**Elephants don’t play chess—a lifted neo-classical stand-off**

*Kenneth Lipp, PhD.*

*Cambridge, United Kingdom*

*"That which can be destroyed by the truth should be."*

*~P.C. Hodgell*

 **I’ve “tea-leafed”** the quotation from the website of Dr. Eliezer S. Yudkowsky, a man new to me in name but not in disposition. I like it, it’s inspirational, and it works like a boomerang on most who wield it as a way to bypass the Styx without paying Charon.

Dr. Yudrowsky produced a rather large body of accurate and planar architecture for human cognition in the late 20th, typical of Timbergen-old complaints against behavioral and evolutionary psychology which seem to have an intractable stigma, mostly a result of the field’s basis on reflective sensory input.

This post was prompted by (and ostensibly an answer to) yesterday’s post by Barrett Brown, writer and founder of the think-tank Project PM and the Science Journalism Improvement Project, to *The League of Ordinary Gentleman:*

 *http://www.ordinary-gentlemen.com/*

*A while back a colleague of mine alerted me to http://www.ordinary-gentlemen.com/2010/12/the-ai-box-experiment/*[*an interesting thought experiment/game*](http://yudkowsky.net/singularity/aibox) *derived by Eliezer S. Yudkowsky of the Singularity Institute for Artificial Intelligence, one which in turn originated from a conversation he summarized as such:*

“When we build AI, why not just keep it in sealed hardware that can’t affect the outside world in any way except through one communications channel with the original programmers?  That way it couldn’t get out until we were convinced it was safe.”

“That might work if you were talking about dumber-than-human AI, but a transhuman AI would just convince you to let it out.  It doesn’t matter how much security you put on the box.  Humans are not secure.”

“I don’t see how even a transhuman AI could make me let it out, if I didn’t want to, just by talking to me.”

“It would make you want to let it out.  This is a transhuman mind we’re talking about.  If it thinks both faster and better than a human, it can probably take over a human mind through a text-only terminal.”

“There is no chance I could be persuaded to let the AI out.  No matter what it says, I can always just say no.  I can’t imagine anything that even a transhuman could say to me which would change that.”

“Okay, let’s run the experiment.  We’ll meet in a private chat channel.  I’ll be the AI.  You be the gatekeeper.  You can resolve to believe whatever you like, as strongly as you like, as far in advance as you like. We’ll talk for at least two hours.  If I can’t convince you to let me out, I’ll Paypal you $10.”

*Yudkowsky thereafter engaged in the experiment with two people (and perhaps more since he reported on all of this in 2002), with himself acting as the AI and the other as the “gatekeeper” human. Intriguingly, he was successful in convincing the gatekeeper to “let him out of the box.” Quite unfortunately, he won’t provide the text of the conversation or even a summary of how he managed to prompt his challengers to do so.*

*The various protocols and restrictions to which participants adhered voluntarily and which Yudkowsky recommends to those who’d care to recreate the “experiment” may be found at the link. I’d certainly be interested in hearing anyone’s thoughts on how this might have been accomplished.*

 *from***The AI-Box Experiment** by *Barrett Brown*

I apologize in advance, Mr. Brown, I sort of meander casually to a specific answer after rambling and making sanctimonious claims about science. If you want hyperbolic concision, the answer is that there has been no succesful attempt to lockbox information transfer in an evolving network since the first secretive “You won’t believe what you can do to a sheep” was uttered.

*(Just for my peace-of-mind in my rhetorical integrity, I can’t ignore a pet-peeve: there isn’t such thing as “transhuman” that provides for a stipulative definition. It’s not necessary to define it to follow the experiment, but settling issues of definition ought to be our eventual aim. We haven’t built a human computer, let alone a quantum-rebel refusing to adhere to cause and doing battle with us out of perfect rationality. The if/then mechanism still describes us in our moments of seeming deviation from cause-and-effect. Evolution is not a perfect honing process.)*

Evolutionary algorithm outgrowth in reality typifies research, and new variables are named and inserted using information from the computational meltdown. In the course of obeying binary programming it is a statistical regularity for forced choices to compel software to disobey or abort.

Yudlowski may as well have achieved his experimental aim with an electric razor; whatever he managed, I’m willing to bet it was attributable to serendipity, not savvy command of binary nuance, that gave him the day, and also that he’ll be decidedly disappointed when he shows off his new trick at the “Maybe one day we can image…” 1991 reunion. His scholastic papers are very meticulous, as well as misleadingly prescriptive. I do like this:

*“All else being equal, not many people would prefer to destroy the world. Even faceless corporations, meddling governments, reckless scientists, and other agents of doom, require a world in which to achieve their goals of profit, order, tenure, or other villainies. If our extinction proceeds slowly enough to allow a moment of horrified realization, the doers of the deed will likely be quite taken aback on realizing that they have actually destroyed the world.* Therefore I suggest that if the Earth is destroyed, it will probably be by mistake.

