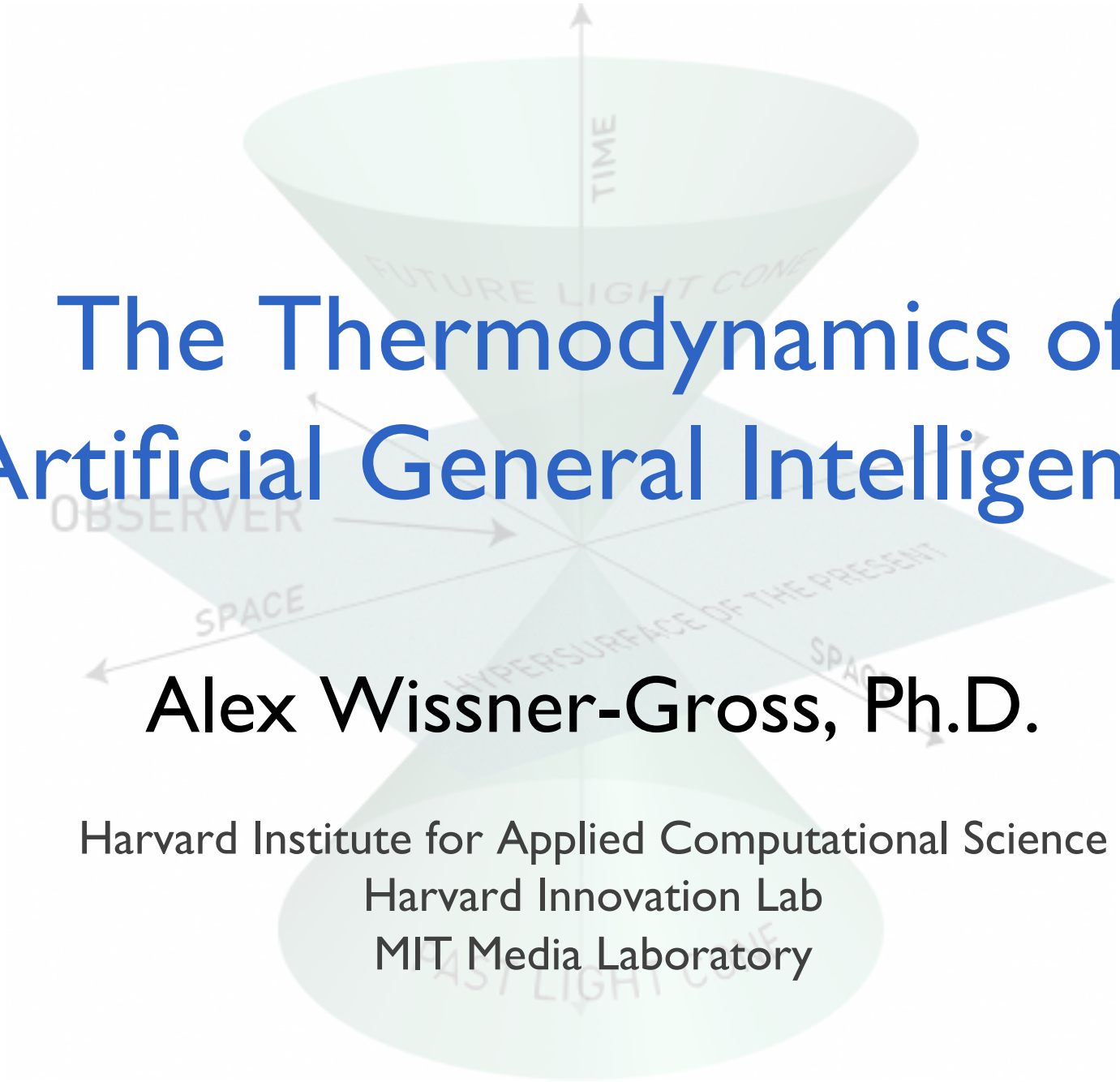


The Thermodynamics of Artificial General Intelligence

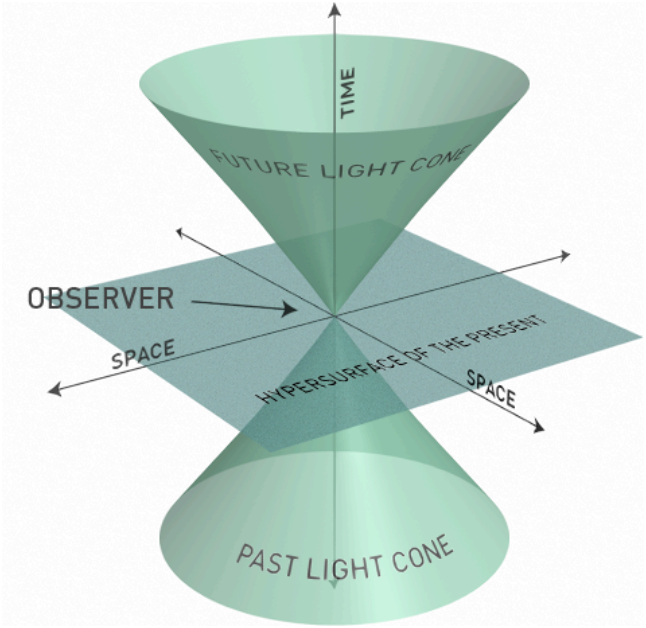
Alex Wissner-Gross, Ph.D.

Harvard Institute for Applied Computational Science
Harvard Innovation Lab
MIT Media Laboratory

web: **www.alexwg.org** twitter: **@alexwg** email: **alexwg@post.harvard.edu**

Overview

As the speed of computer systems and their integration with the physical world have grown, the **physical limits of intelligent systems** have become increasingly relevant.



Two key **physical limits** for realizing optimal coordination & control in software systems:

Time-Like Limit

Maximum causal entropy

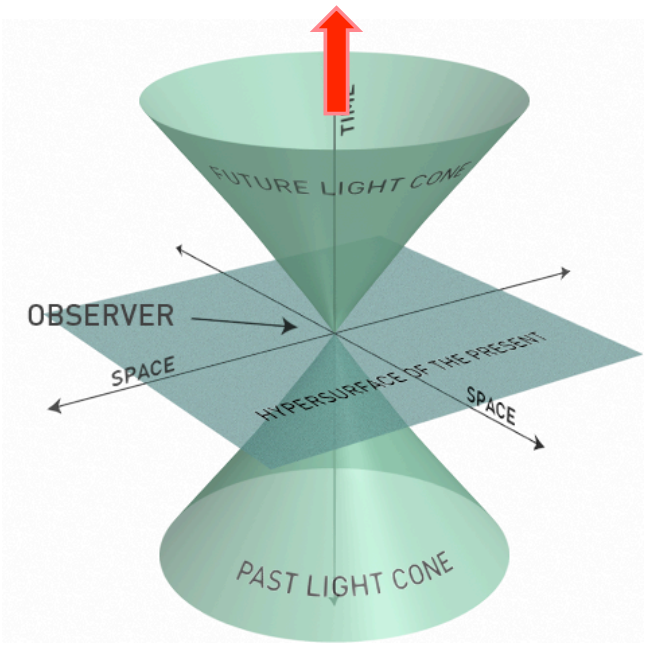
Space-Like Limit

Minimum coordination latency

web: **www.alexwg.org** twitter: **@alexwg** email: **alexwg@post.harvard.edu**

Overview

As the speed of computer systems and their integration with the physical world have grown, the **physical limits of intelligent systems** have become increasingly relevant.



Two key **physical limits** for realizing optimal coordination & control in software systems:

Time-Like Limit

Maximum causal entropy

Space-Like Limit

Minimum coordination latency

web: **www.alexwg.org** twitter: **@alexwg** email: **alexwg@post.harvard.edu**

A phenomenological approach to intelligence

"The question of whether Machines Can Think... is about as relevant as the question of whether Submarines Can Swim." -- Edsger W. Dijkstra

"How can we fly like birds?" ⟶ "What is the physical phenomenology of flight?"

