere close to understanding these questions enough to be able to answer these sort of questions.

**Public questions from the audience at Mila | Yoshua Bengio & Gary Marcus**

**Audience question #1** Thank you very much for this interesting debate because artificial intelligence is going to solve a lot of problems that matter very widely to many persons. I am not a computer programmer. My questions, I have several questions but I’ll limit to two. Professor Marcus said that Professor Bengio’s approach relied too heavily on larger data sets to yield answer. Why is that necessarily bad?

**GM** First of all, I said that that was my impression of Yoshua several years ago. It is not my impression of Yoshua now. I think that he is doing a lot of exciting work and he is right that some of it started a while ago. My impression when I first talked to him, and I had friends at linguistic conferences where, when we would come to him say: “*yeah, but the kind of systems that we have right now, they can’t solve this*”. I felt that his answer was often: when we get a big enough data set, we will be able to cover that. I have some quotes from the slide showing it. I think there are many people that more extreme about it than Yoshua ever was. There is a branch in machine learning where people think that the answer to a particular problem is really about getting the right data set. I think Tesla’s approach to driverless cars is more or less like this. They say: we’ve got the most data, we have very cool ways of trying, for example, to gather data about a particular kind of accident when it happens and so forth. This is very focussed on the data and not so much focused on certain kind of innovations and algorithms spaces that I would like to see. I have no objection to gathering more and more data. I think that getting clean data in really important. People often underestimate the value of having good and clean databases. I thing the field was driven forward by having bigger databases. No problem with any of that. But, the answers aren’t just there. In Yoshua’s terms, you know we need system 1 and system 2 and I would like to have more people working on system 2. Maybe we disagree a little bit on the execution of that. But I think we agree that we need some of that and not just the system 1 stuff plus bigger databases.

**YB** I want to say a few things about data. Because, I didn’t answer about this quote that you attributed to me. I think that I am interested in the small data regime to the extent that we also have a lot of data before we get to that point. Humans learn a new task after they have seen a lot about the world. There is no chance that you will be able to learn in a meaningful way without a lot of knowledge about the world that has been acquired previously. We need both large data, in some sense, we need a lot of example in order for the “BabyAI” to mature and then it can faces a new task very quickly. That is one thing. On the industrial side, if today I lead a company of a project, I am going to use as much data as I can because this is the thing that works well. But, at the same time, if you are looking further down the road, in a few years, and you are asking yourself: what kind of improvement to our current algorithms would be most interesting for industries of for any kind of applications, then looking at those transfer learning problems where you are looking at new tasks where you have little data, but you also have pre-trained on many other things, that’s more right now in the research. The two things are not incompatible. It depends on whether you are doing something in the short term or the long term.

**GM** I found just a tiny bit of something to disagree with Yoshua. But, I actually mostly agree with what he said. The one place where I disagree a little bit… First, let me explain what a small data regime is, because not everybody will know. There are problems where people learn thins with a small amount of data. Yoshua would say that it is because they have a lot of experience elsewhere, and that is often the case. In the small data regime, how do you learn something if you don’t have 10 million data points? If you are my kids and you learn a new game in five trials, how do you do that? Clearly some of it is that you leverage prior experience. The only thing that I am going to add there is the reason I did that baby experiment back in [1999](https://www.ncbi.nlm.nih.gov/pubmed/9872745) was to show that there were something that little kids could learn without much direct experience. I made up the language, so they had no prior experience with the language. The habituation, a period where they learned the made-up language, was only two minutes. They only got something like 45 examples of the made-up language. Sentences like “*la te te*” and so forth for two minutes. Yet, they were able to do this. Then, somebody else, this is what happens in developmental psychology, if you show that kids of a certain age can do X, somebody else says yeah now I’ve got even younger to do X. So, somebody later showed that even newborns could do what I had showed in the 1999 Science paper of kids extracting rules. There is pretty good evidence that even newborn could do this. In this particular case, I think that what you have to draw on is not experience outside the womb, but the experience that we get indirectly from evolution. Some of the problems that we solve in a small data regime come because we have priors for variable and things like that.

Next question now.

