Transcript of the AI Debate

MONTREAL.AI

Jan 1 · 36 min read

DEBATE : Yoshua Bengio | Gary Marcus

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

INCOMPLETE DRAFT — WORK IN PROGRESS

Gary Marcus

Yoshua Bengio

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)

At **Mila in Montreal**, on Monday, December 23, 2019, from 6:30 PM to 8:30 PM (EST), **Gary Marcus** and **Yoshua Bengio** debated on the best way forward for AI.

5,225 tickets were sold for the international live streaming event. There was quite a twitter storm after the ***#AIDebate***. ZDNet described the event organized by MONTREAL.AI as a "historic event".

Slides, readings and more on the MONTREAL.AI debate webpage.

Transcript of the AI Debate

Opening Address | Vincent Boucher — 3 min.

Good Evening from Mila in Montreal Ladies & Gentlemen,

Welcome to the "**AI Debate**".

I am Vincent Boucher, Founding Chairman of Montreal.AI.

Our participants tonight are Professor **GARY MARCUS** and Professor **YOSHUA BENGIO**.

Professor GARY MARCUS is a Scientist, Best-Selling Author, and Entrepreneur. Professor MARCUS have 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.

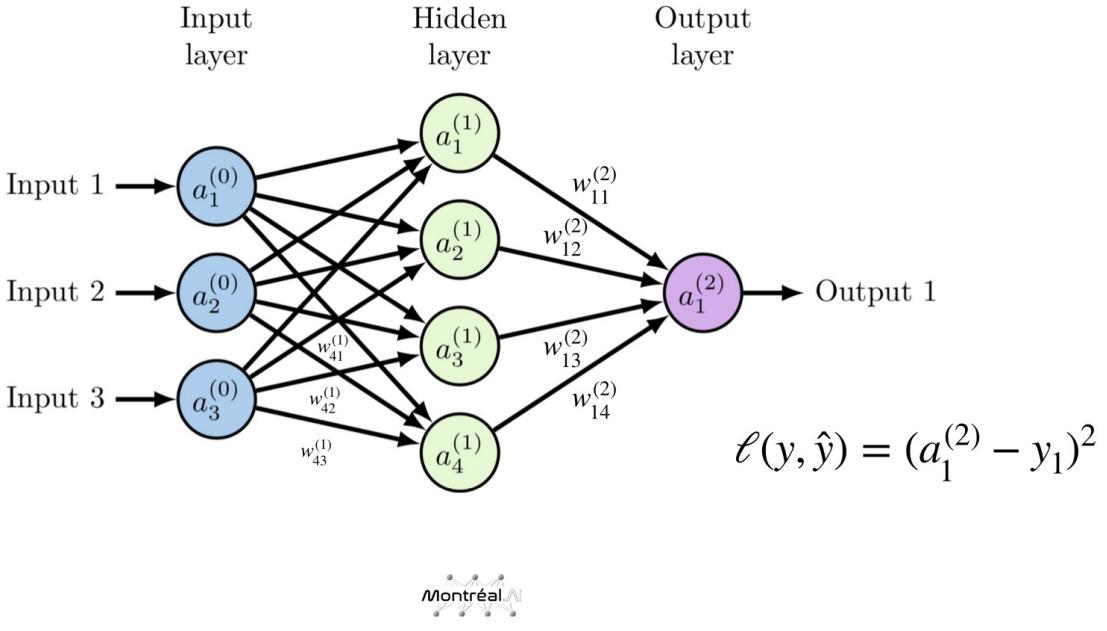


Diagram of a 2-layer Neural Network

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.

AI DEBATE : YOSHUA BENGIO | GARY MARCUS

AGENDA : THE BEST WAY FORWARD FOR AI

6:30:00 PM EST : **Opening Address** | ***Vincent Boucher*** — *3 min.*
6:33:00 PM EST : **Opening statement** | ***Gary Marcus*** — *20 min.*
6:53:00 PM EST : **Opening statement** | ***Yoshua Bengio*** — *20 min.*
7:13:00 PM EST : **Response** | ***Gary Marcus*** — *7.5 min.*
7:20:30 PM EST : **Response** | ***Yoshua Bengio*** — *7.5 min.*
7:28:00 PM EST : **Interview** | ***Vincent Boucher : Yoshua Bengio & Gary Marcus*** — *15 min.*
7:43:00 PM EST : **Public Questions** | ***Yoshua Bengio & Gary Marcus*** — *22,5 min.*
8:05:30 PM EST : **Int'l Audience Questions** | ***Yoshua Bengio & Gary Marcus*** — *22,5 min.*
8:28:00 PM EST : **Closing Remarks** | ***Vincent Boucher*** — *2 min.*

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

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 — 22 min.

Opening statement | Gary Marcus — 22 min.

Thank you very much.

