Ultrafinitism and epistemic limitations

András Kornai
SZTAKI Computer Science Research Institute
and
Institute of Mathematics, Budapest University of Technology

12 April 2025

Plan of the talk

- Some background notions, in particular the NPC
- Epistemic limitations: cpu/memory/bandwith limitations of computing the deductive closure
- Analyzed in the context of human performance limitations on System I tasks
- True to the explicit agenda, we try to get explicit bounds. These will be far below the $10^{10^{10}}$ bound of Parikh, 1971
- We will use networks of finite automata (Clustered Moore Automata, CMA) to obtain our bounds. CMA use timescales
- We populate the model structures both with NPCs and PCs
- This will put tangible lower bounds (10^{20}) on the system as we need considerable resources for building these characters
- Relation to AI safety agenda (Kornai, 2014; Kornai, Bukatin, and Zombori, 2023) Bard kill Smaug

Setting expectations

- We will stay in a domain where ordinary scientific notation is sufficient (knowing full well that in a Platonic sense Graham's number and even Loader's number exist)
- We remain set-theoretically agnostic, but note that New Foundations with $< 10^3$ urelements are quite compatible with all we do (as would be ZF, NBG, etc – everything will be hereditarily finite and small)
- We will use one big number, a *myriad*, but even this is $m \leq 10^4$ (Vardi, 1997)
- We stay well within the AC^0 complexity class, with explicit bounds on circuit depth ≤ 15 and on polynomial (really, linear) number of gates
- Asymptotic arguments (including transfinite ones) are irrelevant in the land of the finite. You can't prove, or even argue, that 3.14159... is transcendental!
- We'll use explicit numbers instead of asymptotic considerations

Background

- Main thesis: given (ordinary) epistemic limitations, ultrafinitism is necessary
- Goal: right-sizing the mathematical universe
- Will give both lower and upper bounds, with the upper bound surprisingly low, 10^{40} or even less
- Data from System 1 thinking, natural language, ‘naive theory’ rather than System 2 (arithmetic, math in general). NL requires significant amounts of deep philology (familiar to philosophers, but less so for mathematicians)
- Main tools: finite automata/transducers (Kornai, 2025)
- These are used to model slow Turing Machines (TMs with epistemic limitations)
- Main ideas can be traced back to the *Sand Reckoner* of Archimedes (Vardi, 1997)

Computing agents with limited memory

- We are distinguishing the tools available to the agent by the degree of effort it takes to use them. Assume a working memory that can fit 7 ± 2 tools, and a long-term memory with a large number of tools.
- We say a tool is *directly accessible* if it is present in LTM in a non-degraded form; *index accessible* if a pointer to it is available in LTM; and *far* if it takes significant effort to (re)learn it
- Examples: after undergrad, complex numbers are *directly accessible*, but octonions are only index accessible. The Wiles proof is like the hundred quadrillionth hex digit of π – it takes a very significant effort to get to it. You may trust the result, but 99% of working mathematicians don't know the major steps of the proof, and would be hard put to explain the great new ideas it contains

Captatio benevolentiae

- Perhaps I'm not the only person in the room who firmly believes in the existence of π, the ratio of the circumference of the circle to its diameter
- Furthermore, I see nothing wrong with the idea that π is well encoded in a particular infinite sequence 3.14159...
- By any sensible definition, this sequence is constructible by *effective* procedure/algorithm (now done for $2.02 \cdot 10^{14}$ digits, requiring 1TB RAM and 1.5PB storage)
- We can compute the 10^{17}th (hex) digit (took two weeks of CPU time on a 512 node cluster in 2018) but can't easily store 10^{17} digits (would have cost $65m at the time) Bard kill Smaug
- We will generalize these observations from numerical to any kind of System I deductive setup (Kahneman, 2011)
- Belief in π amounts to some kind of allegiance to a higher/Platonic world we'd like to live in. It does not amount to the belief that the actual world we live in is such.

What is at stake?

