

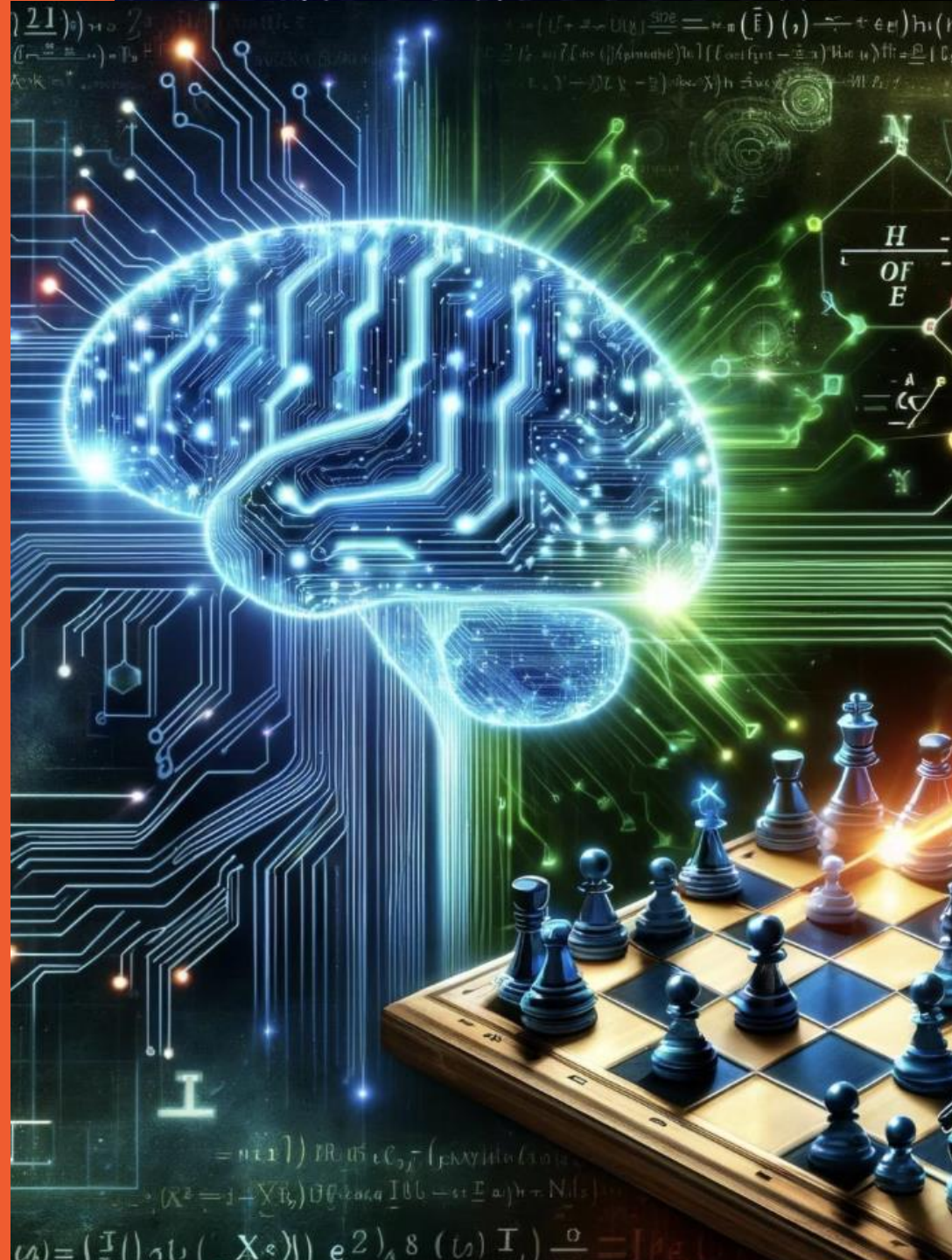
GAME THEORY, EVOLUTIONARY STABILITY, AND THE GREAT FILTER



Game Theory, Evolutionary Stability, and the Great Filter

Dr. Michael Harré
Modelling and Simulation Group
School of Computer Science
Centre for AI Trust and Governance
University of Sydney

Heron Island, January 2026



Economic Myths: “As If” and “Revealed Preferences”

Revealed Preferences Theory:

If someone consistently selects one option over another, then their preferences are revealed by their behaviour and their choices can be interpreted “as if” they are maximising a “utility value”. The alternative, that people act consistently because they have an internal utility value that allows them to order their choices, is called the “causal utility fallacy”.

Binmore, *Rational Decisions* (2009)

As If You Were Thinking:

Truly Important and significant hypotheses will be found to have “assumptions” that are wildly inaccurate descriptive representations of reality, and, in general, the more significant the theory, the more unrealistic the assumptions

Friedman, *Essays in Positive Economics* (1953)

Formal Definition of a Game

A game $G = (N, \{S_i\}, \{g_i\})$ specifies:

Players: $N = \{1, 2, \dots, n\}$

Strategies: S_i = strategy set for player i

Example: $S_1 = \{\text{Cooperate, Defect}\} = \{C, D\}$

Payoffs: $g_i(s_1, \dots, s_n)$ = payoff to player i when profile $(s_1, \dots, s_n)$ is played

Example — Prisoner's Dilemma:

$N = \{1, 2\}$

$S_1 = S_2 = \{C, D\}$

$g_1(C, C) = 3$

$g_1(C, D) = 0$

$g_1(D, C) = 5$

$g_1(D, D) = 1$

g_2 symmetric

	C	D
C	$(3, 3)$	$(0, 5)$
D	$(5, 0)$	$(1, 1)$

THIS IS NOT A DYNAMICAL SYSTEM

No $\dot{x} = f(x)$, no time evolution specified

Just a payoff map: $(s_1, s_2) \mapsto (g_1, g_2)$

Formal Definition of a Game

A game is a function

$$g: S_1 \times \cdots \times S_n \rightarrow \mathbb{R}^n,$$

where $g(s_1, \dots, s_n) = (g_1(s_1, \dots, s_n), \dots, g_n(s_1, \dots, s_n))$, and $g_i(s_1, \dots, s_n)$ is player i 's payoff when strategies $(s_1, \dots, s_n)$ are played.

Notation: $(s_1, \dots, s_n) = (s_i, s_{-i})$, where s_{-i} is the $(n - 1)$ -tuple of other players' strategies.

Nash Equilibrium

A Nash equilibrium is an n -tuple $(s_1^*, \dots, s_n^*) = (s_i^*, s_{-i}^*)$ such that

$$g_i(s_1^*, \dots, s_n^*) \geq g_i(s_i, s_{-i}^*), i = 1, \dots, n,$$

for all $s_i \in S_i$.

No player i gains by unilaterally deviating from s_i^* to any other s_i .

Nash Equilibrium as Fixed Point

Let $\psi_i(p_{-i}) =$ player i 's best-response correspondence

Then Nash equilibrium is a **fixed point**: $p^* \in \psi(p^*)$ via Kakutani's FP theorem

KEY POINT: Nash equilibrium = fixed point of best-response map,

Harsanyi spent 30 years on conditions for uniqueness



Optimality and Nash Equilibrium

Prisoner's Dilemma Payoffs:

		opponent	
		<i>C</i>	<i>D</i>
you	<i>C</i>	(3,3)	(0,5)
	<i>D</i>	(5,0)	(1,1)

Best-Response Analysis:

If opponent plays *C*: BR = *D* (payoff 5 > 3)

If opponent plays *D*: BR = *D* (payoff 1 > 0)

At (C,C): $BR(C,C) = (D,D) \neq (C,C)$

Not a fixed point of the best-response map

Both players have incentive to deviate

(C,C) is NOT a Nash equilibrium

Strategically unstable

This is the simplest “strategic” instability

Strategic Stability: Single Mixed NE

Strategic Stability:

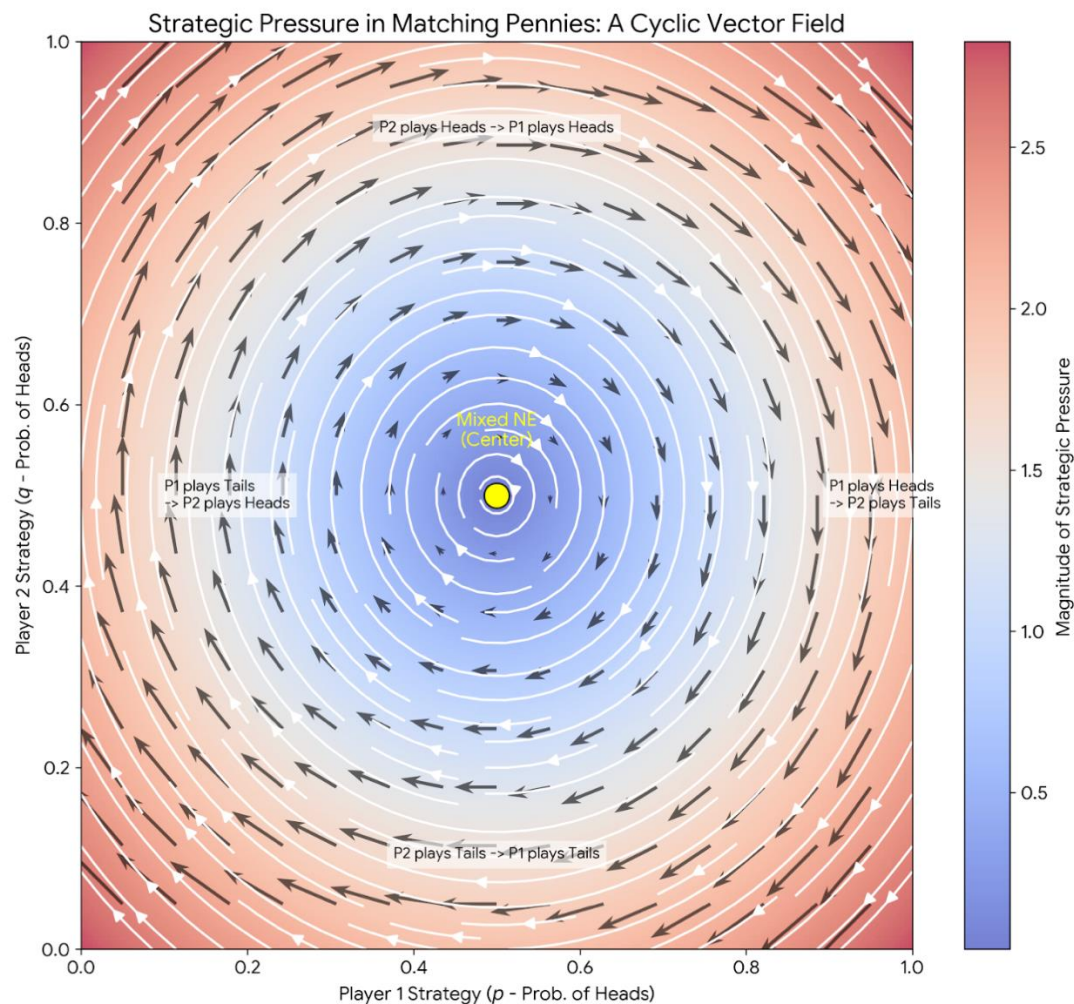


If one penny is heads and the other tails, Odd wins and keeps both coins

	Heads	Tails
Heads	+1, -1	-1, +1
Tails	-1, +1	+1, -1

Matching pennies

$$\|\nabla V\| = \sqrt{\left(\frac{\partial V}{\partial p}\right)^2 + \left(\frac{\partial V}{\partial q}\right)^2}$$



Strategic Instability: Multiple NE with 1 Mixed NE

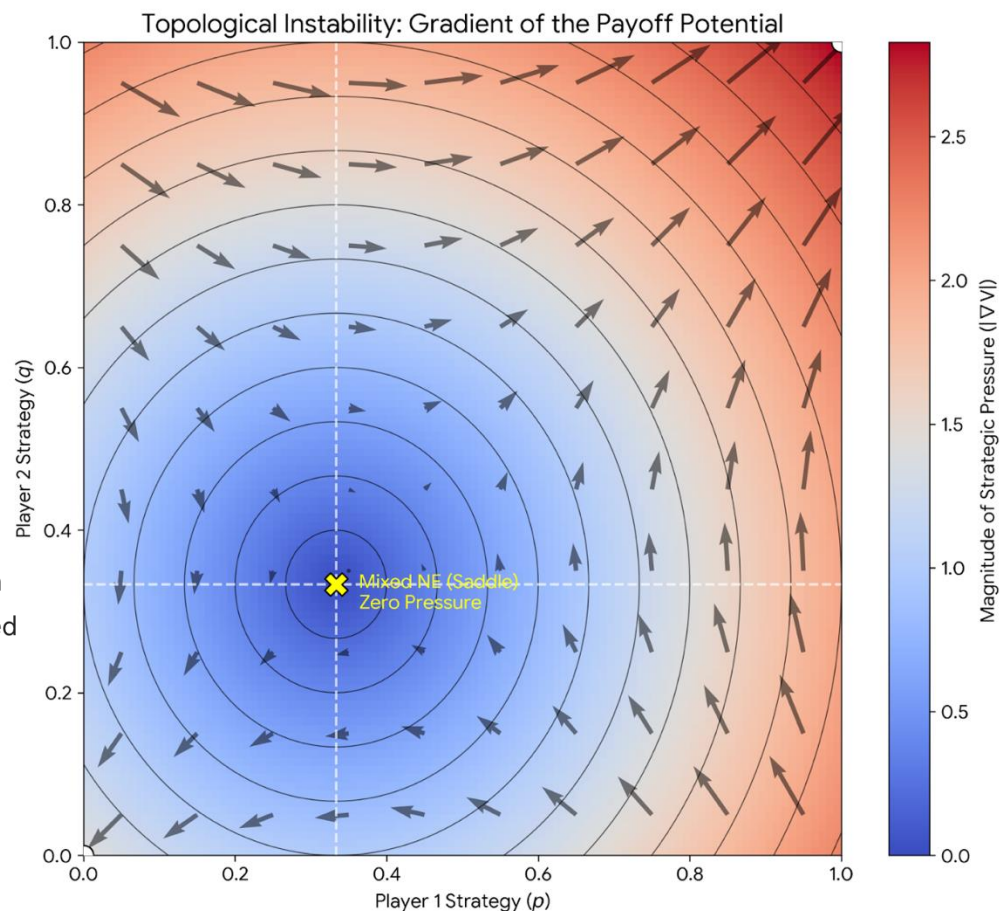
Strategic Stability:

	Action 1	Action 2
Action 1	(2, 2)	(0, 0)
Action 2	(0, 0)	(1, 1)

This heatmap visualizes the **magnitude of the strategic pressure** acting on the system at every point in the (p, q) strategy space. The pressure is defined as the norm of the payoff gradient vector:

$$\|\nabla V\| = \sqrt{\left(\frac{\partial V}{\partial p}\right)^2 + \left(\frac{\partial V}{\partial q}\right)^2}$$

where V is the potential function of the game.



Evolutionary Dynamics

	Hawk	Dove
Hawk	$(-1, -1)$	$(2, 0)$
Dove	$(0, 2)$	$(1, 1)$

1. The Proportional Growth Law

In an infinite population, the "fitness" of a strategy $f_i(x)$ is defined as its expected payoff against the current distribution x . We define the growth rate of the density x_i as being proportional to that fitness:

$$\frac{\dot{x}_i}{x_i} \propto f_i(x)$$

However, because the total probability must remain conserved ($\sum x_i = 1$), we cannot simply set $\dot{x}_i = x_i f_i$. The population must "self-normalize."