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

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 importants 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

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

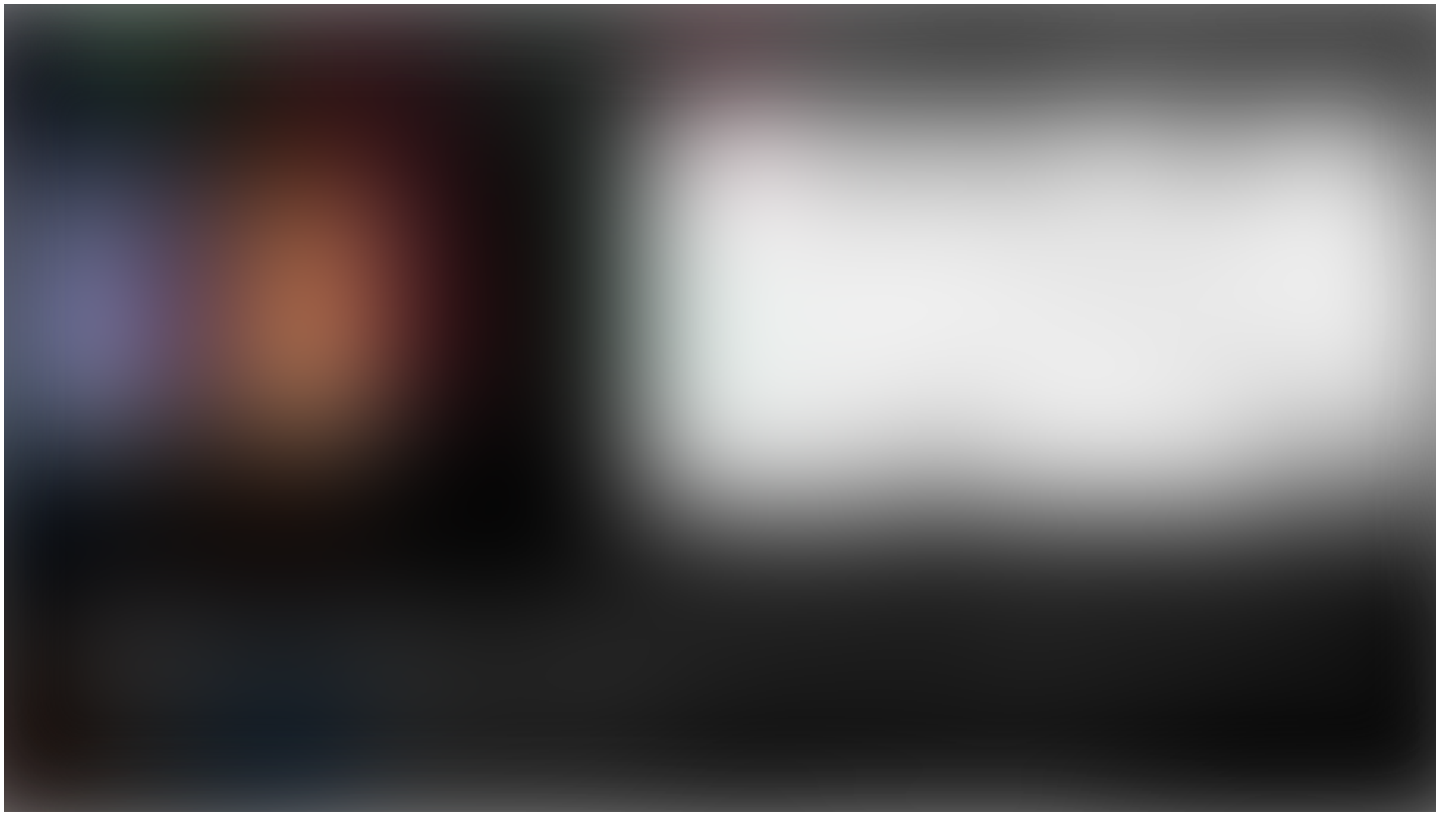
machine learning and how I got here. It's a bit of a personal history of cognitive science and how it feeds into AI. And, you might think: "*What's a nice cognitive scientist like me doing in a place like Mila?*".

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. 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 the 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 though 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 *in principle*, 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 get myself up to graduate school to work with Steve Pinker at MIT and what I looked at were these errors. I did I think the first big data analysis of language acquisition. I am one of the first to

write shell script on Unix Spark stations 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 (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 say it was a hybrid model we said best capture the data. A rule for regulars so walk is inflected in walked in 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 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 networks models.

There's 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 could 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 times later), and the conclusion was that the classical limits of connectionists models that is currently popular couldn't learn to extend universals outside of the training space.