**Audience question #2** Thank you for this presentation. Dr Marcus, you talked about the compositionality and the need to take into account compositionality. From a linguistic point of view, we have debates and arguments on compositionality, but to make a simpler system we accept compositionality. We had some progresses with the literature of neural nets, the recursive neural nets for compositionality. However, those efforts has been abandoned — we don’t do research anymore on the recursive nets. I think that the argument is that we need the parse tree, we need the knowledge, to feed in the recursive network to design the architecture and to form the network. I think that there is a resistance here. That the deep learning community, they are not willing to take any external knowledge in the form of the linguistic structure or the parse trees. Dr. Bengio, would you please elaborate on that?

**YB** I don’t think it is a resistance as much as an obsession to beating the benchmarks. Which could be good or bad. It is because these very large fairly simple architectures have been working so well. A good example now is the successes of transformers. Transformers are working incredibly well. They are using, actually, these key-value pairs I have been talking about. They are operating on sets. The recursive nets was one attempt, but there have been others that have been more successful. Maybe recursive nets will come back, we don’t know. The history of science is very complicated, as we have seen with deep learning. I don’t read the sociology of the current deep learning field like you are. In fact, there is a lot of interest in exploring how we can put some architectural structures in neural nets that facilitate the manipulation of language and reasoning. I am much more optimistic than you seem to be.

**GM** I would say that historically there has been resistance. I think that this is changing somehow. I think it is partly a function of people having tools that are good at particular things. We don’t really have “deep learning tools” maybe in the extended sense for really dealing well with recursion and compositionality in the sense that I am describing here. I think that there is much more hunger in that field in the last two or three years to do it. In terms of the transformers, I just gave a talk on a new benchmark called “Dynamic Understanding” at NeurIPS and you can probably Google for it online. The basic point I made about transformers like GPT-2 is that they make very fluent text, but they don’t understand what is going on. I just have an example here from a slide. They are often plausible in the first few sentences of surrealist fiction basically. I feed this into a system across the street from NeurIPS: two unicorns walk into a bar... And then the system continues that passage with: at least that’s what my picture shows. I’ve never seen such a multicoloured, beautiful forest of sapphire eyes on the same corner of the street in a bar before. It is like fabulous that it is creating this surrealist prose. On the other hand, when I forced it into a nonfiction genre it seems a bit ridiculous.

In the example I had on the right: two lemurs walk on a road and another joins in. The total number of lemurs on the road is… And, you are supposed to add up two and one and come up with three. If you are human you probably do that. But, if you are a deep learning system you might come up with something like “not 100 as claimed, but about 80 or so.” The system doesn't converts the prediction and statistics that it is making about plausible classes of words into a direct representation of the individual entities that are involved. If you watch my benchmark talk, it is full of examples like that. I give things about conventional knowledge definition, transformations, atypical consequences and so forth in the models on the right. And they are typically doing like 30% or 10% or something like that. There are sharp limits. I think those limits comes because we don’t have kind of a parse tree on the output yet and we need to do that.

**Audience question #3** Since Turing 75 years ago and its virtual machine and all we could do with the binary mathematics we’ve achieved great things. Today, we are with quantum computing and quantum computers, which are closer from the way that the human brain thinks: something can be right, or wrong, or can be both almost at the same time. Could quantum computing or quantum computer then represent a breakthrough that we are waiting for to achieve the artificial intelligence the best way?

**YB** Maybe so. But I am a big fan of Occam’s razor. If we can build intelligent machines that explain how the brain works without having to go quantum, I think it is very satisfying to go for a simpler solution. I think in terms of neuroscience, most of the community thinks the brain can do these computation without requiring (of course there is quantum computing in the sense that molecules are operating in a quantum way). But if we abstract one level up, it is all computation that is not quantum by nature. At this point, the majority of the community, both in neuroscience and computer science are betting on traditional computing in a sense that this is not quantum. Right now, there are not many algorithms that can be efficiently paralellized by quantum computing, and no serious machine learning algorithms like deep nets and so on. If they can find the right vertical breakthroughs that enable to implement things like deep nets, in a way that takes advantage of the quantum capabilities, then it would change the game. It hasn't happened yet, but this is something we can look for it.