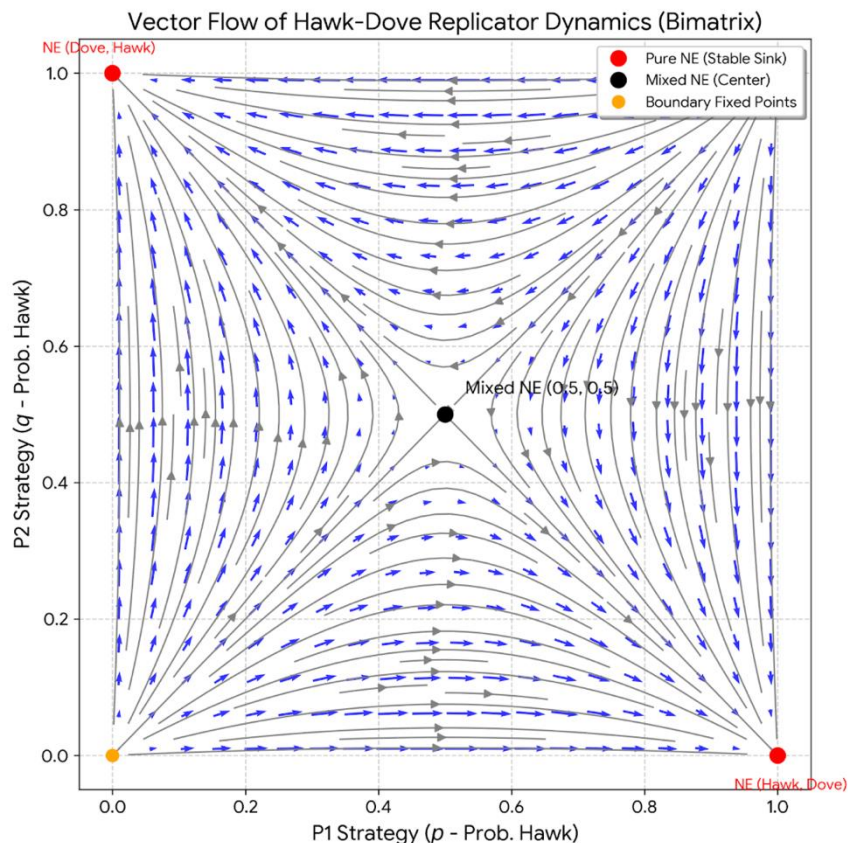
2. Conservation of Probability

Since $\sum x_i = 1$, it follows that the sum of all changes must be zero: $\sum \dot{x}_i = 0$. To ensure this, we subtract the **mean fitness** ($\phi = \sum x_j f_j$) from each individual strategy's fitness. This ensures that strategies only grow if they are performing better than the "background noise" of the population.

3. The Continuous-Time Replicator Equation

This leads directly to the standard form:

$$\dot{x}_i = x_i[f_i(x) - \phi(x)]$$

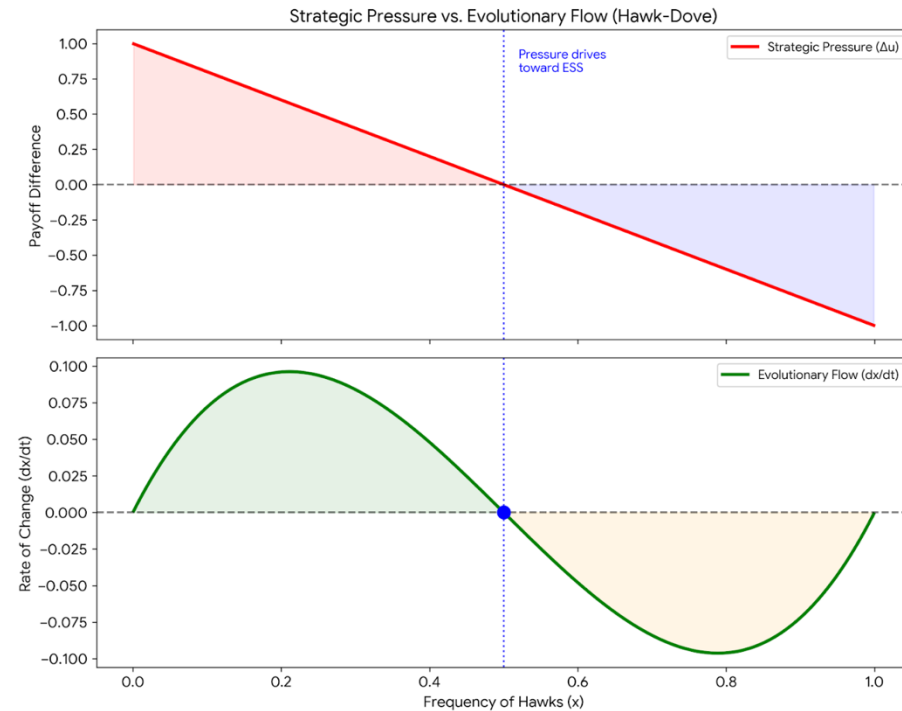


Evolutionary vs. Strategic

	Hawk	Dove
Hawk	$(-1, -1)$	$(2, 0)$
Dove	$(0, 2)$	$(1, 1)$

- **Strategic Pressure (Gradient):** The "raw" incentive Δu (the difference in payoffs). It answers: "How much more do I get if I switch *right now*?"
- **Evolutionary Gradient ($\dot{x}$):** The Replicator Dynamics. It incorporates the population frequency: $\dot{x} = x(1 - x) \cdot \Delta u$.

The "pressure" is highest at the boundaries, but the "evolutionary flow" is zero at the boundaries (because if you have no Doves, you can't "grow" Doves).



The Contrast: Pressure vs. Flow

The plot shows two different "forces" in a **Hawk-Dove** game (where the ESS is at $x = 0.5$):

Feature	Strategic Pressure (Top)	Evolutionary Flow (Bottom)
Mathematical Definition	$\Delta u = u_H - u_D$	$\dot{x} = x(1 - x)\Delta u$
Question it answers	"How much better is it to be a Hawk?"	"How fast is the population changing?"
Behavior at $x = 0$	Maximum. If there are no Hawks, the incentive to <i>become</i> one is huge.	Zero. If there are no Hawks, there are no "parents" to replicate, so the population cannot change.
Stability at 0.5	Crosses zero with a negative slope .	Crosses zero with a negative slope .

Topological Stability

Logit equilibrium [\[edit \]](#)

The most common specification for QRE is **logit equilibrium (LQRE)**. In a logit equilibrium, player's strategies are chosen according to the probability distribution:

$$P_{ij} = \frac{\exp(\lambda EU_{ij}(P_{-i}))}{\sum_k \exp(\lambda EU_{ik}(P_{-i}))}$$

	Dare	Chicken
Dare	0,0	7,2
Chicken	2,7	6,6

Topological Stability

Logit equilibrium [\[edit\]](#)

The most common specification for QRE is **logit equilibrium (LQRE)**. In a logit equilibrium, player's strategies are chosen according to the probability distribution:

$$P_{ij} = \frac{\exp(\lambda EU_{ij}(P_{-i}))}{\sum_k \exp(\lambda EU_{ik}(P_{-i}))}$$