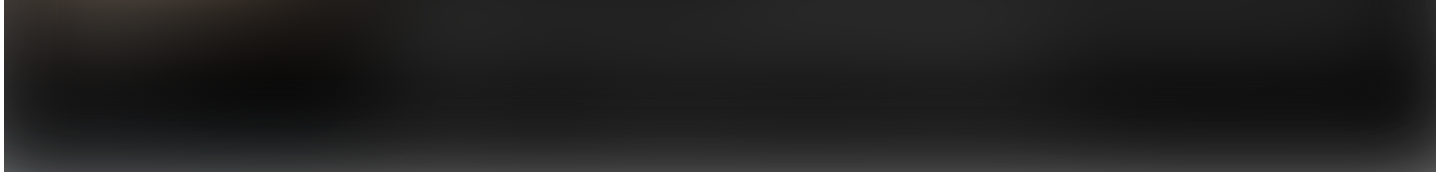
In my view this is the thing that I'm the most proud of having worked on. Some details for later…

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 fee*". 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

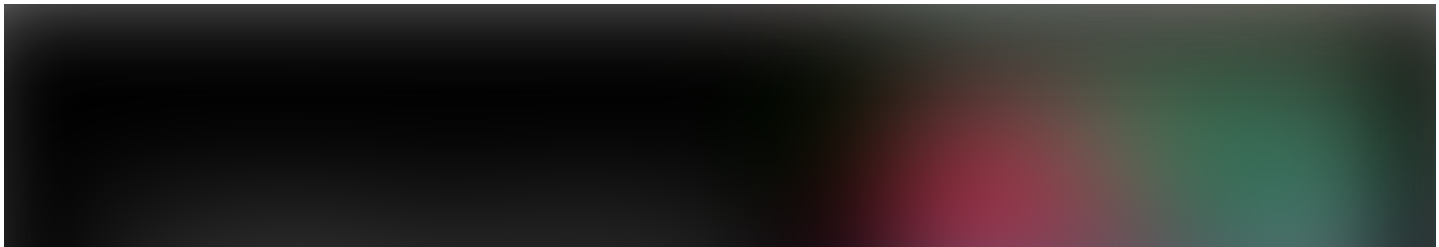
And the key component of the thing I was defending, which I called symbol-manipulation (I didn't invented it, but I tried to explicated it and argue for it), are *variables*, *instances*, *bindings* and *operations over variables*. You can think in algebra where you have a variable like x, you have an 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 and 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 integrating 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 with a lot of which has been using these symbols and so for. And so 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

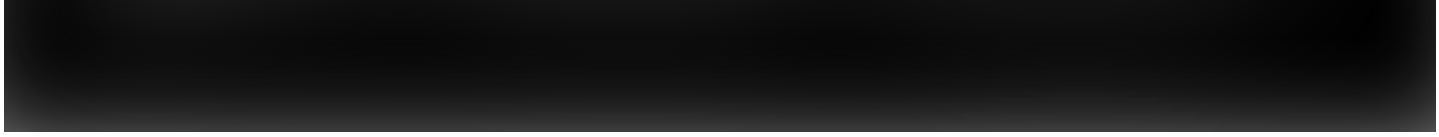
I was arguably ignored I think in candor until a year or so ago. People I think started paying attention to the book again. But, it did inspire a seminal book on neuro-symbolic approaches which I hope some people will take a look at, called *Neuro-Symbolic Cognitive Reasoning* and I'm going to try to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

2012: The Rise of Deep Learning

I stoped 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 machinerie.

2018: 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*). It outlines ten problems for Deep Learning (I think it was on the suggested readings for here) 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 flacks 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

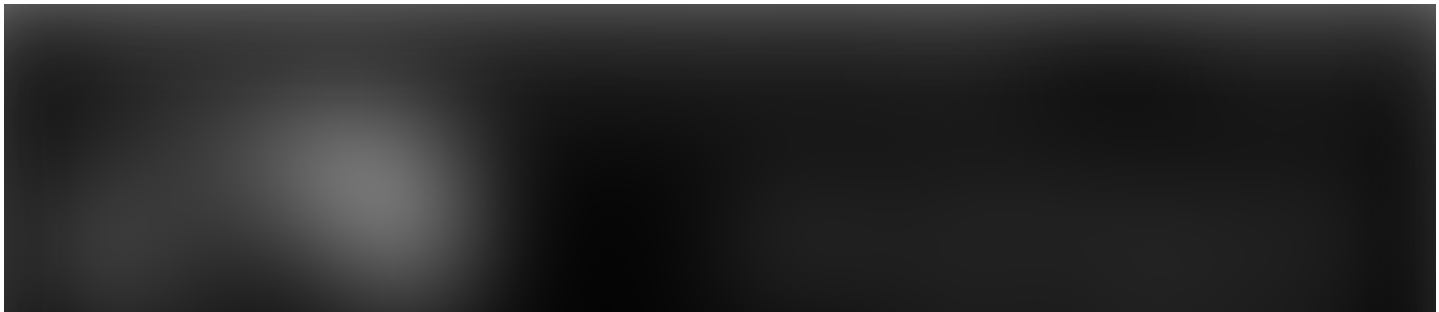
Part II: Yoshua

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.

First 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, squiring 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.