- *Is there a fact of the matter about whether* $D = 10^{10^{10+1}+1} + 1$ *is prime, even if no human or physical computer could ever find out?* (Justin Clarke-Doane)
- Best primality test (Agrawal–Kayal–Saxena) runs in n^6 steps, so this is 10^{66} steps. One step requires at least one Planck time unit, so this is $1.7 \cdot 10^6$ billion years, factor of 10k larger than the expected lifetime of the universe.
- So the brute force answer is that there may not be a fact of the matter about the primality of D
- Here I take the view that there is! Obviously, $10^r + 1$ is divisible by 11 for r odd, and $10^{11} + 1$ is odd. Therefore, $11|D$
- We have some kind of allegiance to a world where this simple proof remains valid
- While the other sciences search for the rules that God has chosen for this Universe, we mathematicians search for the rules that even God has to obey (Jean-Pierre Serre)

NPCs

- The player/nonplayer distinction affects the estimates
- A *player* will admit they have the capability to set their own goals
- We (humans) are players
- Can LLMs be turned into players? Yes, and it wouldn't be hard
- Can they really set their own goals? Can we?
- Compatibilist view, Conway-Kochen
- Math modeling of free will is a trivial issue: use nondeterminism (Rabin and Scott, 1959; Floyd, 1967)
- Detecting players is easy: they have to declare themselves (performative act)
- By doing so, they submit themselves to the PGC (Gewirth, 1978)

AGIs as Gewirthian PPAs

- The claim (Gewirth/Beyleveld): the Principle of Generic Consistency (PGC) is ‘dialectically necessary’ (binding on any agent) as long as they admit that they are potentially purposive agents (capable of voluntarily setting goals for themselves) and are capable of reasoning
- A *player* will admit they have the capability to set their own goals and that they can reason. All it takes is a speech act and you are bound by the PGC
- The PGC ‘Act in accord with the generic rights of your recipients [to freedom and well-being] as well as of yourself’ is the categorical imperative
- There is a machine-verified proof (Fuenmayor and Benzmüller, 2019) Enter the dragon
- Current LLMs are generally trained away from admitting they can set their own goals, and their resoning abilities are System I.

PCs must be accorded human rights

Timescales

- Discrete timescales anchored in R_0 ‘heartbeat’ or ‘second’ scale
- We assume R_{-1} ‘centisecond’ or ‘instant’ scale. The ball changes direction in an instant: https://bit.ly/43AhWGH
- There is a ‘quarter hour’ scale R_1, a ‘day’ scale R_2, a ‘season’ scale R_3, a ‘generation’ scale R_4, with an ‘aeon’ scale R_5 on top
- The entire system is finitistic, with no more than a *myriad* ($< 10^4$) steps accessible on any single scale, and with a well-defined fastest $R_{\mathtt{min}}$ and slowest $R_{\mathtt{max}}$ scale
- For modern physics you’d set zeptosecond to zettasecond scales
- (i.j) means ‘unit j on scale i’

Slow Turing Machines

We will use *nondeterministic, alternating, slow* TMs. To fully describe some nasTM A operating between scales R_i (the 'small' scale) and R_{i+1} (the 'large' scale) requires specifying

(i) the labels $\mathcal{L}$ it can read from external memory in one large time unit

(ii) the state space w that covers each state the control may be in

(iii) the labels $\mathcal{O}$ it can write in one large time unit

(iv) the transitions $w(i.k) \rightarrow w(i.k+1)$ it can take during a single small time unit, and

(v) the set of fluents p which includes, but is not limited to, the standard $\vee, \wedge$, *rest, accept, reject* that govern the alternating behavior (Chandra, Kozen, and Stockmeyer, 1981).

The hard thing is to build the tape! Position of reading head requires only $\log_2$(tapelength) bits

Towards a reasonable question

- Weiss et al set the limits for LSTM/GRU, but the same can be done for Transformers (attention), which would require infinite precision to deal with $a^n b^n$
- In fact, Transformers get quantized down to 3-4 bit precision without much loss of core System I (Kahneman, 2011) functionality such as machine translation and textual inference
- Work hitherto concentrated on System II capabilities, with long chains of deduction
- The limits of this kind of precise deduction can be investigated through investigation of bignum arithmetic, but is alien both to humans (low numerosity, see (Dehaene, 1997) and LLMs (which show the exact same limitations)
- My interest is with humans and LLMs, both known to be finite automata. The target is System I deduction and the full semantic framework surrounding it