	Dare	Chicken
Dare	0,0	7,2
Chicken	2,7	6,6

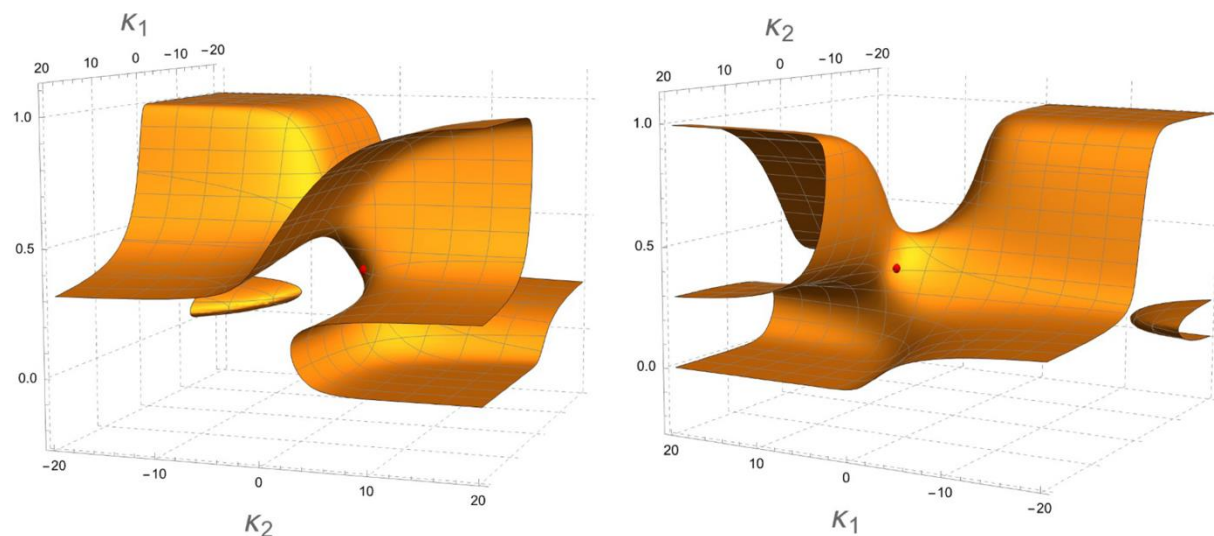


Fig. 6. Two perspectives of the quantal response equilibrium surface Σ_ω for $\omega_2 = 0.25$, $\omega_1 = 0.75$. Note that $\omega_2 = 1 - \omega_1$ implies the equation $0.75 = 1 - h(x)$ is approximately satisfied by $x \approx 0.373$, in which case the equations (†) imply $\kappa_1 = \kappa_2 = \frac{1}{f(0.373)} \approx 4.28$. The red dot indicates the unique branch point of multiplicity three with these approximate coordinates. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

“On the Smooth Unfolding of Bifurcations in Quantal Response Equilibria”,
Harris, McCallum, Harre, *Games and Economic Behavior* (2022)

Topological Stability

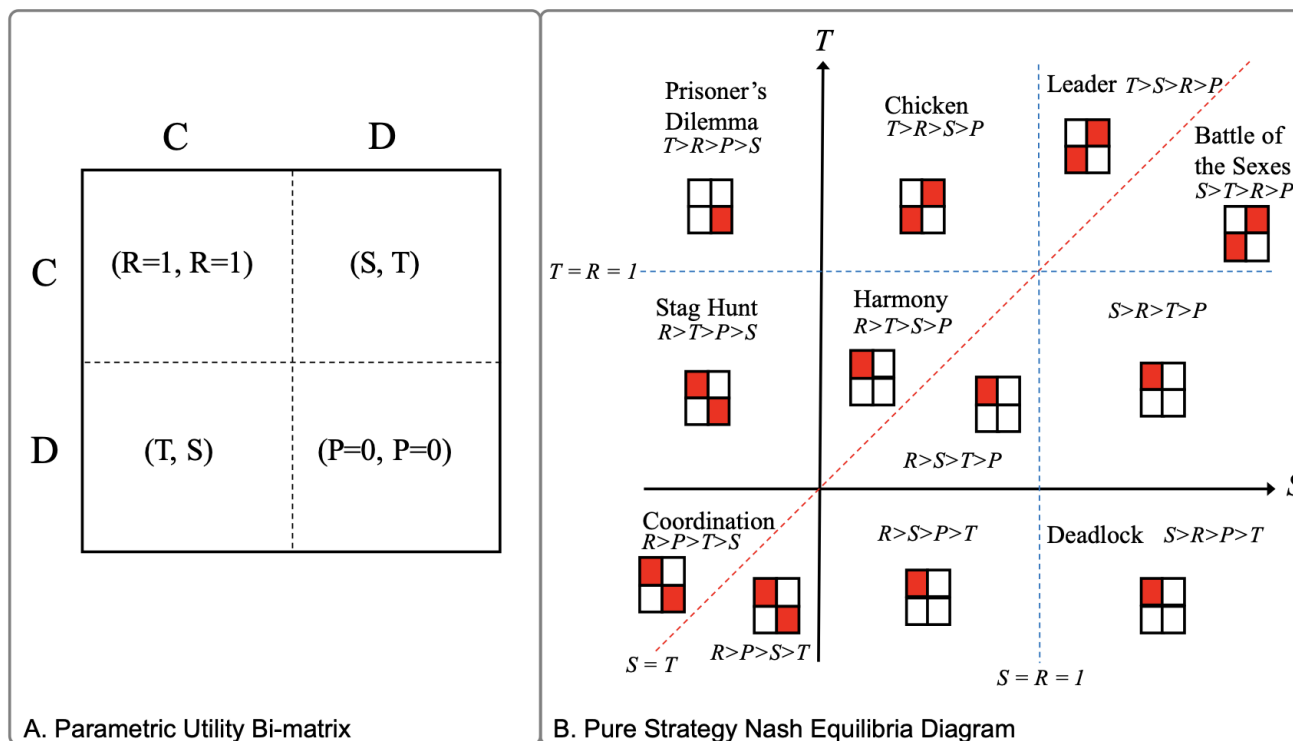
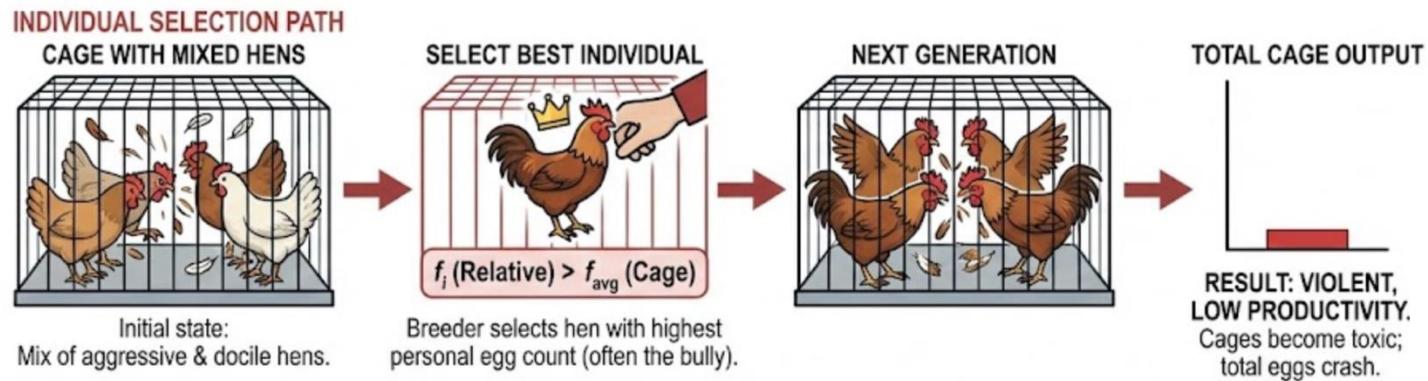


Figure 4: **Parametric games:** **A.** The utility bi-matrix with two free parameters T and S for some symmetrical games. **B.** The bifurcations of NE as T and S vary. The locations (strategy pairs) of the NE are shown as red boxes, e.g. the Harmony game has one NE: (CC) and the Stag Hunt game has two NE: (CC) , (DD) .