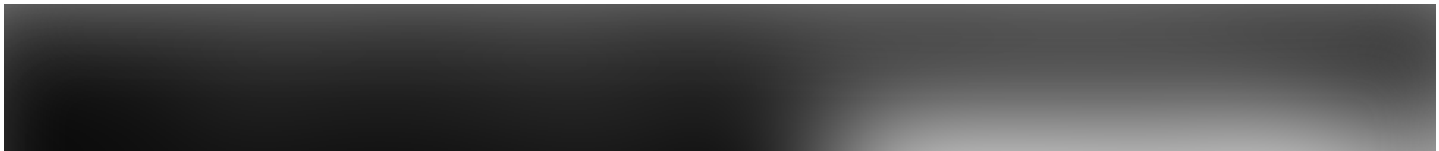


My differences are mainly with Yoshua's earlier (e.g., 2014–2015) views

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.



Recently, 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.

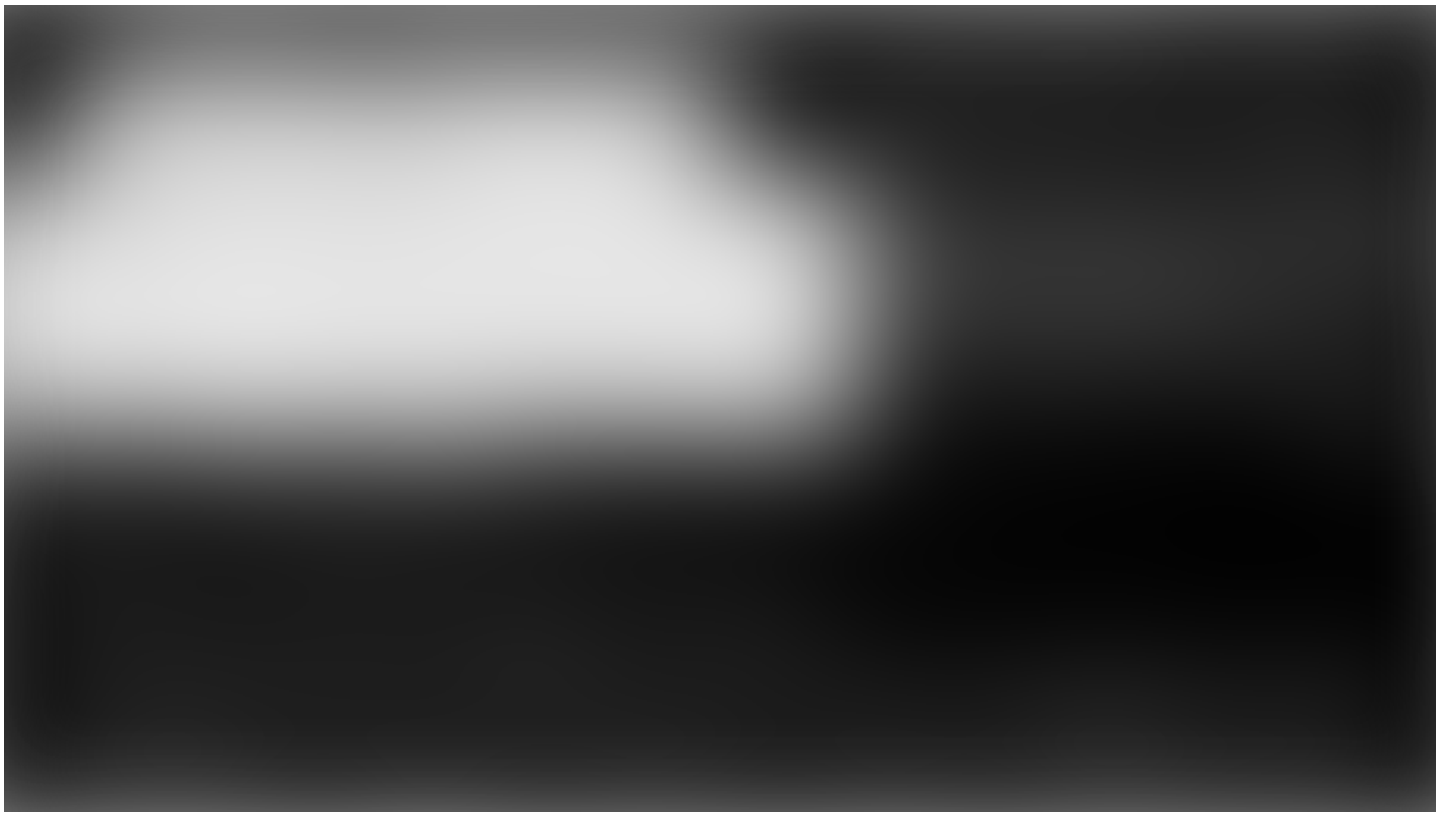
I'll take about my position, the right way to build hybrid models, innateness, the significance of the fact that the brain is a neural network and what we mean by compositionally.