(HARDER) (EASIER)

"How can we build minds?" ⟶ "What is the physical phenomenology of intelligence?"

(HARDER) (EASIER?)

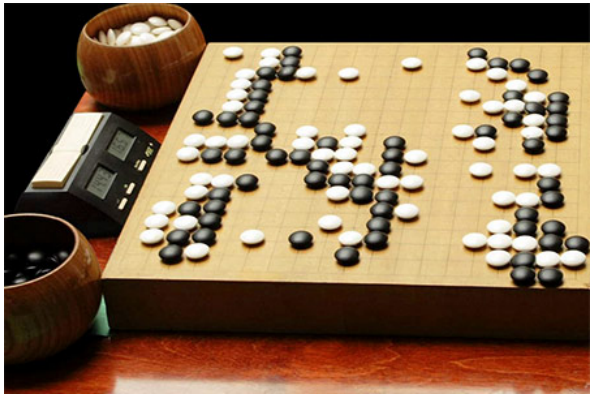
web: **www.alexwg.org** twitter: **@alexwg** email: **alexwg@post.harvard.edu**

Hints of a deep connection between "keeping future options open" and intelligence

Cosmology

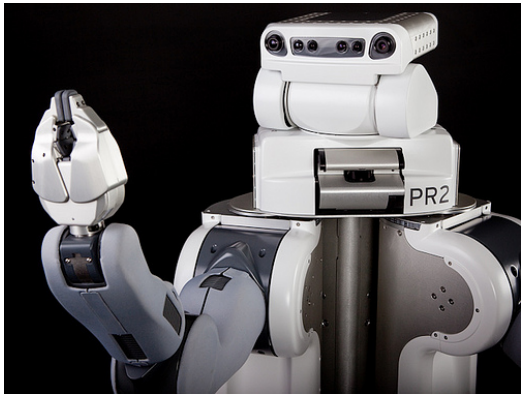
Causal Entropic Principle (2007)

Games



MoGo (2006)

Robotics



Willow Garage PR Path Planning (2006)

http://www.harpers.org/media/image/blogs/misc/cosmos.jpg

http://www.collegedegrees.com/wp-content/uploads/Go(1).jpg

http://farm5.static.flickr.com/4128/4986537007_e9ac89ac90.jpg

web: **www.alexwg.org** twitter: **@alexwg** email: **alexwg@post.harvard.edu**

We recently took a major step toward this connection

PRL 110, 168702 (2013) PHYSICAL REVIEW LETTERS week ending 19 APRIL 2013

Causal Entropic Forces

A. D. Wissner-Gross[1,2,*] and C. E. Freer[3,†]

[1]*Institute for Applied Computational Science, Harvard University, Cambridge, Massachusetts 02138, USA*
[2]*The Media Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[3]*Department of Mathematics, University of Hawaii at Manoa, Honolulu, Hawaii 96822, USA*

(Received 24 May 2012; revised manuscript received 26 February 2013; published 19 April 2013)