Multi-level Selection: Competition to Cooperation

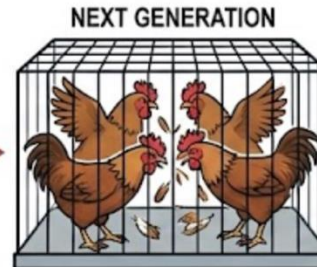
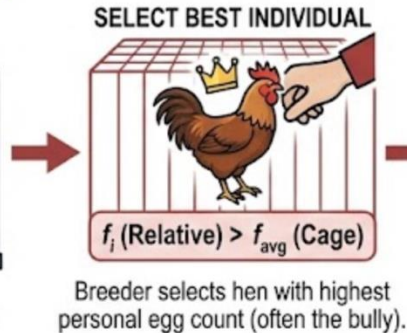
Muir's Hen Experiment: The Level of Selection Matters



Multi-level Selection: Competition to Cooperation

Muir's Hen Experiment: The Level of Selection Matters

INDIVIDUAL SELECTION PATH

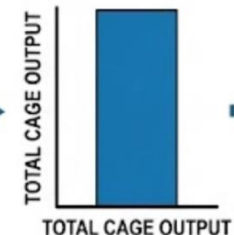
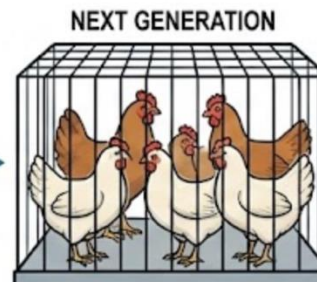
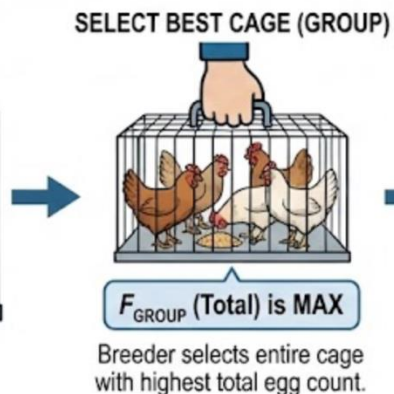


TOTAL CAGE OUTPUT



Shift in
Selection Level
=
Transformation
of Outcome

GROUP SELECTION PATH



RESULT: DOCILE, HIGH PRODUCTIVITY.
Cages become cooperative; total eggs soar.

Multi-level Selection: Competition to Cooperation



1. The Formal MLS Equation

To maintain consistency, let p be the global frequency of cooperation. We partition the population into groups j .

The total dynamics are governed by:

$$\dot{p} = \underbrace{E_j [p_j(1 - p_j)(f_{C,j} - f_{D,j})]}_{\text{Within-Group Selection}} + \underbrace{\text{Cov}(\bar{f}_j, p_j)}_{\text{Between-Group Selection}}$$

Multi-level Selection: Competition to Cooperation

$$\dot{p} = \underbrace{E_j[p_j(1 - p_j)(f_{C,j} - f_{D,j})]}_{\text{Within-Group Selection}} + \underbrace{\text{Cov}(\bar{f}_j, p_j)}_{\text{Between-Group Selection}}$$

The Price Equation: Interpretation of Group Dynamics

Term	Physical Reality	Interpretation
$(p_j - \bar{p})$	Group Diversity	The Signal: Deviation of a group from the population mean; the "raw material" for selection.
$(\bar{f}_j - \bar{f})$	Group Performance	The Reward: Competitive advantage of a group; shows if a collective structure yields more "fruit."
$\text{Cov}(\bar{f}_j, p_j) > 0$	Selection Alignment	The Mechanism: Systematic coupling of cooperation and success; the "engine" of group-driven evolution.

Multi-level Selection: Competition to Cooperation

$$\Delta p_{total} = \underbrace{\Delta p_{within_groups}}_{L1} + \underbrace{\Delta p_{between_groups}}_{L2} + \underbrace{\Delta p_{between_meta-populations}}_{L3} \dots$$

The Price Equation: Interpretation of Group Dynamics

Term	Physical Reality	Interpretation
$(p_j - \bar{p})$	Group Diversity	The Signal: Deviation of a group from the population mean; the "raw material" for selection.
$(\bar{f}_j - \bar{f})$	Group Performance	The Reward: Competitive advantage of a group; shows if a collective structure yields more "fruit."
$\text{Cov}(\bar{f}_j, p_j) > 0$	Selection Alignment	The Mechanism: Systematic coupling of cooperation and success; the "engine" of group-driven evolution.

Multi-level Selection: Competition to Cooperation

$$\Delta p_{total} = \underbrace{\Delta p_{within_groups}}_{L1} + \underbrace{\Delta p_{between_groups}}_{L2} + \underbrace{\Delta p_{between_meta-populations}}_{L3} \dots$$

The Price equation [23, 24] provides a formal framework for understanding how selection operates across multiple levels. In its simplest form for economic organizations, the change in average organizational performance ($\Delta \bar{\pi}$) can be decomposed into [25]:

$$\bar{w} \Delta \bar{\pi} = \underbrace{\text{Cov}_g(w^g, \pi^g)}_{\text{between-group}} + \underbrace{\mathbb{E}_g [\text{Cov}_i(w_i^g, \pi_i^g)]}_{\text{within-group}} \quad (1)$$

For AI-augmented organizations, we must explicitly distinguish between human and AI agents, as they may face different selection pressures and operate under different constraints:

$$\begin{aligned} \bar{w} \Delta \bar{\pi} = & \underbrace{\text{Cov}_g(w^g, \bar{\pi}^g)}_{\text{between-group}} + \underbrace{\mathbb{E}_g [\text{Cov}_{h \in g}(w_h^g, \pi_h^g)]}_{\text{within-group (human)}} + \underbrace{\mathbb{E}_g [\text{Cov}_{ai \in g}(w_{ai}^g, \pi_{ai}^g)]}_{\text{within-group (AI)}} \\ & + \underbrace{\mathbb{E}_g [\text{Cov}_{(h, ai) \in g}(w_h^g, \pi_{ai}^g)] + \mathbb{E}_g [\text{Cov}_{(h, ai) \in g}(w_{ai}^g, \pi_h^g)]}_{\text{cross-agent interaction terms}} \end{aligned} \quad (2)$$

Multi-level Selection: Cooperation to Competition

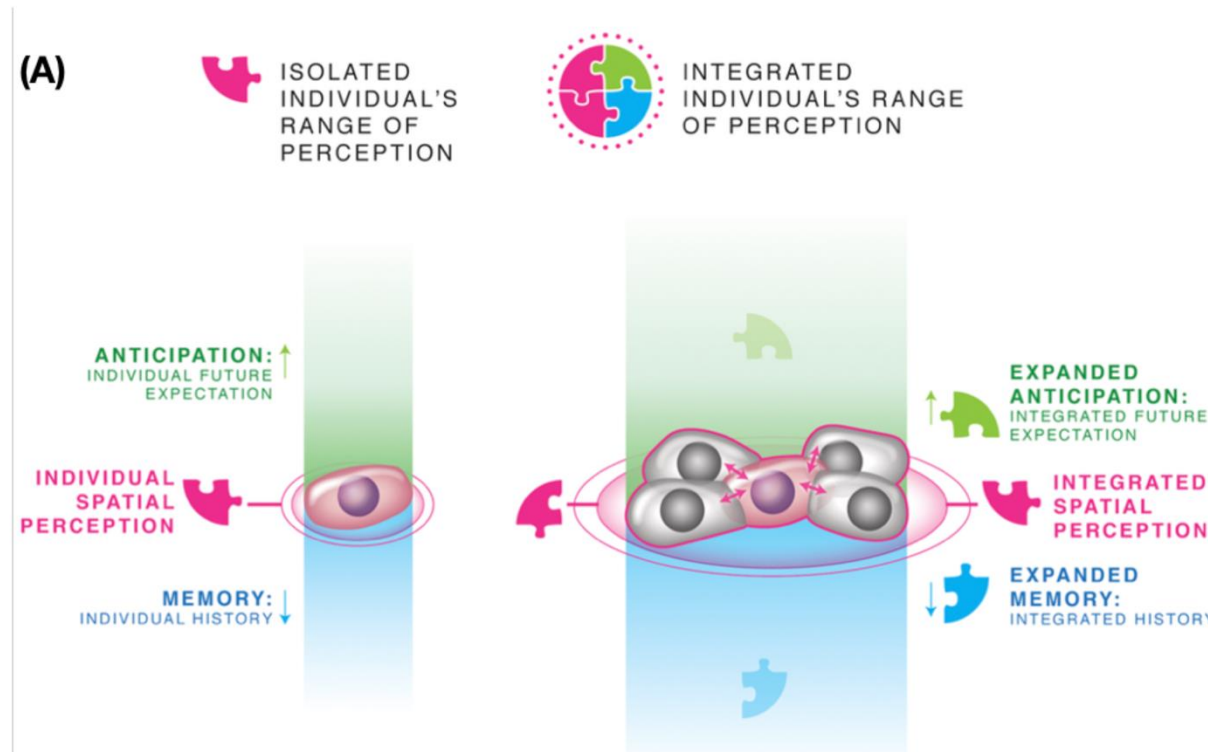
1. The Disruption of the "Multicellular Container"

In a healthy organism, **Between-Group Selection** (the survival and fitness of the organism) is the primary evolutionary force. The individual-level replicator dynamics are suppressed by tumor-suppressor genes and immune checkpoints, ensuring that internal competition remains near zero.