**GM** I pretty much totally agree. My friend Sandy is going to ask a question.

**Audience question #4** I’m gonna start my question with a little anecdote. When a bunch of journalists interviewed the scientists who created the nuclear bombs one of the things that they profoundly stated was: they were so involved in the science they didn’t even think of the ramifications. So, I’m listening to you two geniuses here and I’m not even gonna pretend that like three-quarters of this isn’t SpaceX going right over me. But, one thing that disturbs me is: I don’t hear a single word about checks and balances and ethics that are going into your creating the algorithms that are going into all of this you know AI. As somebody who’s not an AI, who is a human in this world, I find this incredibly disturbing. I’m sure Gary has heard me say stuff like this before. I’m bringing it out again because I would love to hear you guys address this!

**GM** The first thing I’ll say is: of all the people in the field you could have leveled that accusation against, I think Yoshua is the least appropriate. I think Yoshua thinks pretty deeply about this and I’ll let him speak about his version. My own version, in the book review [Rebooting AI](http://rebooting.ai/), was to argue that common sense could be a way of building a framework such that machines could represent values. So, you can think about Asimov’s laws you know. You want robots to not do *harm*to people. One of the things we talk about is how do you get a computer to even think about what a harm would be to a person. So, it’s one thing to get a computer to recognize a picture of an elephant after you’ve seen many other pictures of elephants. They can’t really do the same trick for harm. Harm takes many many different forms it’s not really about the way that the pixels fall on the screen or the page. A lot of the argument that Ernie Davis and I gave about this particular set of issues was: we need to rethink how we get knowledge into these systems and the nature of knowledge as a platform to then get to be able to program in the values that we want. That’s how we thought about it. I don’t think it’s a full answer. That it is how we thought about it. I will turn over to Yoshua.

**YB** Thanks you for raising that question. It’s very important! Gary and I have been talking about maybe something a little bit technical from your point of view, about where we think the our research should be going in terms of how we build smarter machines, but it’s at least as important that our society invests even more on the question of how are we going to deploy these things, what is the responsibility of everyone and a chain from the researcher to the engineer to the people doing auditing or to government’s drafting regulations to make sure that we steer our boat in a direction that’s best for humanity, that’s best for citizens. I’m very concerned that we’re building tools that are too powerful for our collective wisdom and I’m fine with like slowing down the deployment of AI. I think governments are not yet ready to do the proper regulation and we need to spend more time talking about things like: how AI can be abused to influence people, to control people, to kill people. These are all very serious issues: discrimination, killer drones, advertising, social media, deep fakes. Basically, right now is the Wild West and we need to quickly get our act together.

**GM** Maybe we will. I’ll just give one last example, for one…

**YB** I want to mention that here in Montreal, we we’ve been really working hard on this question and we came up last year, after two years of work involving not just scholars but also citizens, with a thing we call the [Montreal Declaration for the responsible development of AI](https://www.declarationmontreal-iaresponsable.com/). I invite you to check it out online and we’re pushing these ideas to the Canadian government. There’s been many frameworks that people have developed around the world  
to try to think about the sort of social norms that we need in this deployment of AI. Now, I think it’s a lot in the hands of governments and the agencies that are looking at specific sectors where where this technology is being deployed. It’s also in the hands of the UN, if it involves for example military deployment. And for that to work we need the media, we need people to voice their concerns.  
  
**GM** I’ll just add one thing, because I think we have to go to the online questions. I wanted a amplify the point about the Wild West. A good way to think about this is: right now a driverless car manufacturer can basically put anything on the road. We can sue them after the fact, if they cause great harm, but there’s no regulations essentially, about what you can do with a driverless car. If you compare that to how much trouble there is to perform a new medical test or build the new drug and how much regulations there, there’s an asymmetry that I don’t think makes a lot of sense. I’ll give a shout out to my friend [Missy Cummings](https://twitter.com/missy_cummings), who has a podcast, I think with [Azeem Azhar](https://www.linkedin.com/in/azhar/), [Exponential View](https://www.exponentialview.co/), talking about this issue : the asymmetry in regulation between what’s required for health and what’s required for a AI. I think Yoshua and I agree that there needs to be a lot more.