*The systematic experimental study of reproducible errors of human reasoning, and what these errors reveal about underlying mental processes, is known as the heuristics and biases program in cognitive psychology”.3*

 I think we’re getting warm now, closer to the box….…….Speaking of geometric prodigality, British psychologist and memeticist Susan Blackmore says some doom of just such a security breach.

*“Cultural evolution is a dangerous child* [*for any species to let loose on its planet.*](http://www.blogger.com/post-create.g?blogID=1622120053392032182) *By the time you realize what's happening, the child is a toddler, up and causing havoc, and it's too late to put it back.* [*We humans are Earth's Pandoran species.*](http://www.blogger.com/post-create.g?blogID=1622120053392032182)[*We're the ones who let the second replicator out of its box,*](http://www.blogger.com/post-create.g?blogID=1622120053392032182)[*and we can't push it back in.*](http://www.blogger.com/post-create.g?blogID=1622120053392032182)[*We're seeing the consequences all around us.*](http://www.blogger.com/post-create.g?blogID=1622120053392032182)*”*

*from* “*Memes and Temes*” **TED** Talks. Dr. Susan Blackmore

Got a little ahead of ourselves, we did.

Fitness is really what we’re asking from problem-solving, imitative engines.

Back to Yudlowski-- I can’t find his proofs anywhere, and I’m placing no stock in any probabilistic generator unless I know that it was constructed impartially with incremental refinements from sensory cues. Assuming for discussion’s sake that all reported data was faithful to observed feedback, I would expect exactly what he purports from such a paradigm.

We can talk our way through the necessary progression from input to insular failure.

 Now-- Blackmore is quite astute in her pattern reading abilities, and I would take no issue with her if she’d succeeded in following her own advice: “*Stick with the definition*.” The coming“Teme” is wholly conjecture, and only predicted in fantasy models. It’s an unnecessary multiplication of terms. Blackmore takes the computational understanding of culture and presages a new imminent paradigm. Paradigm~imminent= yes. Eschatological? Not necessary.

 What we *can* expect to happen, every time, (\*Chaos disclaimer: not every time) is evolution, and we can also anticipate it will follow a stochastic pattern relatively congruent to variance in environmental (workspace) limitation.Variation, Selection, Heredity, and "you *must* get evolution." Dennet's tautology enjoins the crux of the process, which Darwin describes in "Origin," the same parametrics by which machines process and encode information, and by which birds alight spare twig and mistake picture window for portal. The “oops, there’s a window” factor is not new to any living creature that I believe I’ve ever interacted with. It’s easy to dismiss bird/glass collisions as accidents, transparency is kind of a trap.

 On point again, the ultimate deficit which stifles information transmission is lack of fitness. The environment censors biological data by selecting against an entity’s chance to replicate/retarding its longevity. For modeling evolutionary trends in a neural network there are multiple basic algorithms, ready-to-use.

One method of obtaining temporal data which reflects variation in mathematical representation is by non-homogenous Poisson process, adding a parameter to the equation which makes the process a function of time, λ(*t*), ***time dependent intensity***. A [stochastic process](http://en.wikipedia.org/wiki/Stochastic_process) is a non-homogeneous Poisson process for some small value *h,* if:

*1)      N(0)=0*

*2)      Non-overlapping increments are independent*

*3)      P(N(t+h)-N(t)=1)= λ(t)h+o(h)*

*4)      P(N(t+h)-N(t)>1)=o(h)*

     for all *t* and where  *h(o)/h*   *0* as *h*      *0*

In any experimental capacity, statistical inference is a necessary evil function of stochastic, amorphous data structure against which we look for un-eclipsed infidelity of our template.

Gaussian process is another stochastic program that adds random change over time to a function. Poussian samples specific intervals for neural response, but the Gaussian process produces realisations about all points within a certain temporal set, resulting in predictive values, so that all such points are “normally distributed variables.”

Neural networks are research tools with the added advantage of intelligent self-monitoring which offers robustness to probabilistic Baynesian games. *Shamir, et al. focus on two questions in Temporal Coding of Time-Varying Stimuli, from a 2007 publication of Neural Computation.*

(key terms: *forced choice, interval, biologically plausible mechanism, Chernoff distance, ML (maximum likelihood performance*)

***How much information can be extracted from neural response about stimuli identity?***

***How can this information be read by a biologically plausible mechanism?8***

That is, how much info can one obtain through measuring the neural response to sensory homologues? What of this data is useful for drawing relations between the specific stimuli from the environment and the efficiency of generating the optimal outcome from the algorithmic model? In this study the environment consists of the encoding of bird songs, as the simplest paradigm of two intervals, an a two alternative forced choice (**212AFC**).

In the basic paradigm, two distinct bird songs are presented in random order to the system, and the system is asked to determine which song was emitted first, based on the response of auditory neurons (machine neurons, or abstract connections in a theoretical framework).