- **Healthy State:** The "cost" of cooperation is low, and the "punishment" for individual defection is high, keeping the population of cells in a stable, cooperative state.
- **Cancerous State:** Mutations break the structural and bioelectric "seal" of the group. The cell reverts to its ancestral unicellular logic, where its **relative fitness** within the local tissue environment becomes the only driver of its replication.

Multi-Level Selection Acts at Arbitrary Scales

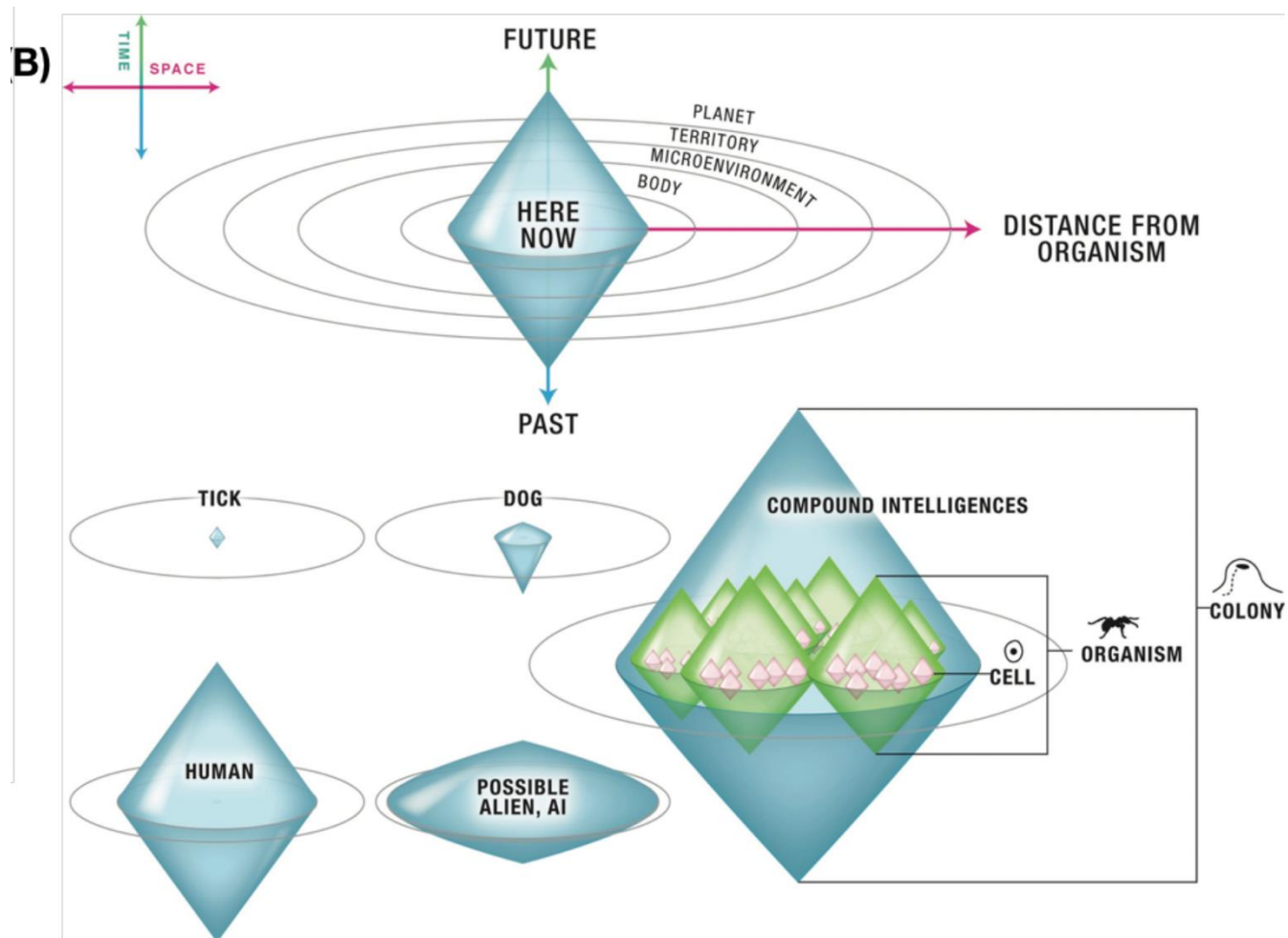
$$\Delta p_{total} = \underbrace{\Delta p_{within_groups}}_{L1} + \underbrace{\Delta p_{between_groups}}_{L2} + \underbrace{\Delta p_{between_meta-populations}}_{L3} \dots$$



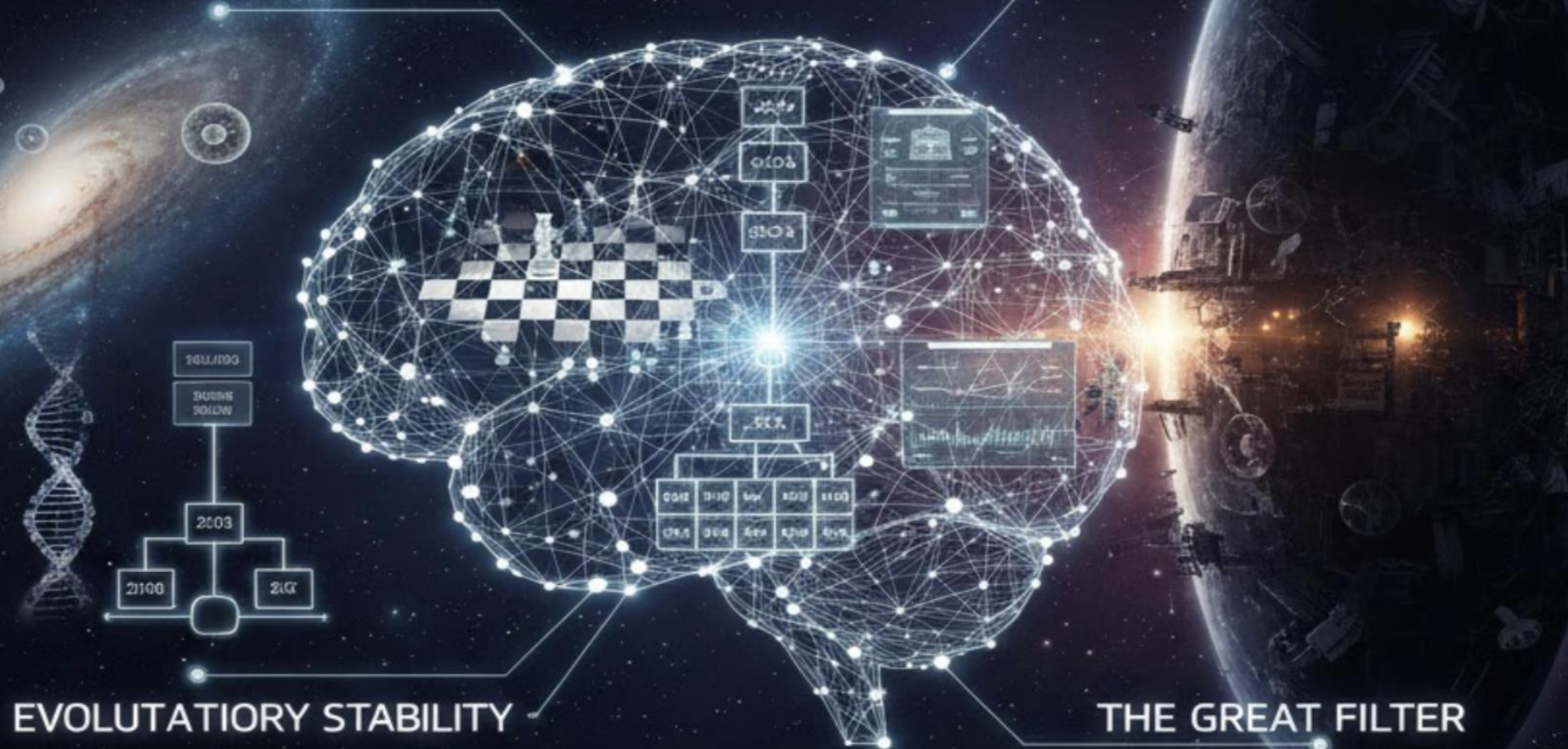
“Ingressing Minds: Causal Patterns Beyond Genetics and Environment in Natural, Synthetic, and Hybrid Embodiments”

