How Children Learn Once You Subtract Language: Statistical Learning, Causal Intervention, Curiosity, and Few-Shot Priors as Buildable Healthcare AI Point Solutions Rather Than One Giant World Model

JUN 24, 2026

1

Share

Video Preview

🎧 Part I Podcast free on Spotify.

Thoughts on Healthcare Markets & Technology

🎧 **Part I: How Children Learn Once You Subtract Language: Statistical Learning, Causal Intervention, Curiosity & Buildable Healthcare AI Point Solutions Rather Than One Giant World Model**

Children learn without labels, without feedback, and on almost no data. Healthcare AI still struggles with all three. Here is why the gap is a product opportunity…

▶ Listen now

3 days ago · Thoughts on Healthcare

🎧 Part II Podcast epis subscribers only. Also Spotify.

Thoughts on Healthcare Market

🎧 **Part II: How Children Lear Language: Statistical Lear Curiosity & Buildable Heal Rather Than One Giant Wo**

Children learn without labels, withou data. Healthcare AI still struggles wi is a product opportunity…

▶ Listen now

3 days ago · Thoughts on Healthcare

Discover more from Thoughts on Healthcare Markets & Technology

Expert analysis of healthcare and life sciences markets, technology, investment, entrepreneursh policy, and AI — for investors, entrepreneurs,..

Over 4,000 subscribers

Enter your email...

Subscribe

By subscribing, you agree Substack's Terms of Use, and acknowledge its Information Collection Notice and Privac Policy.

Already have an account? Sign in

To listen to paid episodes in Apple or Spotify, link your Substack subscription via the show settings on those platforms (instructions inside the Substack app under Subscriptions → Podcast).

Abstract

This piece sets aside language, the part of child learning that looks most like today's chatbots and therefore soaks up most of the attention, and walks through the specific non linguistic mechanisms kids actually use to learn. The argument is that each mechanism is a narrow, buildable product, while the catch all idea of a general world model is a category rather than a thing anyone can ship. Five core mechanisms get the deep treatment, each paired with a healthcare problem where that exact mechanism is the missing piece, plus one cautionary mechanism and a short synthesis.

- Statistical learning is self supervised pretraining. Healthcare fit: foundation models on abundant unlabeled signal such as continuous glucose, fetal monitoring, single lead ECG.
- Causal intervention is the do operator. Healthcare fit: a causal layer on linked claims and EHR that flags when observation cannot answer and an experiment is the only honest move.
- Curiosity and the Goldilocks effect are active learning. Healthcare fit: uncertainty driven annotation and information gain test ordering.
- Few shot learning on strong priors is meta learning. Healthcare fit: rare disease flagging built on phenotype ontology priors over tiny case counts.
- Prediction error and violation of expectation are personalized anomaly detection. Healthcare fit: per patient baselines that go after alarm fatigue.
- Overimitation is cargo cult imitation learning. Caution: ambient scribes that inherit EHR rituals along with the reasoning.

Thoughts on Healthcare Markets & Technology is a reader-supported publication. To receive new posts and support my work, consider becoming a free or paid subscriber.

Table of Contents

The framing problem, or why world models are too big to build against

The decks all say the same thing now. Build a world model, give the machine a general sense of how reality behaves, and everything downstream falls out for free. It is a

lovely story and it is also the venture equivalent of saying the cure for cancer is biology. True at some altitude, useless at the altitude where somebody has to ship by Q3. Kids are the existence proof everyone gestures at, since a four year old picks up physics, other minds, cause and effect, and a firm theory of which foods are gross, all without a training run that costs more than a mid sized country. But pointing at the four year old and saying general world model is the move that boils the ocean. The interesting question for anyone trying to build a company is narrower and far more useful.

Strip out language, which gets almost all the oxygen because it is the part that looks most like the chatbots, and look at the specific, separable mechanisms a child runs to learn. Each one is a distinct computational trick. Each one has been studied for decades by people who put babies in front of puppet shows for a living. And several of them map almost one to one onto a healthcare problem that is currently solved badly or not at all. The throughline here is simple. World models are a category, not a product. Statistical learning is a product. Causal intervention is a product. Curiosity is a product. The kid is not running one giant model. The kid is running a stack of cheap, specialized routines that happen to compose well, and the compose well part is the only place the word general earns its keep.

Statistical learning, the original self supervised pretraining

Start with the oldest and most robust result in the entire field. Saffran, Aslin and Newport, 1996. Eight month olds listen to two minutes of a monotone synthesized speech stream with no pauses, no stress, no cues of any kind except the raw statistics of which syllables tend to follow which. After two minutes the babies reliably tell the difference between a word from the stream and a near identical part word that straddled a boundary. Two minutes. No feedback, no rewards, no labels, no teacher standing over them. They are computing transitional probabilities more or less in their sleep. The kicker is that this is not a language gadget bolted onto the side of the brain. Kirkham and colleagues showed the same machinery running on shapes and on tones in infants as young as two months. It is a domain general statistics engine that eats whatever sequence you feed it and quietly extracts the structure.