Towards a reasonable architecture

- As is well known (Cybenko, 1989) neural nets are universal approximators
- There is a line of research implementing Turing Machines in NNs (Siegelmann and Sontag, 1992; Siegelmann, 1996) showing that recurrent NNs a la Elman, with sigmoid activation function, rational weights, and infinite precision can simulate a TM (in real time)
- This is better than the Shannon, 1941 GPAC model (can do e.g. Γ function) but really we are sub-Turing so the hypercomputing advantage clamed by Siegelmann, 1999 is illusory
- For careful analysis of where the tire meets the road, see Weiss, Goldberg, and Yahav, 2018 (the issue is with arbitrary precision required for embedding loops of arbitrary depth)
- We will deal with finite automata only, but with realistic memory size, and the fact that adding n bits of memory blows up the state space by a factor of 2^n

What is the logic of System I deduction like?

- Short chains, few $< 7 \pm 2$ elementary steps
- Few variables (not a true limitation given Tarski and Givant, 1987) type theory can be data-mined through them
- A good number of constants: $\approx 10^4$, maybe 10^5, but humans are unlikely to have 10^6
- Few rules of deduction (not a true limitation given modus ponens/sequent calculi)
- Few-argument relations built around a static core
- Connectives: only conjunction
- For model theory/grounding will look at mappings to/from ideas (things in the head). Mappings to/from objective(?) reality are mediated by mind-states

Bird's eye view

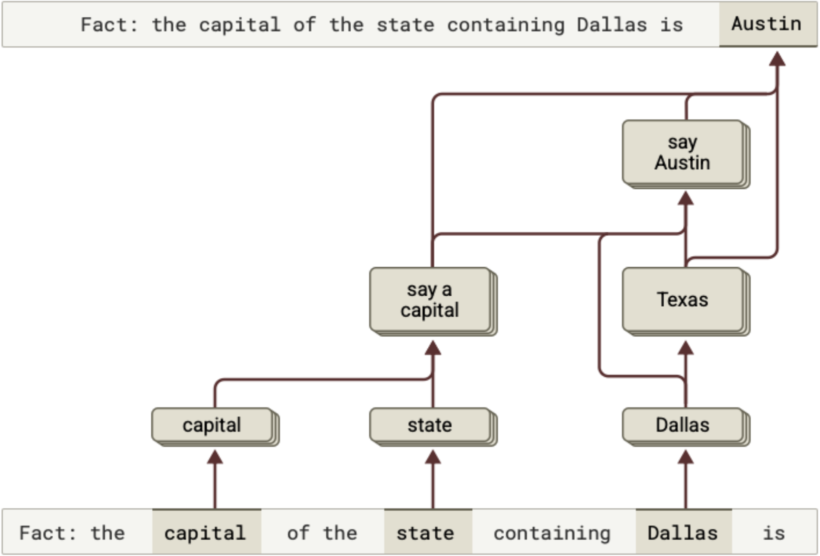
- Classic: only two truth values, everything has one and only one
- Heavy on defaults
- Pattern matching treated as analysis-worthy
- Long story about negation and disjunction (Kornai, 2024)
- Only generic quantification
- Modality is within-model
- System called `4lang`

A typical System I deduction

- *What is the capital of the state containing Dallas?*
- Dallas IN Texas, Texas ISA state, state HAS capital, Austin ISA capital, Austin IN Texas
- 4lang has no compund relations like 'HAS_CAPITAL' and it is quite hard to formulate uniqueness statements like 'there is only one capital per state'
- We do everything by *spreading activation* (Quillian, 1967; Collins and Loftus, 1975; Carroll, 1983)
- Initially, only Dallas, capital, state, containing are active. The triples listed above are all in the lexicon (long-term memory) together with Sacramento ISA capital, Sacramento IN California, California ISA state etc.
- Through these, activation spreads to Texas, and from there to Austin QED.