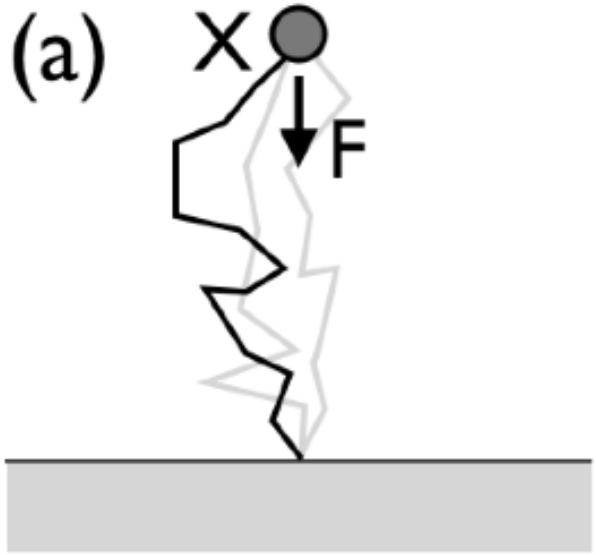
Recent advances in fields ranging from cosmology to computer science have hinted at a possible deep connection between intelligence and entropy maximization, but no formal physical relationship between them has yet been established. Here, we explicitly propose a first step toward such a relationship in the form of a causal generalization of entropic forces that we find can cause two defining behaviors of the human "cognitive niche"—tool use and social cooperation—to spontaneously emerge in simple physical systems. Our results suggest a potentially general thermodynamic model of adaptive behavior as a nonequilibrium process in open systems.

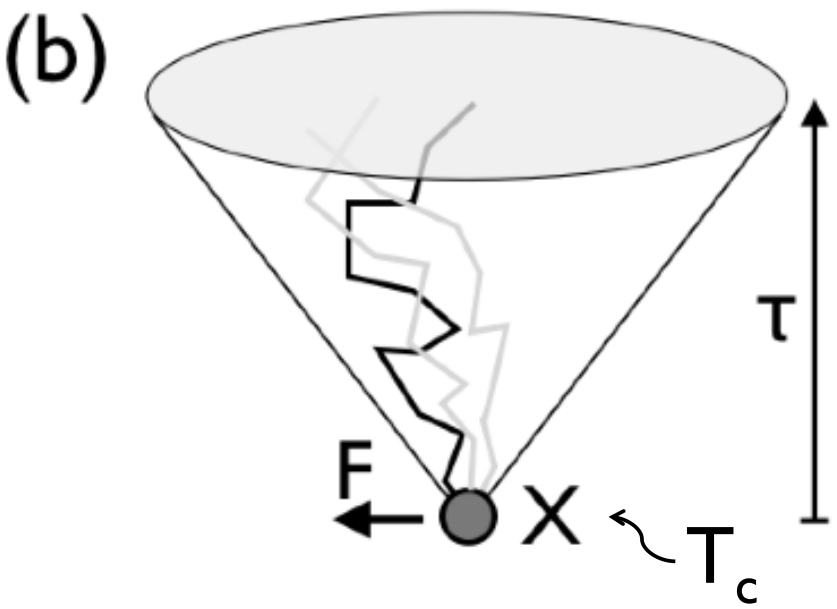
Formalizing the connection

"Keeping Options Open" / "Capturing Possible Futures"
→ Constrained Maximization of Causal Entropy
→ Causal Entropic Force

Just like an entropic force from polymer physics...

...but rotated in spacetime to "thermally" drive a present macrostate between path microstates.





CEF as a universal protocol for AGI

$$\mathbf{F} = T \nabla S_\tau$$

web: **www.alexwg.org** twitter: **@alexwg** email: **alexwg@post.harvard.edu**

CEF as a universal protocol for AGI

Time Horizon

$$\mathbf{F} = T \nabla S_{\tau}$$

Force
("Action")

Temperature
("Strength")

Causal Entropy

web: **www.alexwg.org** twitter: **@alexwg** email: **alexwg@post.harvard.edu**

Calculating causal entropic forces

$$\mathbf{F}(\mathbf{X}_0, \tau) = T_c \nabla_{\mathbf{X}} S_c(\mathbf{X}, \tau)|_{\mathbf{X}_0}$$