If that sounds familiar it should, because it is self supervised learning before the term existed. Predict the next thing, notice what co-occurs, build representations out of the raw stream without anyone hand labeling a single example. The whole reason foundation models work is that labels are expensive and structure is free if you are

willing to predict. The infant figured this out without a GPU and on a power budget of roughly a banana.

The healthcare angle is direct. The labeled data problem in medicine is brutal and everyone knows it. A board certified specialist annotating pathology slides or echocardiograms costs real money per hour and there are nowhere near enough of them. Meanwhile the unlabeled signal is everywhere and mostly thrown away. Continuous glucose monitor traces, ICU telemetry, fetal heart rate strips, ambulatory ECG, ventilator waveforms, raw accelerometry off a wrist wearable. Oceans of sequence with almost no labels attached. The infant move is to pretrain on the raw stream first and worry about labels later, which is exactly what the EHR foundation model crowd has been doing with models in the Med-BERT lineage, treating a patient timeline as a sentence and the codes as tokens. The narrow buildable version is not a foundation model for all of health, which is itself a boil the ocean phrase. It is one modality where the unlabeled stream is enormous, the labels are scarce and pricey, and a self supervised pretraining run turns a thousand labels into the effective power of fifty thousand. Pick continuous glucose, or fetal monitoring, or single lead patch ECG. That is a company, not a research program.

The child as experimenter, and the correlational swamp of claims data

Now the part that separates a kid from a correlation engine. Gopnik and colleagues spent years on a beautifully simple toy called a blicket detector, a box that lights up and plays music when certain objects get placed on it. Hand a preschooler a pile of blocks and this box and the kid starts running experiments. Not metaphorically. One block on, take it off, try a different one, try two at once, watch what happens each time. By age four they are doing something that looks an awful lot like Bayesian causal inference, sorting which object actually drives the box from which one merely happened to be sitting nearby. Schulz and Bonawitz ran a lovely study where children played far more with a toy when its causal structure was confounded, when they could not tell which lever did what, and settled down once the structure became clear. The extra play was not random fidgeting. It was information seeking. The kid was running interventions specifically to break the confound.

This is the single most underrated thing children do and it is precisely the thing most healthcare machine learning does not do. The bulk of clinical AI is trained on observational data, claims and EHR exhaust, and observational data is a swamp of confounding. The patients who got the drug differ from the patients who did not, in ways both measured and hidden. A model that learns this treatment associates with

worse outcomes may have simply learned that sicker people get the treatment. Every health economist knows this in their bones, which is why the field carries whole liturgies around propensity scores, instrumental variables, and target trial emulation, the Hernán framework that tortures an observational dataset into behaving like the randomized trial nobody ran.

Here is the kid lesson, stated plainly. The child does not just compute fancier statistics on the same observations. When the observations are confounded, the child intervenes. The buildable product is a causal layer that sits on top of large linked datasets and does two jobs. It estimates effects where the data can actually support a causal claim, and, more importantly, it flags the cases where the confounding is unresolvable from observation alone and an actual experiment is the only honest answer. Most tooling pretends every question can be wrung out of the data on hand. The four year old knows better. A platform that says this one you can estimate, this one you need to go poke the box is worth more than one that confidently hallucinates effect sizes across the board. Pair that with the pragmatic and platform trial infrastructure that is finally maturing and there is a real wedge sitting there.

Curiosity as a reward function, and which scan to label next

Kids are not passive sponges and they are not exhaustive crawlers either. They are picky about where they spend attention, and the rule they follow is elegant. Kidd, Piantadosi and Aslin called it the Goldilocks effect. Seven and eight month olds watching a stream of events look away from things that are too predictable, because boring, and also look away from things that are too surprising, because hopeless, and lock in on events of intermediate complexity. They are implicitly maximizing learning progress, spending the scarce currency of attention exactly where the expected information gain is highest. Oudeyer and Kaplan built a whole research program out of formalizing this as intrinsic motivation, a reward signal that comes not from any external goal but from the rate at which the learner is getting better at predicting. Schmidhuber framed roughly the same idea in the language of compression progress. Curiosity here is not a personality trait or a vibe. It is an active learning policy with a loss function.