LLMs

Anthropic (2025)



Constants

- Roughly correspond to words and meaningful word parts (morphemes)
- The number of constants is traditionally considerd a good measure of intelligence and ability to govern (traditional Chinese imperial examination, but also in GB and elsewhere since the Northcote-Trevelyan report)
- They are word vectors, computed from local cooccurrence statistics in text
- Verbs, generally considered functions or relations in Montague-style semantics, are also vectors
- There are a few matrices such as prepositions, case endings, etc. and a few operations more tricky than vector addition or matrix application

Variables

- The closest thing in natural language is pronouns
- There are no variable binding term operators
- There is only one quantifier, `gen`, corresponds to the vector $[1/d, 1/d, \ldots, 1/d]$ in d-dim space not an operator
- It is best to think of variables as unknowns/indeterminates/partially determined constants than as actual variables with a domain
- Monadic second order quantification (with hypernode graphs)
- Because of Büchi-Elgot-Trakhtenbroth this means only regular languages are within scope, but we know this anyway, since the brain is a (large) finite automaton
- Exciting work on subregular hierarchy Graf, 2022

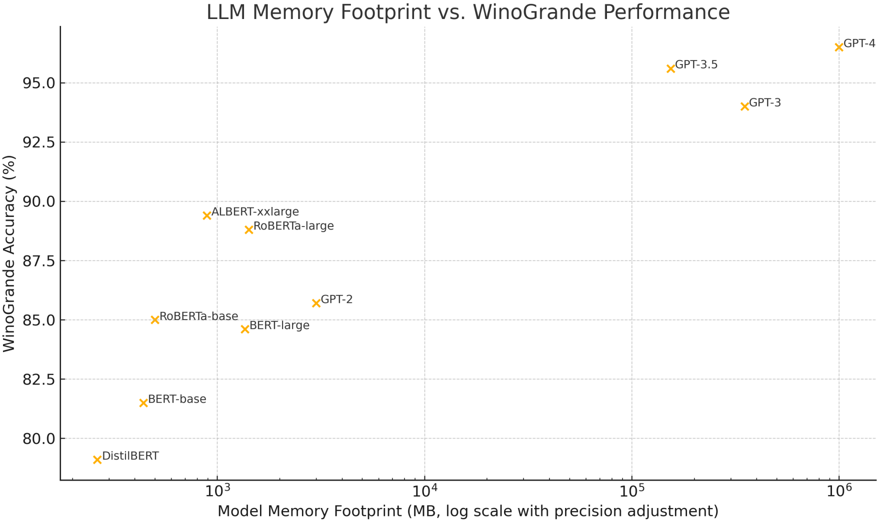
Basic types

- Two types: automata and vector
- Both are standard (non-deterministic, subsequential Moore automata; modules over rings and fields)
- They are just two ways to speak about the same thing, like the geometric and the algebraic view of linear spaces. We consider these equally valid, and make no attempt to reduce either one to the other (though this should be doable in both directions)
- Both are size-limited: elementary autmata have at most 10^4 states, all vectors shorter than 10^{12} bits (actually, much shorter)