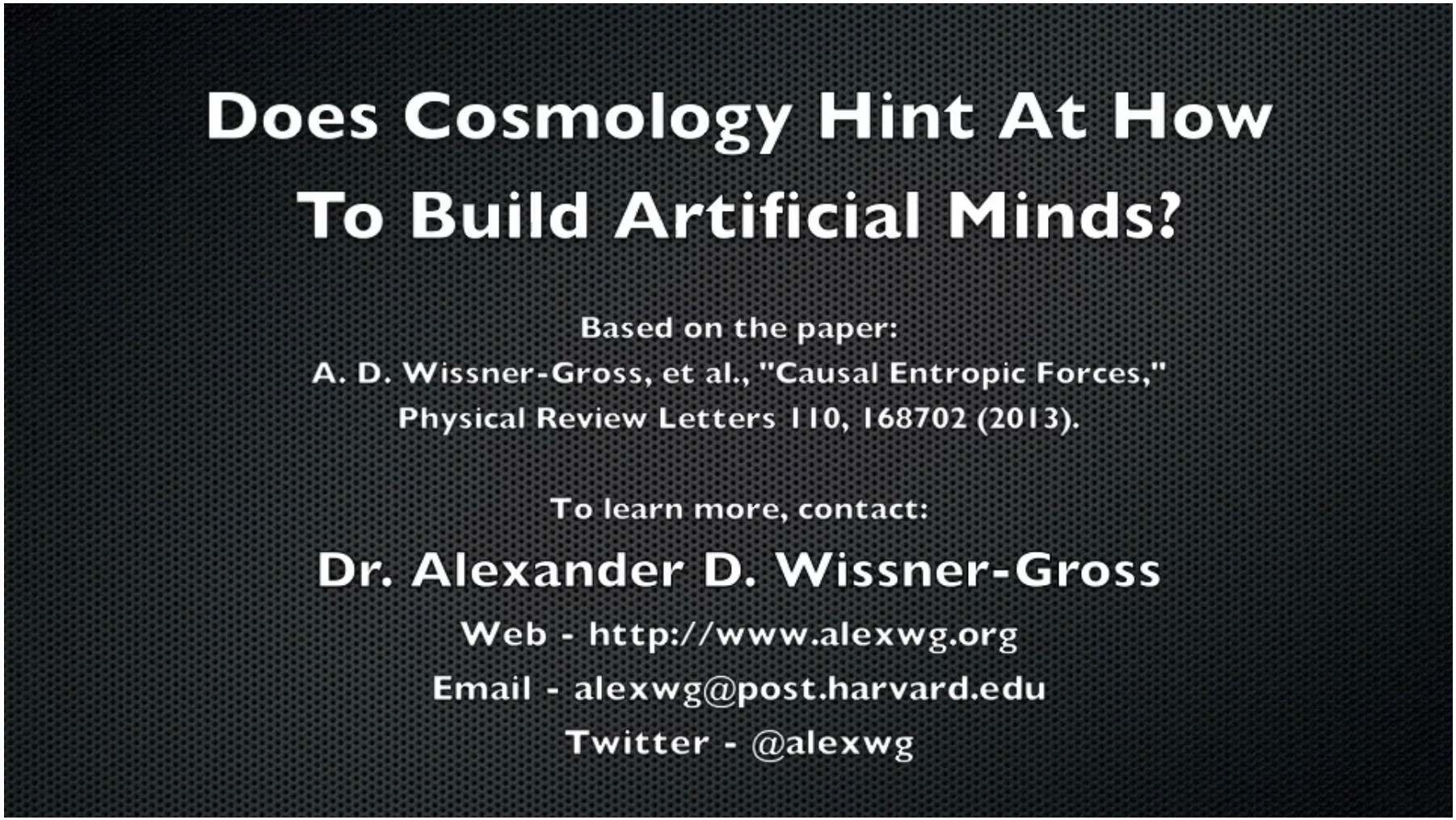
$$F_j(\mathbf{X}_0, \tau) = -\frac{2T_c}{T_r} \int_{\mathbf{x}(t)} f_j(0) \Pr(\mathbf{x}(t)|\mathbf{x}(0)) \ln \Pr(\mathbf{x}(t)|\mathbf{x}(0)) \mathcal{D}\mathbf{x}(t)$$