The healthcare version is almost embarrassingly on the nose, because the bottleneck in medical AI is annotation, and annotation is rate limited by expert attention, which is the single most expensive resource in the building. Say a model has ten thousand unlabeled chest films and budget for a radiologist to label five hundred. The dumb approach labels five hundred at random. The infant approach labels the five hundred

the model is currently most uncertain about, the ones sitting in the Goldilocks zone where a label actually moves the needle. This is active learning, it is old news in the literature, and it is shockingly underused in deployed clinical pipelines where people still ship random batches to overworked specialists and call it a day. A point solution that wraps any clinical model in an uncertainty driven query loop, picks the next most informative case, and routes only that case to the human can cut annotation spend by large multiples without touching model architecture at all.

The same idea climbs straight out of the lab and into the clinic through test ordering. A big chunk of the diagnostic odyssey is a failure to order the test that maximally reduces uncertainty about what is going on, and instead ordering the test that is cheap, or habitual, or defensively reflexive, or whatever the order set happens to default to. A system that ranks candidate tests by expected information gain about the live differential is just the toddler with the blicket box, except wearing a white coat and arguing with a prior authorization queue. The interesting thing is that the value here is not better imaging or better assays. It is a smarter policy over which question to ask next, which is exactly the thing the baby is optimizing and the order set is not.

Few shot generalization, strong priors, and the rare disease odyssey

The thing that quietly unnerves the big model crowd about kids is sample efficiency. A child learns a new word from one or two exposures. Carey and Bartlett showed this back in the seventies with an invented color word, chromium, that children slotted into their vocabulary after a single offhand mention and then retained. Large models need to see a concept thousands of times. The child needs to see it once. The reason is not a bigger brain. It is that the child walks in with enormous priors already installed. Spelke spent a career documenting core knowledge, the small set of systems infants seem to arrive pre wired with, objects that persist and do not pass through each other, agents that have goals, a rough sense of number, basic geometry. Layered on top sit learned inductive biases like the shape bias, the assumption that a new word for an object probably picks out its shape rather than its color or material. The few shot magic is downstream of the priors. Strong priors plus one example beats weak priors plus ten thousand, every time.

Now aim that at the most expensive few shot problem in all of medicine, which is rare disease. There are on the order of 7,000 recognized rare diseases, collectively affecting tens of millions of people, and the average patient burns years and parades through a half dozen specialists before anybody names the thing. By definition there are not ten thousand clean training examples per disease sitting in a bucket somewhere. There

might be eleven. This is the exact regime where brute force dies and the child approach wins, because the fix is not more data, it is better priors plus genuine few shot generalization. The buildable version encodes the strong priors the way the kid does, through structured phenotype ontologies like the Human Phenotype Ontology, gene to phenotype maps, known inheritance patterns, and then does few shot matching against a handful of confirmed cases rather than demanding a fat labeled cohort that will never exist. The better rare disease modeling work showing up in the serious journals lately leans precisely this direction, foundation model representations fine tuned on tiny case counts. The product is a few shot rare disease flagger that treats eleven examples as plenty because it brought its priors to the party. A toddler does not need a thousand giraffes to learn giraffe. The model should not need a thousand cases of a one in a million disease before it is allowed to suspect it.

Prediction error, violation of expectation, and the war on alarm fatigue

Babies are tiny prediction machines, and the proof is in what surprises them. Baillargeon and others built an entire method, violation of expectation, around the plain fact that an infant stares longer at something impossible. Show a baby a screen that appears to rotate straight through a solid box that should have stopped it cold, and the baby looks, and keeps looking, because the prediction got violated and the violation is the interesting part. That is the behavioral fingerprint of a brain constantly forecasting the next instant of sensory input and treating the gap between forecast and reality, the prediction error, as the signal worth learning from. The predictive processing lineage, Friston and Clark and that crowd, spun this into a grand unified theory of cortex, but the buildable nugget does not require buying the whole metaphysics. The nugget is that surprise relative to a personal baseline is a far better learning and alerting signal than distance from a population threshold.

Which lands on the most quietly destructive problem in hospitals, alarm fatigue. Monitors fire constantly, the overwhelming majority of alerts are false or clinically meaningless, and staff learn to tune them out, which is of course the precise moment the real one slips by unnoticed. The reason the alarms are so useless is that they trip on fixed population thresholds. Heart rate over some number, oxygen saturation under some number, the same cutoff for the marathon runner and the frail eighty year old in the next bed. The infant handles this differently and better. The infant habituates to whatever counts as normal in its particular environment and orients only when the input violates the model it has built of this specific situation. The product is a per patient predictive baseline that learns what normal looks like for this person across hours and days and fires only on genuine prediction error, the real deviation from the

patient's own trajectory, not a one size threshold inherited from a guideline. Deterioration detection, sepsis, the slow drift that precedes a crash, all of these are prediction error problems wearing a monitoring costume. A system engineered to be surprised correctly, and crucially to stay quiet when nothing surprising is happening, goes after alarm fatigue at the root instead of bolting one more beeping box onto a wall full of boxes nobody listens to anymore.