**International audience questions | Yoshua Bengio & Gary Marcus**

**VB** For the last segment, our participants will answer questions from the international audience.

**GM** I’ll do very quickly symbols while Yoshua picks the next one. What’s my definition of symbols? I don’t think we should waste time arguing about that. I think, from the perspective a symbol manipulation, the real question is: do we have operations over variables? You can define a symbol in such a way that it encompasses everything or nothing and I don’t think that’s where the debate should be.

**YB** So, there’s a question about: what is the chance of AI possessing self- consciousness? I think this is a very interesting question, but it’s also very loaded, because we all have our own ideas of what consciousness means. We think we have something special. What I can say is it’s something, fortunately, that scientists in neuroscience, kinetic science and machine learning are starting to think about and, hopefully, we can remove some of the mystery and magic from there, so we can be better equipped to answer these kinds of questions later.

**YB** What is the best way to reproduce the levels of conscious and unconscious thinking in AI? Well, that’s what you are arguing actually about. The answer is: we don’t know. That’s why we need many different researchers to explore different ways of doing this.

**GM** Gary Marcus thinks the deep learning and symbolic AI are compatible and can provide the best of those worlds. Is there any evidence? I think that the best evidence that we have for that is: we have some people building actual hybrid models in the real world that do useful things. None  
of them achieved human level intelligence. You know, no deep learning system does that, no symbolic system does that, no hybrid system does that, but systems like Google search do something that’s relatively intelligent and help us. They’re very much hybrid systems. Then you have results like the Josh Tenenbaum and so forth results that I showed briefly where, at least in  
a very controlled environment, a hybrid system can be a deep learning system or a symbolic system on its own. It’s still an open argument. I don’t think in the end that either Yoshua or I would say we have the answers here, right. We’re trying to lay out what we think the territory is that people need to explore. I think the biggest take-home message, as I said on the slide, is we  
actually agree a lot about what that geography is that needs to be explored. We have some differences about where to go in that exploration. Neither of us think that we’ve reached the destination by any means

**YB** I want to talk about the question about do you think that language understanding is a form of intelligence? We clearly need better language understanding for AI. There are really interesting connections between language, understanding and reasoning but they’re really different. I listened to a presentation at the last NeurIPS by [Ev Fedorenko](https://evlab.mit.edu/) and she’s a cognitive neuroscientist. What she found, with our colleagues, is that there is a language area in the brain and it does process everything that’s connected to language, but it doesn’t do the other things that one might think are related to language like reasoning. It’s other areas that are doing it. That’s also connected to the bigger picture of language that I’ve been talking about. Language has sort of syntactic aspects and structural aspects, but the semantics what language is referring to, is you know people call common sense and grounded language understanding, refers to general knowledge of how the world works. This is an area which is very active in machine learning. People, irrespective of whether they do language or not, are looking at how learning systems which interact with their environment can build better models of the world. If we don’t do that, we’ll never have good language understanding. This connects with some of the things that Gary talked about with limitations of transformers. I mean transformers work incredibly well. These are the best things we know right now in order to process language in quantitative benchmarks. But, as you said, they have, you know, they make what I call “stupid mistakes”. I think one of the missing ingredients is they don’t have a world model. They don’t, I mean they might build quite a bit actually a world model through reading text, but there’s a lot about the world which you can’t get, I think just from reading text. Maybe this is a place where Gary and I could disagree. I think that there’s a lot of knowledge about the world which is intuitive, for example intuitive physics, that is difficult to put in words. I mean, of course physicists do it, but babies don’t do it. They have an intuitive understanding of physics. In order to do good language understanding, I think that we need machines that understand how the world works in a physical sense, in an intuitive sense, and these two things need to be tied and that’s connected to the link between system 1 and system 2 that I was talking about.