Multi-Level Selection Acts at Arbitrary Scales

$$\Delta p_{total} = \underbrace{\Delta p_{within_groups}}_{L1} + \underbrace{\Delta p_{between_groups}}_{L2} + \underbrace{\Delta p_{between_meta-populations}}_{L3} \dots$$



GAME THEORY, EVOLUTIONARY STABILITY, AND THE GREAT FILTER



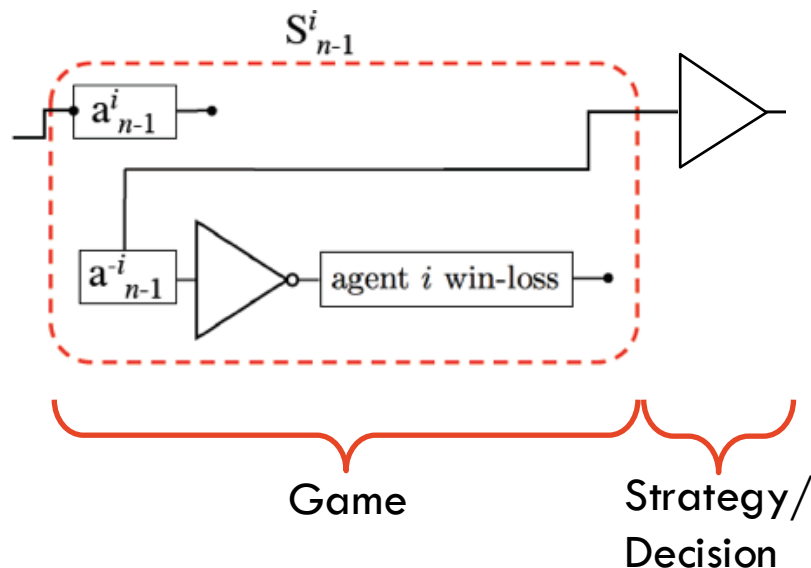
EVOLUTIONARY STABILITY

THE GREAT FILTER

On Games and Game Theory: Circuits

Decision strategies used in iterated games are also logic gates

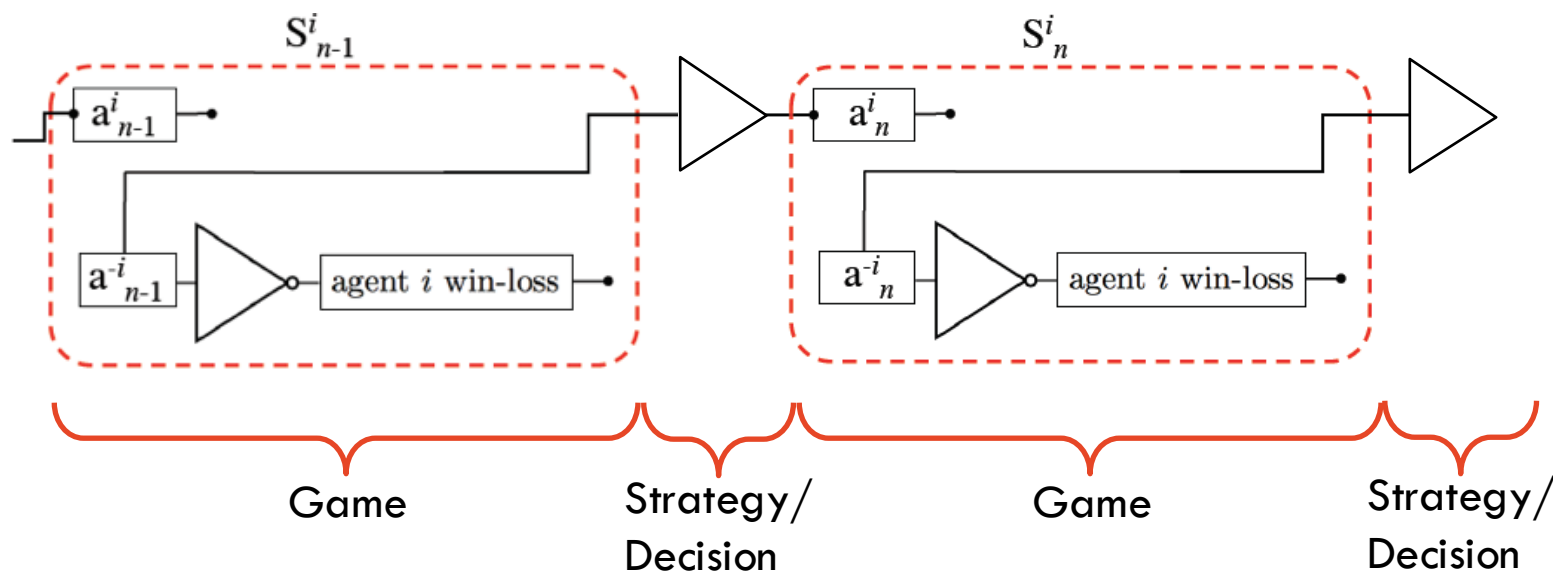
- Tit-for-Tat: A buffer (COPY) gate
- Win-Stay, Lose-Switch: An XNOR gate