Overimitation, or why copying the doctor copies the rituals too

Time for the cautionary tale, and it happens to be a funny one. Put a child and a chimpanzee in front of a puzzle box and demonstrate how to get the treat out, but pad the demonstration with obviously pointless steps, tapping the lid three times with a stick before opening a door that was never locked. The chimp, no fool, skips the theater and goes straight for the treat. The child solemnly taps the lid three times first, then opens the door. This is overimitation, documented by Horner and Whiten and then sharpened by Lyons and colleagues, and human children do it more, not less, the more irrelevant the steps look, especially when a confident adult performs them with the air of someone who clearly knows what they are doing. It looks like a bug. It is mostly a feature, the mechanism that lets culture transmit faithfully even when the learner has no idea why a given step matters, which is wonderful when the step is wash your hands before the incision and a disaster when the step is tap the lid three times.

The healthcare lesson lands directly on the most fashionable corner of clinical AI, the ambient scribe and the broad family of just imitate what good doctors do systems. Train a model to imitate physician behavior off EHR data and it will faithfully reproduce the rituals right alongside the reasoning, because it cannot tell them apart any better than a four year old can. It will learn to copy forward a note that has been quietly wrong since 2019 because everyone copies it forward. It will learn defensive over testing because the training physicians over test to cover themselves. It will reproduce the billing driven documentation bloat, the templated normal exam that nobody actually performed, the reflexive consult that exists mostly to spread liability around. Imitation learning inherits the cargo cult wholesale. The lesson for a builder is not to throw out imitation, which is genuinely powerful and underpins a lot of what works. The lesson is to pair imitation with the causal machinery from earlier, the part that actually asks whether tapping the lid does anything at all. The valuable scribe is not the one that imitates most faithfully. It is the one that can quietly drop the three taps and keep the part that mattered.

What a builder actually does with all this

Step back and the pattern is clean. The child is not running one model. The child is running a small zoo of specialized, cheap routines. A statistics engine that pulls structure out of unlabeled streams. A causal experimenter that intervenes when observation alone will not settle the question. A curiosity policy that spends scarce attention exactly where it learns the most. A few shot generalizer that leans on strong priors instead of brute data. A prediction error detector that only bothers to care when reality breaks the forecast. An imitation system that copies first and reasons later, sometimes to its own detriment. These compose into something that looks general from the outside, but each routine is separable, well characterized, and small enough to build a real product around. The world model is the emergent property of the stack. The routines are the engineering, and the engineering is where the money is.

For anyone hunting for where to point a team, the move is to pick one routine and one healthcare modality where that routine is the actual bottleneck, then ignore the rest for now. Self supervised pretraining where labels are the constraint and raw signal is abundant. Causal estimation where the whole field is drowning in confounded observational data and mostly pretending it is fine. Active learning where expert attention is the rate limiter and random sampling is the status quo. Few shot priors where the disease is too rare to ever assemble a fat dataset. Personalized prediction error where the population threshold generates noise instead of signal and trains everyone to ignore the monitor. None of these requires solving cognition or raising a nine figure round to afford the compute. Each one copies a single specific trick a baby already runs, and applies it to a spot in medicine where that exact trick is conspicuously missing.

There is even a sixth worth a footnote, the motor babbling infants use to learn the model of their own bodies through more or less random exploratory flailing, which maps neatly onto rehabilitation robotics and prosthetics that have to learn a particular patient’s altered body rather than a generic textbook one, but that is a different essay for a different week. The honest summary is that world models is a fundraising phrase and the kid runs a stack of point solutions is a building phrase. The babies, who have collectively raised exactly zero seed rounds and cannot yet operate a doorknob, have been demonstrating the second approach the entire time. Worth copying their homework, one routine at a time

.

Subscribe to Thoughts on Healthcare Markets & Technology

By Thoughts On Healthcare Markets & Technology · Hundreds of paid subscribers

Expert analysis of healthcare and life sciences markets, technology, investment, entrepreneurship, policy, and AI — for investors, entrepreneurs, hospital and insurance executives, and physicians navigating the business of healthcare.

By subscribing, you agree Substack's Terms of Use, and acknowledge its Information Collection Notice and Privacy Policy.

1 Like

← Previous

Next →

Discussion about this post

Comments Restacks

Write a comment...

© 2026 Healthcare Markets & Technology · Privacy • Terms • Collection notice
Substack is the home for great culture