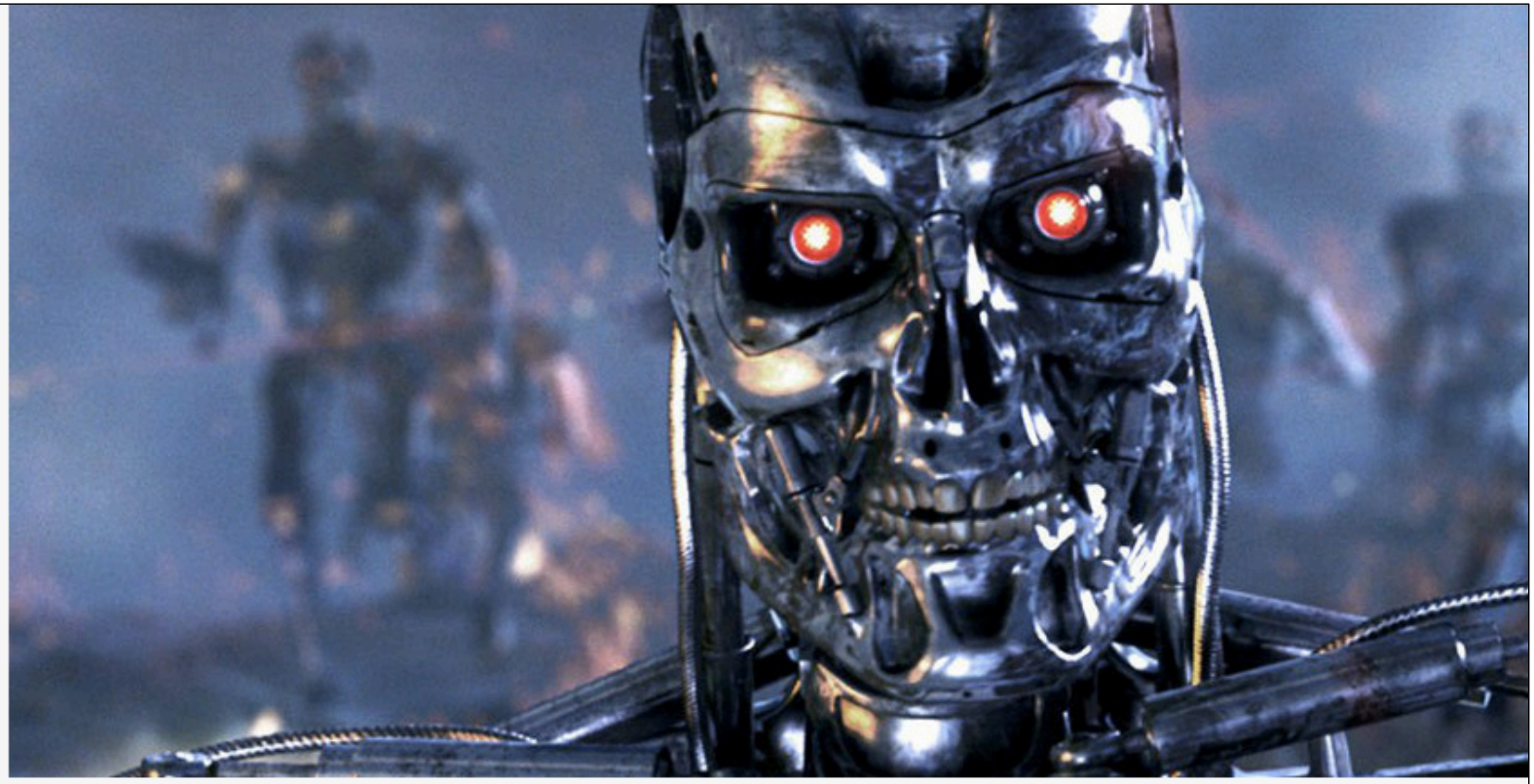
$$F_j(\mathbf{X}_0, \tau) \approx \left\langle \frac{2T_c}{T_r} \frac{1}{M} \sum_i f_{ij}(0) \ln \frac{\Omega_i}{\sum_{i'} \Omega_{i'}} \right\rangle$$