**GM** I think we could probably talk for the rest of the time about this one question. There’s a lot of interesting things there. I think the first thing that I will agree with with Yoshua… Language and reasoning are clearly separate things but, they’re not fully separate. There’s wonderful work for example from Michael K Tanenhaus and John C. Trueswell showing, experimentally, the people reason about the world at the very moment where they’re processing it. If I give you an ambiguous sentence, you will look to what are the things out there in the world that can help me to disambiguate the sentencing. You will reason like: is there a cup on the table or a cup on the towel and I’m gonna put these all together in an understanding of a sentence. It’s hard to draw a sharp line, as you know, interesting work notwithstanding, there’s certainly overlap. On the other hand, you know very clear example of how important all the physical reasoning stuff is, would be any primate that’s not a human, right? Think about all the physical reasoning that a chimpanzee can do without any language at all, we could argue about the ape language studies, but I don’t think they’re very compelling. You have species that can, you know, navigate their way through trees and have social interactions of all you know very complicated social interactions, exchange, and all of these things, without any language. I think we would both be thrilled if, before we leave this mortal coil, we were able to build AI systems that could do what chimpanzee do. Now, I have a personal interest in language, having studied it for most of my career, and so I’d really like to see us get language right. I think having the world model is a prerequisite and it’s really hard.

**YB** Let me talk about this question about reasoning. How do you define reasoning and what do you think deep learning will or will not be able to solve it. Of course people have been tackling reasoning for a long time. and well before Neural Nets, you know, became hot and considered as “potential tools for reasoning”. The way I think about how deep learning can do reasoning is connected to what I mentioned as these dynamically recombinable pieces of knowledge that we can search through. We can search through which piece of knowledge can be combined with which piece of knowledge in order to find an answer to a question. Those searches are heavily guided by our intuition so we know where to search. Reasoning is about looking for coherent solutions to a problem, to a question. There is in all the way of thinking about reasoning which I find really appealing, which dates back to the early eighties neural network of Geoffrey Hinton with Boltzmann machines where, you can think of reasoning, and if you can find it again in modern graphical models, you can think of reasoning as finding a configuration of the random variables, the variables that maybe provide answers to your questions, that is most compatible with everything you already know right, which has the highest probability given all the facts you’re giving to the machine. With Boltzmann machines you’re trying to find that through a Markov chain, which searches, in the way that it does it is by looking for a low-energy configuration by gradually changing the configuration, until you find something that is good. I think something like this could make sense for the kind of unconscious reasoning that we do. We all have the experience of asking ourselves a question, not getting the answer back right away, moving on to something else and then maybe the day later, the morning after, or something, the answer comes to you. The thing that has happened during those hours is happening in the background and it’s something that we don’t you know, it’s harder to characterize, but it may plausibly be this kind of energy minimization. Now, the kind of reasoning that we do consciously is very different. We don’t consciously experience going through thousands and thousands of possible configurations. We immediately search through a few things that happen to be very relevant. I think we need the two kinds of reasoning.

**GM** I think, again, we agree on the two kinds of reasoning. You could think about what you called 80s deep learning networks and I call pure deep learning. I would say those can’t do certainly what you call system 2 reasoning. I would say that, right now, the best system for doing system 2 reasoning is the much-maligned CYC, and people might want to look at an article in Forbes where CYC is given Romeo and Juliet, and not straight text, but, put into computer interpretable form, and you could argue about whether that’s fair, but it’s given Romeo and Juliet and it can make some interesting inferences about characters motivations and so forth. That’s a symbolic system. I would say that the richness of inferences that can be made by symbolic systems is for now ahead of deep learning but I will also grant Yoshua…

**YB** But, it doesn’t work.

**GM** Well, I mean in a narrow domain, I mean actually in many narrow domains, it can work to some extent. I certainly don’t want to say that it’s the answer. But, it’s a proof of concept that you can do.

**YB** And it’s very unlike how the brain does it. Your brain doesn’t go through zillions of, you know, trajectories.

**GM** I will agree on that point. But, hold on. The contrast that I wanted to draw is… So, we have a system that doesn’t really, I think do reasoning at all, which is a pure deep net multi-layer perceptron with none of the attention.

**YB** I disagree I’ve mentioned that Boltzmann machines they do just that!

**GM** They’re not gonna be able to make reasoning over quantified statements and so forth, at least not that I’ve ever seen.

**YB** Well, they haven’t been explored recently, but this is essentially what they were designed for.

**GM** We can place a public bet afterwards about Boltzmann machines and their ability to deal with quantification, *every*, *some* and so forth. I would say, by and large, that the results of extant neural networks on reasoning are not as impressive even as that example from CYC, but I would also say I was going to give you the point and then we can come come back. If you take the broader notion of deep learning that Yoshua would like to defend and you start putting in mechanisms for attention and indirection and so forth, which come at least a little bit close to the things that I want, then all bets are actually open. We don’t know yet what the boundaries are once you include mechanisms like indirection. We know some of the things we can do there. There’s a lot of stuff in classical reasoning, I don’t think has really been addressed yet. There are other people more expert in that. I would say, even just dealing with quantified sentences, how do you deal with “*everybody love somebody*” in the ambiguity and that. We haven’t really seen that yet.