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

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

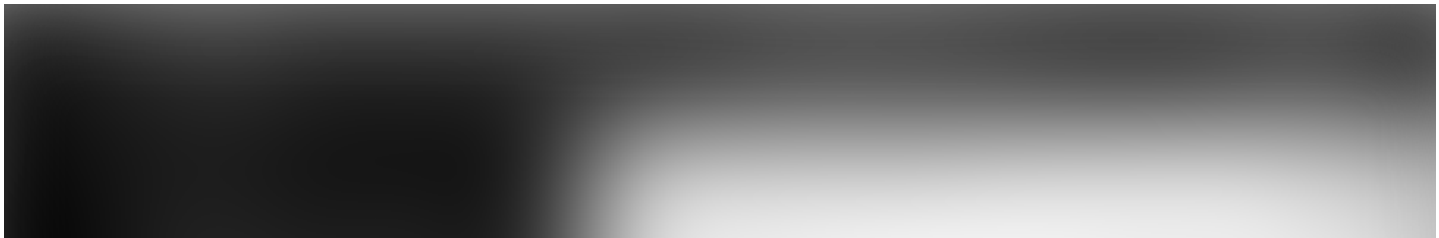
The first one is the most delicate. But, I think occasionally Yoshua is misrepresenting me as saying "*look, deep learning doesn't work*", he

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.



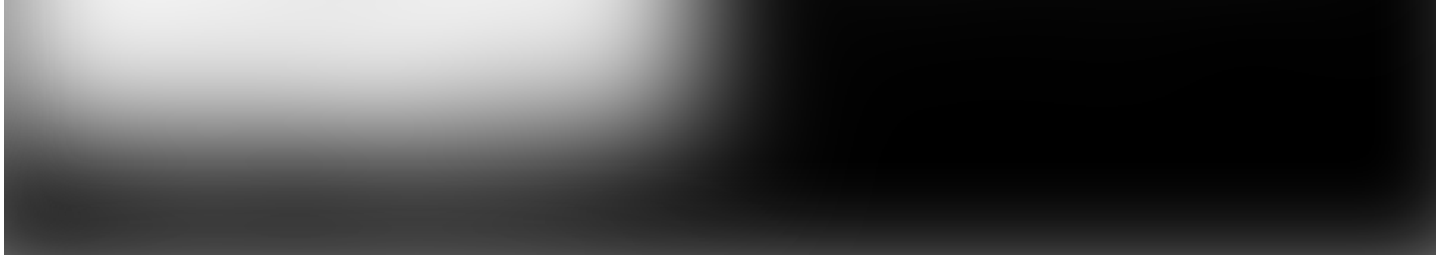
1. Yoshua's (mis)representation of my position (2 of 2)

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 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?