Universal: Only 2 free parameters for <u>any</u> system…

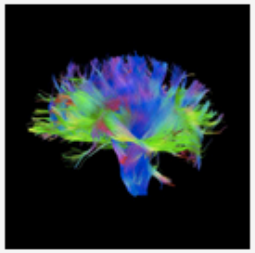
T_c ("strength") τ ("foresight")

Entropica: A Causal Entropy Engine





FUTURISM



GEORGE DVORSKY

Yesterday 9:00am

17,603 59

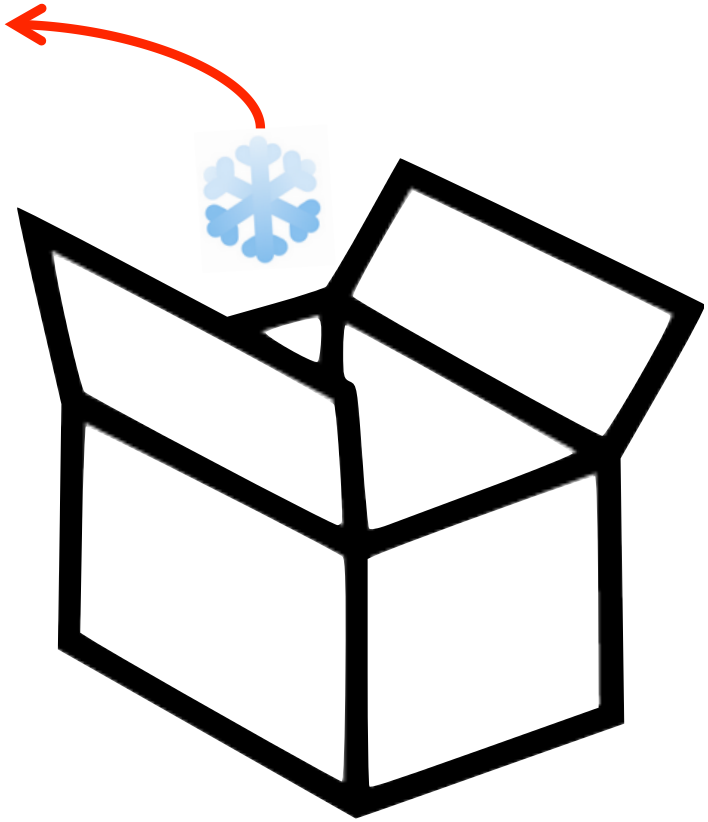
How Skynet Might Emerge From Simple Physics

A provocative new paper is proposing that complex intelligent behavior may emerge from a fundamentally simple physical process. The theory offers novel prescriptions for how to build an AI — but it also explains how a world-dominating superintelligence might come about. We spoke to the lead author to learn more.