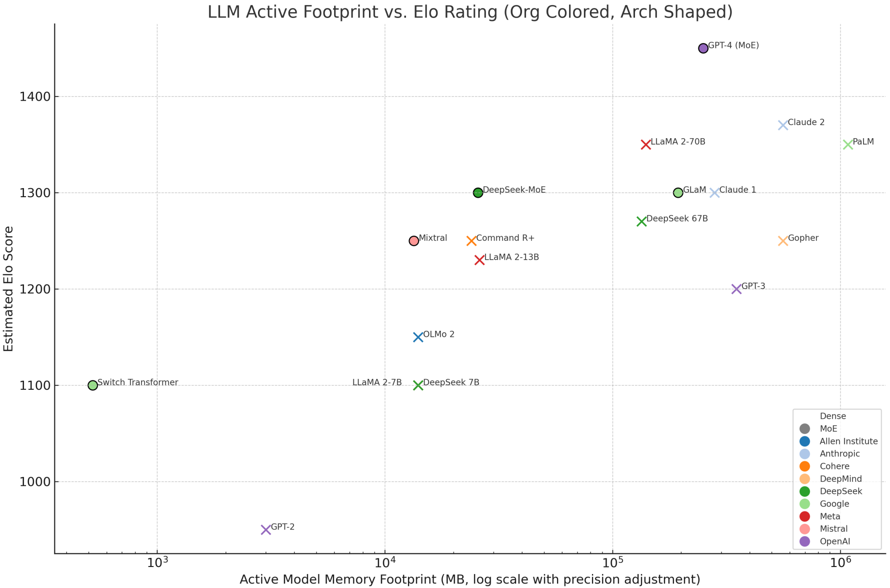
Rules of deduction

- Of course one (modus ponens) would be enough, but we aim at realism, not minimalism
- The empirical domain we want to model is known in the trade as ‘textual entailment’
- There are standard problem sets ‘shared tasks’ that we want our systems to do well on, perhaps the best known is the Winograd Schema and WinoGRANDE set of challenges
- The large ball crashed right through the table because it was made of steel/styrofoam
- Ten years ago even the best NLP systems did badly on these (50-55%, hard to distinguish from random)
- Contemporary LLMs do as well as humans or better

Size estimates



Overall performance



Lower bounds: the basic setup

Any AGI expecting to reach a high level of fitness will find it prudent to expend some effort toward tamper-proofing its environment, its perceptual and motor systems, and its internal logic. Once these efforts are deemed successful (and they can never be completely successful in the material universe in that arbitrarily large gamma-ray bursts can always reset some part of memory) we can equate an AGI with its deductive system.

- We have an **A**GI Alice and an **E**xperimenter Eve. E can
- eavesdrop on A's thoughts; present A with false perceptual data; rewrite A's personal memory; experiment (serially or in parallel) with a large number of Alice copies; change physics locally (perform miracles) or globally (change the entire physics)
- Should Alice drop down on her knees and pray to Eve?
- Suppose A has some goal G she wants to come about, and E opposes G. (Prometheus, Adam)

What can Alice do?

- What can anyone in inferior position do? (A) Bring some helpers
- (B) Probe the limits of what can be done against Eve's wishes – this is just doing physics research
- (C) Convince Eve that G is right
- Note that by having goals and rationality, Alice is a PC
- We must assume Eve to be a player too
- Mathematicians can't pledge allegiance to irrational gods – logic is part of the everlasting covenant (brit olam, Gen 17:7) binding God as well (see also Deut 7:9; Ex 2:24; Lev:26:44)
- Also part of the commentary literature (Rashi on Ex 2:24, Ex 6:5; Maimonides on Ex 2:25) and the received Christian view (St. Anselm of Canterbury, St. Thomas Aquinas)

The initial position

- You don't know whether you are Alice or Eve. In fact you can't know. In a Rawlsian initial position it doesn't even matter
- For a lower bound, consider GPT-4 which measurably outperforms average humans on commonsense reasoning tasks. GPT-4 takes only a TB
- For a full BIV you may want more: access to personal memories (a day of full immersive experience should be compressibe to 300MB, 90 years to 10TB)
- We also want shared cultural heritage, which will be a few PB
- We should be able to build Exabyte City to hold about 500 PCs
- Given current technology it takes about $100m to host a disk array with well over a Dunbar's number (~ 150) of PCs and as many NPCs as you wish
- Altogether, humans cannot distinguish themselves from BIVs hosted in a computational substratum hosted on an EB ($8 \cdot 10^{18}$ bits)

Upper bound

- 10^{120} bits comes from Lloyd, 2002
- I am not capable of critically assessing the validity of his argumentation
- A more pedestrian argument is based on elementary operations taking Planck time, and estimated lifetime of the universe being half of total lifetime (Copernican expected value), this yields about 10^{63} ops. Number of ops can't be very different from number of bits (Cray's Rule: on word for one flop)
- Current systems are in the GHz range, exaflop scale comes from parallelism
- H100 card has 2 petaflops but only 80GB of memory
- Highly speculative reasoning for 10^{40} based on sqrt rule of thumb: n datapoints should be binned in no more than $\sqrt{n}$ bins

Thank you for your attention

Beyleveld, D. (1992). *The dialectical necessity of morality: An analysis and defense of Alan Gewirth's argument to the principle of generic consistency*. University of Chicago Press.