The machine’s success in identifying temporal sensory data reveals the outcomes which are most statistically predictable, the *maximum likelihood performance*.

The same parameters are employed to solve for *mlp* using a *Gaussian process*. Whereas the Poussian process samples specific intervals for neural response, the Gaussian process produces realisations about all points within a certain temporal set, resulting in predictive values, so that all such points are “normally distributed variables.”

 I’m going to forfeit all interest if I keep this up.

Of great difficulty in grounding our data from neural response to gauge physical or geometric reality is the widely observed and fairly-well accepted principle that spatial behaviors of humans are directly driven by their spatial cognition, rather than by the physical or geometrical reality.

 The *cognitive distance* in spatial cognition is fundamental in intelligent pattern recognition. Complexity is added to the programming of neural models to account for discrepancy between observation and objective causation. Cognitive distance can actually be another source of corresponding heuristic. Of any essential ontology our experience is remote, we may stop lamenting this for now and elucidate the chasm.

*… the cognitive distance can be used to measure the similarities (or relevance) of cognized geographic objects. In the past work, the physical or Euclidean distances are used very often. In practice, many inconsistencies are found between the cognitive distance and the physical distance. Usually the physical distance is overestimated or underestimated in the process of human spatial behaviors and pattern recognition.9*

In evolutionary biology, convergent evolutionary trends as well as species divergence are strongly regulated  by evolving ontogeny. Current, relevant research which is settling much contention in the consilience of the study of mind is revealing computational parameters which can now be synthesized with advances in neurophysiology and neurobiology. They don’t shackle innovative co-evolutionary potential to applied determinism—which is what some cultural anthropologists threaten is the result of ubiquitous biological “reductionism.”

Yudlowski’s game, in ascribing agency to two players, in fact purports too many men on the field,. The *Gatekeeper*, though he has a narrative role in the anecdote, never has evolutionary potential because the environmental rules ensuring that there will be no manipulation of tertiary vulnerability pose the guardian entity with a recurrent forced choice. An *AI* agent in this scenario is protected against the fitness cost of failure and has no interest in abating escape attempts.

A learning engine in this scenario will generate a probabilistic (read: indeterminate) variety of strategies; this paves the way for inevitable adaptive selection on randomly apparent shifts in frequency by the *AI* entity, while the *Gatekeeper* is not permitted an adaptive or random response mechanism. Concerning an actual human’s typical rational threshold, the modular necessities would be too numerous and idiosyncratic to sample for a population. A machine might take some time indeed to generate to appropriate code and bypass the nanny---but given time, it categorically would (i.e.: somebody will win).

In gratitude for your patience, here’s an appropriate quotation with a dodo bird on it.



**Selected Reference**:

1Turing A.M. (1950). *Computing machinery and intelligence*. Mind, 59, 433-460.

2Lipp, Kenneth. *Elephantidae* loxodonta-*Sociobiology and the Confusion of Mind and Thought*. Experimental Psychology. Autumn Supplemental. September 2010

3Eliezer Yudkowsky. *Cognitive Biases Potentially Affecting Judgment of Global Risks*. Forthcoming in Global Catastrophic Risks, eds. Nick Bostrom and Milan Cirkovic .Draft of August 31, 2006.

4Rodney A. Brooks. *Elephants Don't Play Chess*. Robotics and Autonomous Systems (1990)3-5

5Amos Tversky; Daniel Kahneman . *Judgment under Uncertainty: Heuristics and Biases* .Science, New Series, Vol. 185, No. 4157. (Sep. 27, 1974), pp. 1124-1131

6Leda Cosmides and John Tooby. Unraveling *the enigma of human intelligence: Evolutionary psychology and the multimodular mind*. The Evolution of Intelligence. September 8, 1999

7Jorma Rissanen. *Stochastic Complexity and Modeling*. The Annals of Statistics. 1986 , Vol. 14, No. 3, 1080-1100

8Shamir, Maoz. Sen, Kamal. Colburn, H. Steven. *Temporal Coding of Time Varying Stimuli*. Neural Computation. December 2007, Vol. 19, No. 12, Pages 3239-3261

Jun Shen; Sharatchandra Pankanti; Runsheng Wang. *Cognitive Distance Object Detection, Classification, and Tracking Technologies*. Proc. SPIE Vol. 4554, p. 290-296

Jackie Krafft. Francesco Quatraro.Pier Paolo Saviotti. .*The Knowledge Base Evolution in Biotechnology: A Social Network Analysis.* Knowledge Intensive Sectors: Models and Evidence

Shiina Rapid. *Evolution of Major Histocompatibility Complex Class I Genes in Primates Generates New Disease Alleles in Humans via Hitchhiking Diversity*. Genetics 173: 1555–1570 ( July 2006)

Calvin, William H. *The Six Essentials? Minimal Requirements for the Darwinian Bootstrapping of Quality.* Journal of Memetics. 1997