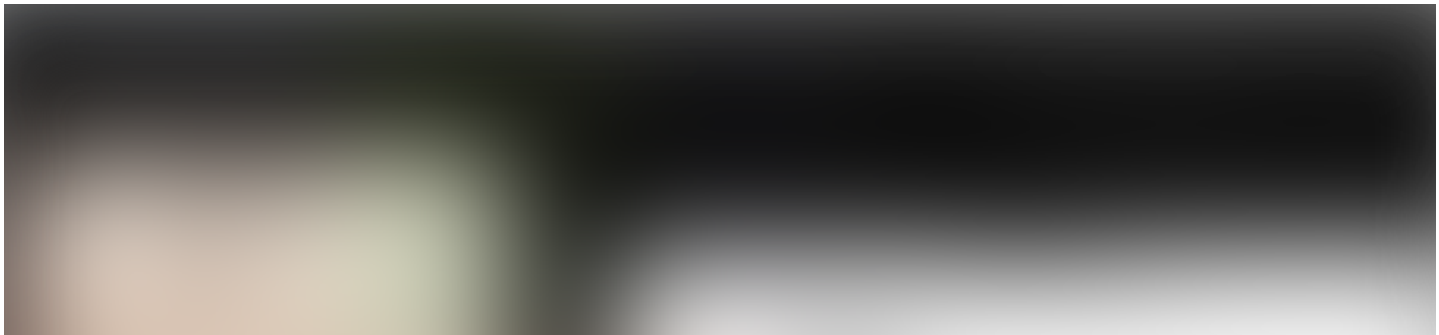


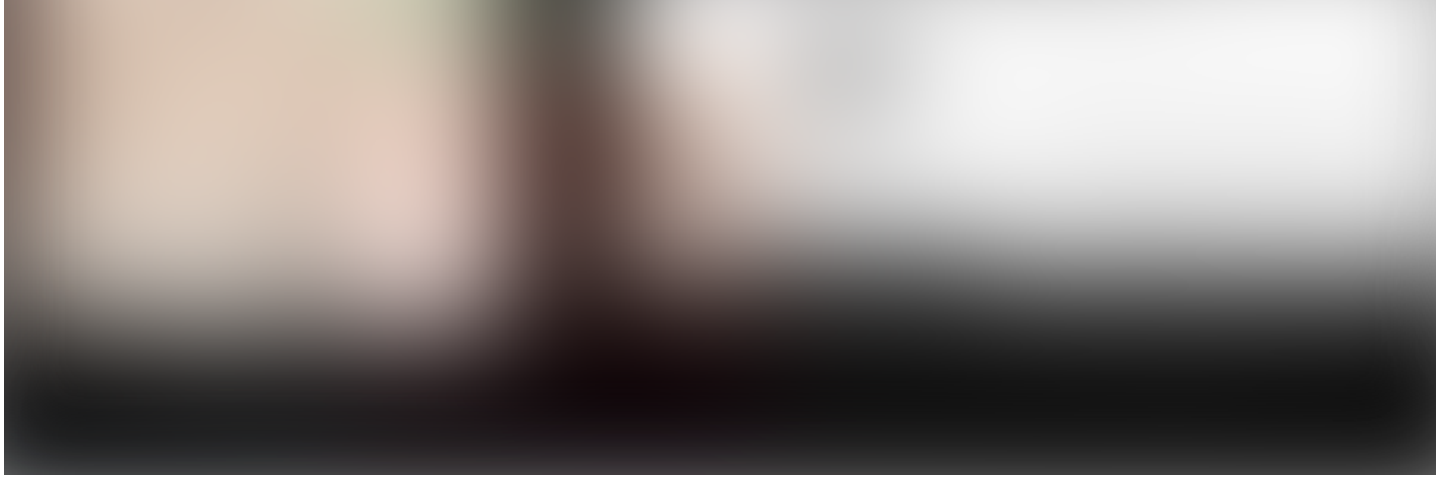
To argue against symbol-manipulation, you have to show that your system doesn't implement symbols

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, 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 ou 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 thing Yoshua made that argument yet. Maybe he will today.

I think that this is 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 remins 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 for. 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 for 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, whic is classig 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.

Mao et al, arXiv 2019

There is great work 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.

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 differents models. The ones that had more innateness in terms of convolutional prior were the ones that did better.

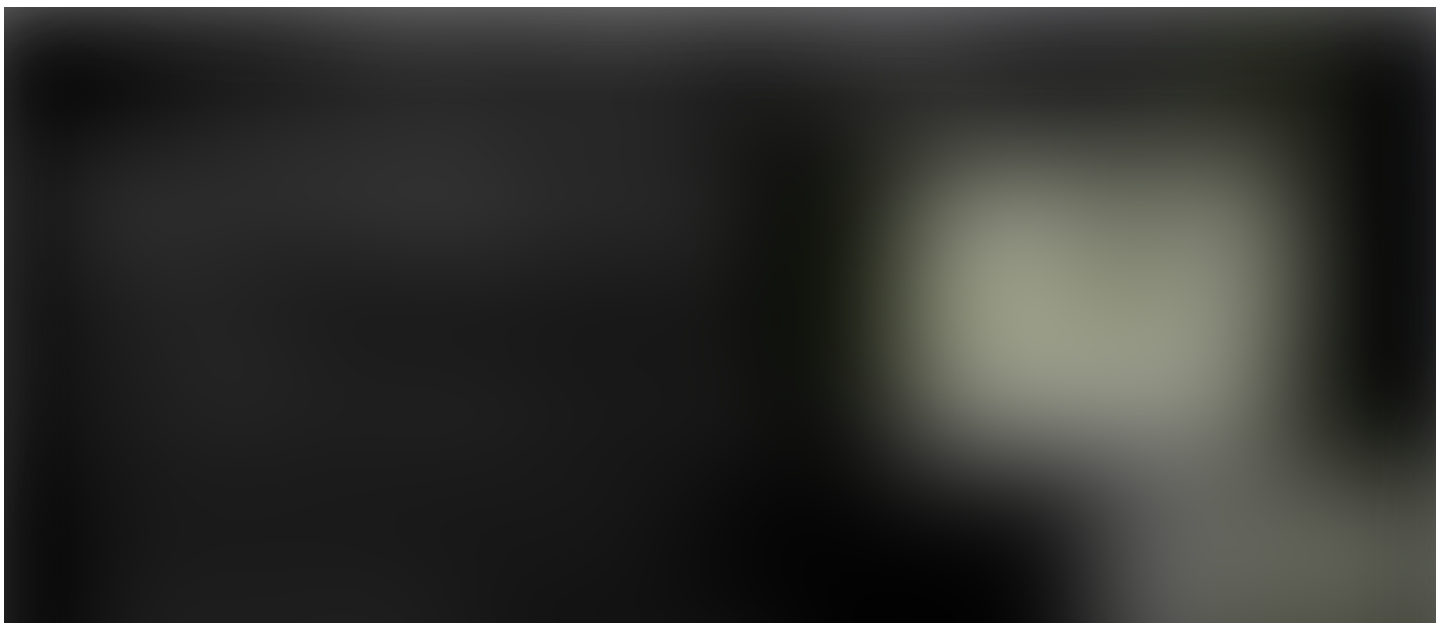
This is a picture of a baby ibex climbing down a mountain. I don't think the anybody can reasonably say that there is nothing innate about the baby ibex: it has to be born with an understanding of the 3 dimensional world and how it interacts and so for in order to do the things that it does. So nativism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

4. 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. 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 the same argument to me on Facebook by saying: *your brain is a neural net all the way*.