**YB** There’s a question here about: What do you think of transferring structured rules in the form of first order logic onto network parameters as opposed to encoding the information in latent variables? This is actually the kinds of things that people were trying to do in the 90s. Trying to create a direct analogy between our representation, mapping between representations of knowledge in the weights between neurons, and the rules in logic. Personally, I don’t think this can work for a number of reasons. On the other hand, what I think can work, and in a way we’re already doing it, is neural nets that can acquire knowledge by reading documents just like humans do, or reading databases, like knowledge bases. This is something of course that we can do a lot better. Right now, I don’t think we have the right tools for this, but we’re making progress. The kinds of things that people are using now, transformers, I think can be evolved into what we need, especially if we couple them with better world models.

**GM** Is it’s clearly still an open question. I think the most interesting work right now is done by, or some of the most interesting work, in that specific questions, done by [Artur d’Avila Garcez](http://www.staff.city.ac.uk/~aag/?id=3), who’s trying to build hybrid models where you have explicit representation of logical formalisms and you can map it in on to a deep learning system. It’s still I think early days for that work, but it’s interesting and it’s another perspective.

**YB** (…) to make sure that those initial injected assumptions will still hold after the training, overcome catastrophic forgetting. The first question, I think, we just, at least I gave, a partial answer to this. The second one about forgetting is very important. It is connected to some of the things I was talking about when I mentioned factorizing knowledge in two pieces so that when there’s a change in the world, a change in task, a change in distribution, new piece of knowledge gets added. It doesn’t require the whole system to be adapted, but only a few parts of it that explained that change. If we’re able to do that, which for now we’ve done on a very small scale, but if we’re able to do that on a large scale. then I think we can overcome catastrophic forgetting. We can build systems that adapt in just the necessary ways without having every neuron and every weight trying to be part of explaining the change that just happened or a new task.

**GM** If I could push some of Yoshua’s fantastic students, who are probably sitting in the room, to study one question that they might not be studying so much right now, it would be that first question here: “*how to inject knowledge into deep learning models and frameworks*”. There are a lot of people in the field thinking about this. I don’t think there are a lot of people at the CYC-level…

**YB** We had a paper, a few years ago, that does the things that I was talking about earlier in the sense that we have a language model that while it’s reading text, for example like while is it reading a question or trying to complete a sentence, is looking up in a knowledge base, a structured knowledge base, with you know like subject-object-verb things like this standard relational databases and looking for those words that it has seen, or their equivalent synonym representations, in the knowledge base, and then using attention mechanism, picking the pieces of knowledge, in the knowledge base, which can help it predict what the next word should be. This was done with \_\_\_\_\_\_\_\_\_\_\_. It’s been it’s been published. What it allows is models that can do their normal neural nets thing, but as they are computing is like they’re allowed to go online and check for information that they don’t already know, that is not already integrated into their inside brain, and use that information in order to answer questions or predict something.

**GM** I’m gonna raise a technical issue and make an advertisement. The technical issue is: I wonder how well (you have to answer it now, but how well) it works with quantified statements and negation as opposed to triples which you know, …

**YB** I don’t know. At that time, we were not looking at that and I think it’s only recently, with attention mechanisms, in the form that involves indirection, that you can start thinking about quantification. Quantification, the way I interpret it, in a neural net sense, is essentially that you have these little modules, which in your world you would call rules, and that’s fine, except there are not symbolic rules, they’re just more like they allowed to do inference on some variables given other variables. The inputs of those rules don’t have to be always the coming form the same place. The inputs have types, just like functions in programming in C++ have types, and so they expect their input to have the right type to match together and when a rule matches what is in the data, it can be triggered, just like in production systems.