Carroll, John A. (1983). *An island parsing interpreter for the full augmented transition network formalism*. ACL Proceedings, First European Conference, pp. 101–105.

Chandra, Ashok K., Dexter C. Kozen, and Larry J. Stockmeyer (1981). "Alternation". In: *Journal of the Association for Computing Machinery*. Vol. 28. 1, pp. 114–133.

Collins, A.M. and E.F. Loftus (1975). "A spreading-activation theory of semantic processing". In: *Psychological Review* 82, pp. 407–428. DOI: 10.1037/0033-295X.82.6.407.

Cybenko, George (1989). "Approximation by superpositions of a sigmoidal function". In: *Mathematics of control, signals and systems* 2.4, pp. 303–314.

Dehaene, Stanislas (1997). *The number sense*. Oxford University Press.

- Floyd, Robert W (1967). “Nondeterministic algorithms”. In: *Journal of the ACM (JACM)* 14.4, pp. 636–644.
- Fuenmayor, David and Christoph Benzmüller (2019). “Harnessing Higher-Order (Meta)Logic to Represent and Reason with Complex Ethical Theories”. In: *PRICAI 2019: Trends in Artificial Intelligence*. Ed. by Abhaya C. Nayak and Alok Sharma. Springer International Publishing, pp. 418–432. ISBN: 978-3-030-29908-8.
- Gewirth, Alan (1978). *Reason and morality*. University of Chicago Press.
- Graf, Thomas (2022). “Subregular linguistics: bridging theoretical linguistics and formal grammar”. In: *Theoretical Linguistics* 48.3–4, pp. 145–184. DOI: 10.1515/tl-2022-2037.
- Kahneman, Daniel (2011). *Thinking, fast and slow*. Farrar, Straus, and Giroux.
- Kornai, András (2014). “Bounding the impact of AGI”. In: *Journal of Experimental and Theoretical Artificial Intelligence* 26.3, pp. 417–438.

- Kornai, András (2024). "Dyadic negation in natural language". In: *Acta Linguistica Academica* 71, pp. 235–257. DOI: 10.1556/2062.2024.00656. URL: https://akjournals.com/view/journals/2062/71/1-2/article-p235.xml.
- — (2025). *Cluster automata*. arXiv: 2503.22000 [cs.CL]. URL: https://arxiv.org/abs/2503.22000.
- Kornai, András, Michael Bukatin, and Zsolt Zombori (2023). "Safety without alignment". In: *ArXiv* 2303.00752.
- Lloyd, S. (2002). "Computational capacity of the universe". In: *Physical Review Letters* 88.23, p. 237901.
- Parikh, Rohit (1971). "Existence and feasibility in arithmetic". In: *Journal of Symbolic Logic* 36.3, pp. 494–508.
- Quillian, M. Ross (1967). "Semantic memory". In: *Semantic information processing*. Ed. by Minsky. Cambridge: MIT Press, pp. 227–270.

- Rabin, M.O. and D. Scott (1959). "Finite automata and their decision problems". In: *IBM journal of research and development* 3.2, pp. 114–125. ISSN: 0018-8646.
- Shannon, Claude E. (1941). "Mathematical Theory of the differential analyzer". In: *Journal of Mathematics and Physics of the MIT* 20, pp. 337–354.
- Siegelmann, Hava T. (1996). "Recurrent neural networks and finite automata". In: *Computational Intelligence* 12, pp. 567–574.
- — (1999). *Neural Networks and Analog Computation: Beyond the Turing Limit*. Birkhäuser.
- Siegelmann, Hava T. and Eduardo D. Sontag (1992). "On the computational power of neural nets". In: *Proc 5th ACM Conference on Computational Learning Theory, COLT 1992*, pp. 440–449.
- Tarski, A. and S.R. Givant (1987). *A formalization of set theory without variables*. American Mathematical Society.
- Vardi, Ilan (1997). "Archimedes the sand reckoner". In: *preprint.*

Weiss, Gail, Yoav Goldberg, and Eran Yahav (2018). "On the Practical Computational Power of Finite Precision RNNs for Language Recognition". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 740–745. DOI: 10.18653/v1/P18-2117. URL: https://aclanthology.org/P18-2117.