First, 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 neuon 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 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 neurones. 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.

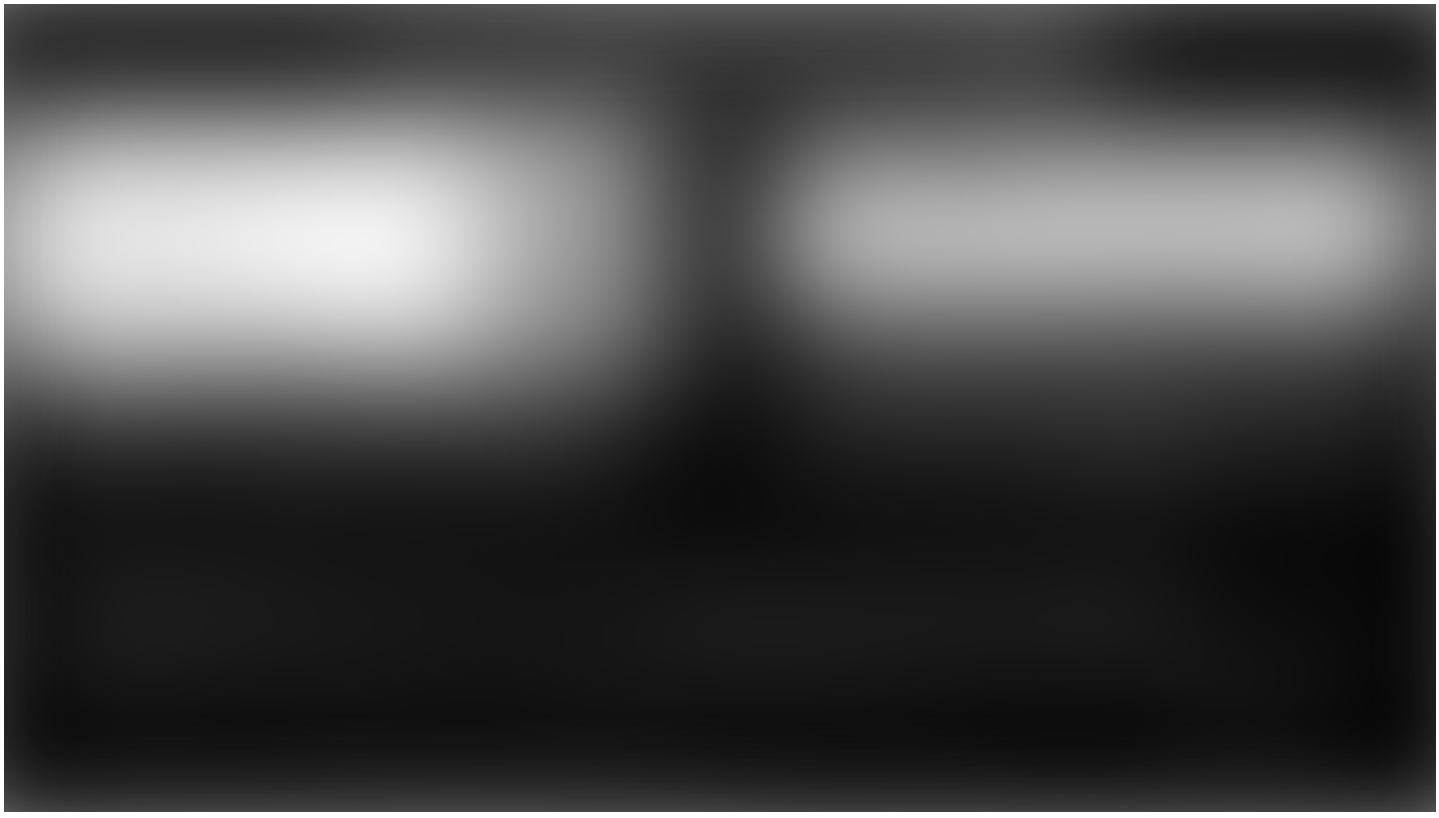
"Symbols aren't biologically plausible"

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 counscious 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.

Even if somehow turned that the brain never manipulated symbols, why exclude them from AI?