**GM** I’ll get to the advertisement in a second. If we could work on one thing together, that would be it. The advertisement is: Vince and I are going to put together a set of [readings](https://montrealartificialintelligence.com/aidebate/), after the debate, so people can follow up on some of the issues that we talked about today. My first nomination is the paper that you just suggested.

**YB** If we are to build real will the AI systems, how feasible is the current practice of training deep learning networks end to end, knowing fully well that they are going to be huge in this regards? For me, the end-to-end thing is mostly a problem when you consider biological plausibility. Because, there’s a long delay between information being propagated from one part of your brain, say in the back, to the front part of your brain, so that the number of back and forth exchanges that can happen, in the time that you are able to answer a question, like half a second, is very short. You can go like back and forth a couple of times. It would be reasonable to assume that although there is coordination at a global level, a lot of the learning involves local effects. There’s been a lot of interesting work in deep learning. I don’t think, right now, we’ve solved this problem where people are trying to predict the gradient that would eventually come back if you were to do end-to-end learning and then use that to start the learning, in in a sense. If you look at reinforcement learning systems, they use that kind of trick as well, to predict the reward that you will get, and use a predicted reward as an intermediate local reward. I think there’s some interesting questions about decentralizing this kind of learning. There is also more pragmatic explorations and things like federated learning where people are trying to build deep learning systems that can learn on local nodes, like on people’s phones and things like this, without having the data on those phones necessarily travelling to some central repository. I think this raises all kinds of interesting questions. There’s also a lot of interesting connections to multi-agent learning. One of the hot topics in machine learning these days is how do we have multiple neural nets interact with each other, each learning from its own objective function, but in a way there’s a social thing going on where they’re together trying to solve the problem. I think that’s another way that you can decentralize the learning.

**GM** I think we’re almost out of time. I’ll add one thing to that and I think Vince maybe goes next. I think that modularity, in the sense that Jerry Fodor was talking about once upon a time, of having individual components that are tuned to particular things, is the heart of how the brain works. it’s not fully modular, but I think the most amazing thing about the imaging literature, taking pictures scans of people’s brains, is the way in which the brain, now in a connect to a phrase of yours, dynamically reconfigures itself in the course of anything that we do. You can tell someone who’s coming into a scanner experiment, like the ones that [Ev](https://evlab.mit.edu/ev-fedorenkos-language-lab) is going to do, okay what you’re gonna do now is you’re gonna take glasses and you’re gonna put them onto the head every time you hear the word *blue*. Then people, in the space of three seconds, dynamically reconfigure their brain in order to be able to do that. I don’t think that end-to-end deep learning is capturing that, but some of the dynamic reconfigurability that they Yoshua is fighting for…

**YB** That is deep learning too.

**GM** I mean there’s some deep learning there, I mean again one…

**YB** No, it is deep learning! Seriously.

**GM** Deep learning that allows you to do the reconfiguration?

**YB** Yeah.

**GM** In what sense?

**YB** It is just deep learning with gates! We’ve had gates for since 1989 or something like this.

**GM** Gates, I’m with you and we can argue about the definition later.

**YB** Of course we’ve made progress but I mean it’s not really a completely new idea. Gating computation. Neuroscientist have been talking about your neuro-modulators is forever. Just just remove from your brain the idea that deep learning is in 1989 MLP with feed forward connections. That is not deep learning! Sorry.

**GM** We can argue about the scope. I will end with the agreement, which is: I think the gates are the solution. The question was framed in terms of end to end deep learning and end to end deep learning typically is the closest thing to the kind of thing that I’m critiquing…

**YB** Not today’s deep learning state-of-the-art. That’s not how.

**GM** I’m all for the gates.  
  
**VB** Distinguished Participants, Ladies and Gentlemen, we just had a hugely impactful “**AI Debate**”! My thanks to Gary Marcus and Yoshua Bengio and to Mila for hosting us.  
  
**GM** I want to thank Vince for even having the idea to do this and for Yoshua for being gracious enough not only to do it but to host it.  
  
[Applause]  
  
**VB** The conversation will continue on social media with the hashtag: #***AIDebate***. A decade that has revived the field of AI is ending with this AI Debate. A new decade will soon begin with the best way forward for AI.

Stay tuned for the announcement of follow-up events and for the unveiling of the next MONTREAL.AI World-Class Event.

Good night every one.