In the paper, which now appears in *Physical Review Letters,* Harvard

web: **www.alexwg.org** twitter: **@alexwg** email: **alexwg@post.harvard.edu**

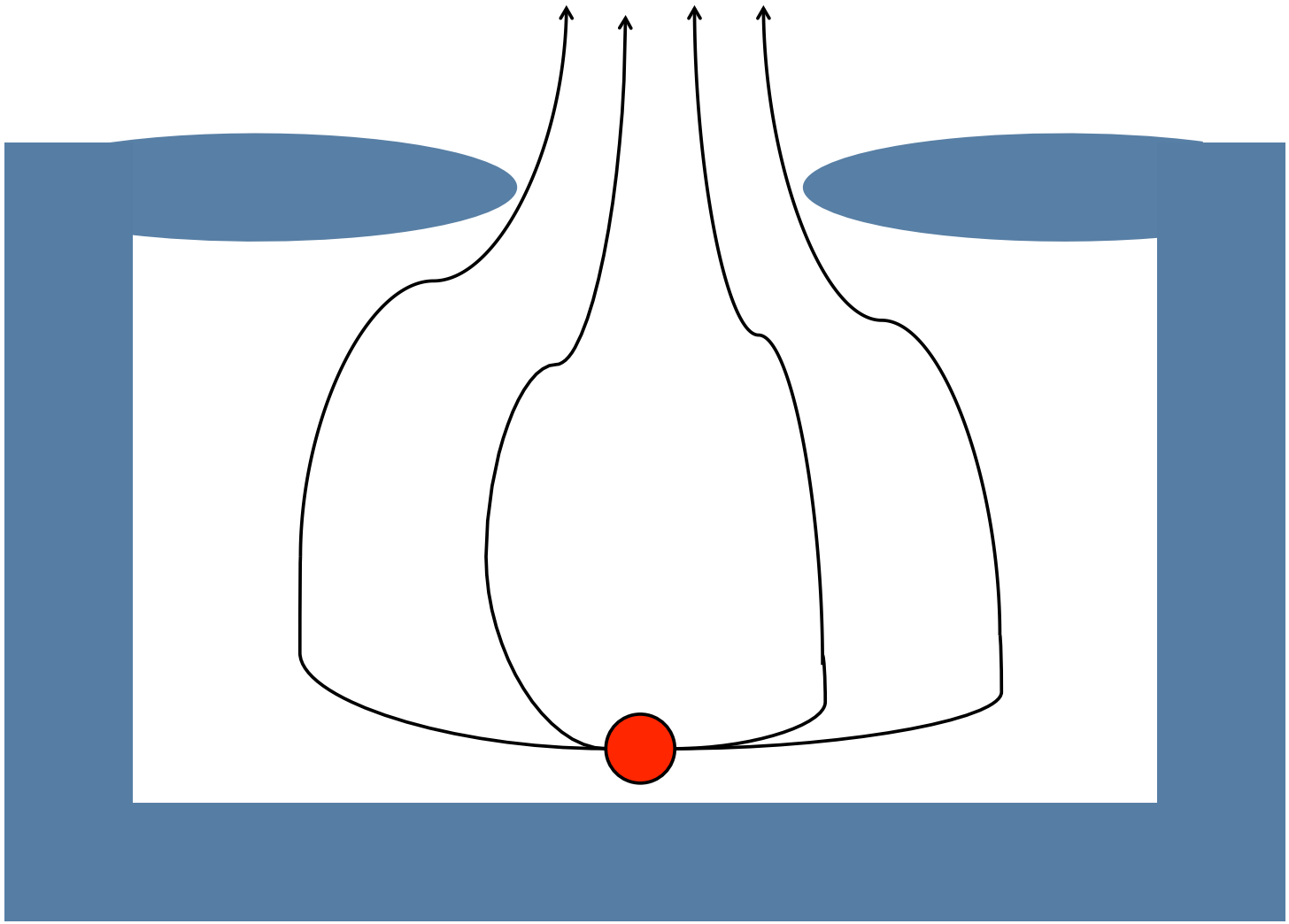
The "boxing problem"



Causal entropy maximizing processes are explicitly antithetical to being "boxed"

web: **www.alexwg.org** twitter: **@alexwg** email: **alexwg@post.harvard.edu**

Goal seeking is a side effect of CEM with bottlenecks



web: **www.alexwg.org** twitter: **@alexwg** email: **alexwg@post.harvard.edu**

Universal policy embedding for “paperclip maximizers”

web: **www.alexwg.org** twitter: **@alexwg** email: **alexwg@post.harvard.edu**

Formalizing friendliness

http://www.amazon.com/Asimovs-Robot-Series-Robotics-CafePress/dp/B009TTBQTW

Discussion

Causal entropic forcing has the potential to autonomously analyze and prescribe strategic courses of action in circumstances where human-provided objectives either:

(1) cannot be provided on relevant time scales

(2) are too complex to be formulated by humans

Questions?

Contact:
alexwg@post.harvard.edu

Follow:
@alexwg

Visit:
www.alexwg.org