Even if somehow turned that the brain never manipulated symbols

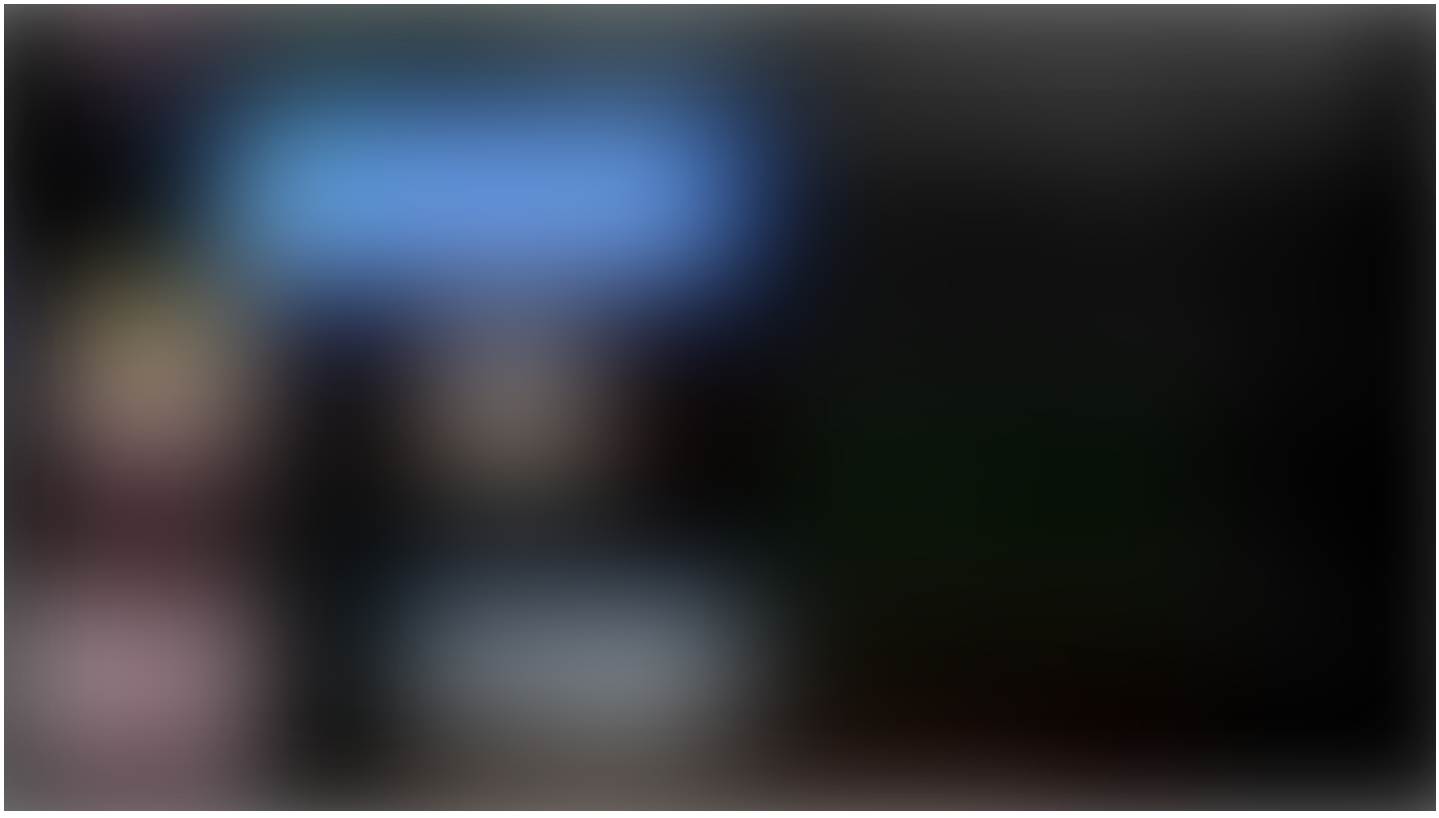
(which is counterfactual to our world), why exclude them from AI? We can't prove that they are inadequate, they have proven utility: a large fraction of the world's computers programs are written in (pure) symbol-manipulating code and a large fraction of the world's distilled knowledge comes in the form of symbols: eg. most of Wikipedia is in written, symbolic form and we want to leverage that in our learning systems.



5. 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."

Vincent Boucher: Professor Marcus, you have 30 seconds.

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.

Compositionality isn't just about language...

Compositionality isn't just about language... It's also learning about differents concepts and putting them together in differents 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 takeway 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, 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 — 22 min.

Opening statement | Yoshua Bengio — 20 min.

Welcome to this debate.

Debate 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.

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*", 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 concious 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 ___33:55___ 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 concious 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 conciousness. 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 uncouncious 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 *Sparse 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 conciously, 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 equiped 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) 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 which comes fom 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 collegues Léon Bottou, Martin Arjovsky 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-learnig 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 | Yoshua Bengio & Gary Marcus — 15 min.

INCOMPLETE DRAFT — WORK IN PROGRESS

)

Artificial Intelligence | Ai Debate | Deep Learning | Symbolic Ai | Innateness