On Games and Game Theory: Circuits

Decision strategies used in iterated games are also logic gates

- Tit-for-Tat: A buffer (COPY) gate
- Win-Stay, Lose-Switch: An XNOR gate



On Games and Game Theory: Circuits

Decision strategies used in iterated games are also logic gates

- Tit-for-Tat: A buffer (COPY) gate
- Win-Stay, Lose-Switch: An XNOR gate

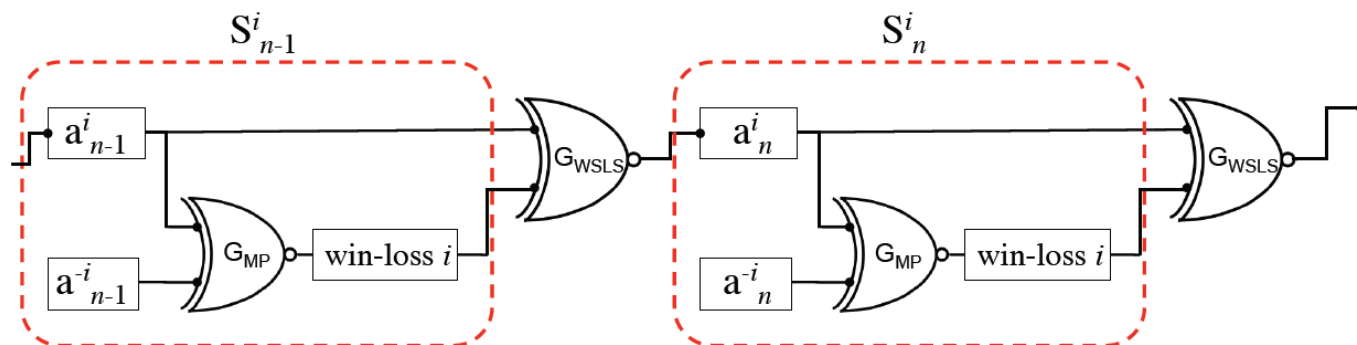
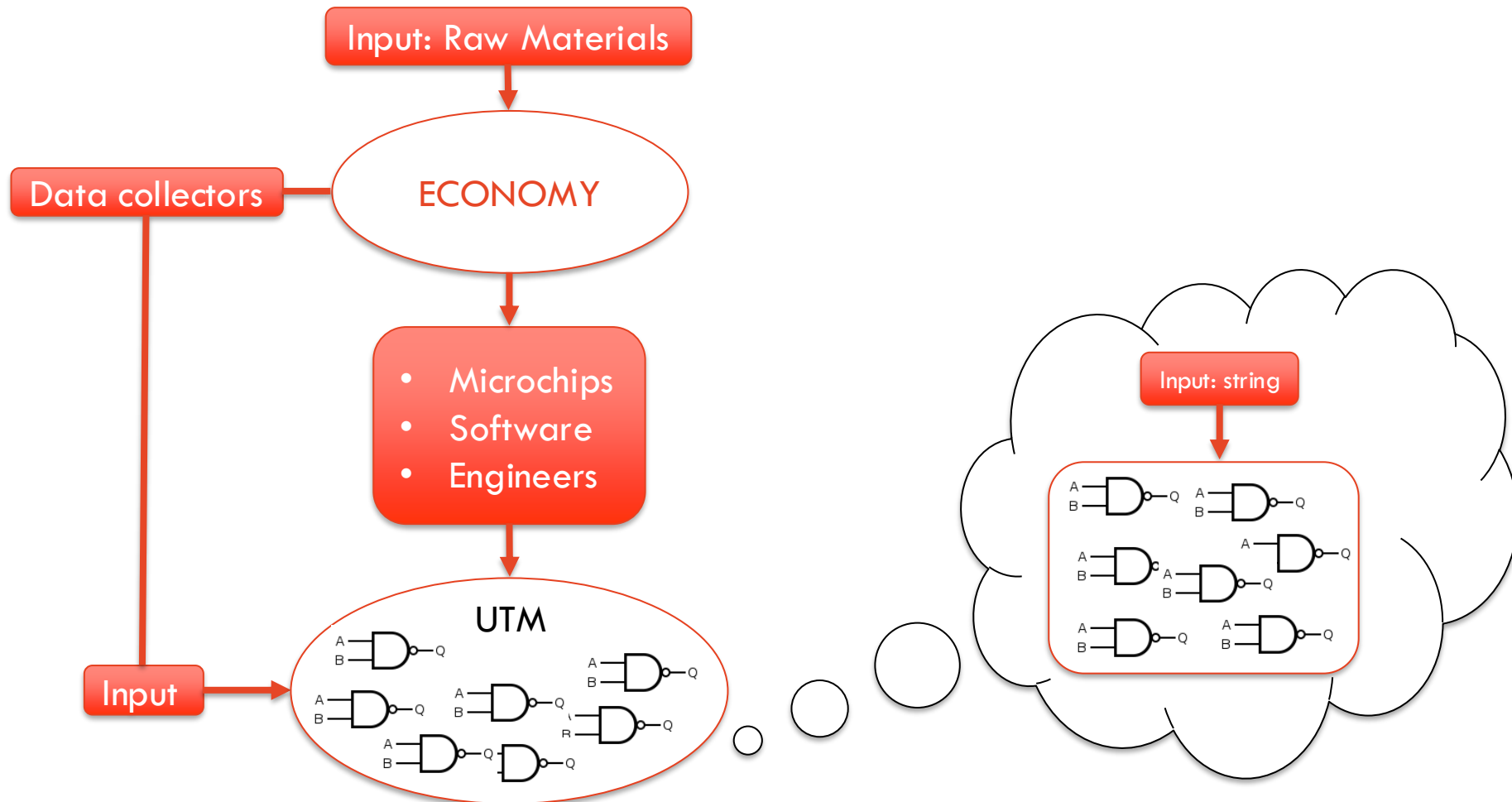


Figure 4. Two steps in the MP iterated game, as for Figure 3 the elements of S_n^i are related to one another via an XNOR logic gate G_{MP} representation of the MP game. The WSLS strategy is represented as an XNOR logic gate G_{WSLS} between inputs $\{a_{n-1}^i, \text{win-loss}\}$ and output a_n^i . For agent $-i$ the G_{MP} gate would be an XOR gate.

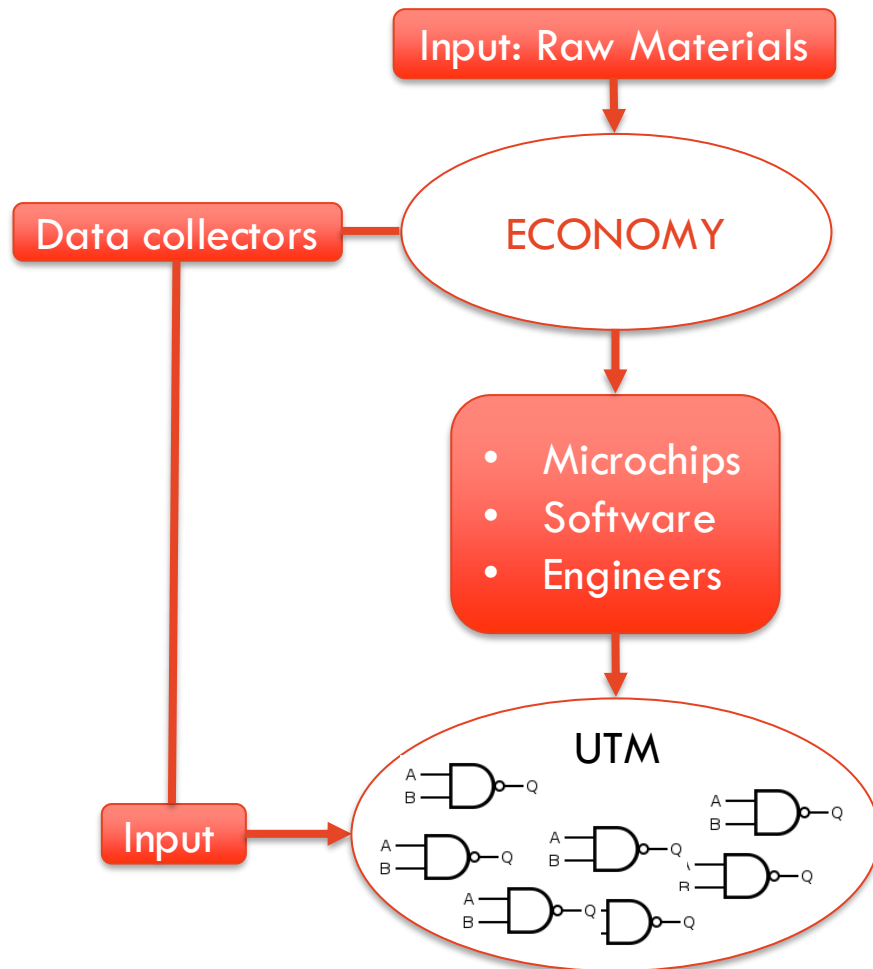
Economics, Computation and Turing Machines

Why is universal computation carried out by the economy?



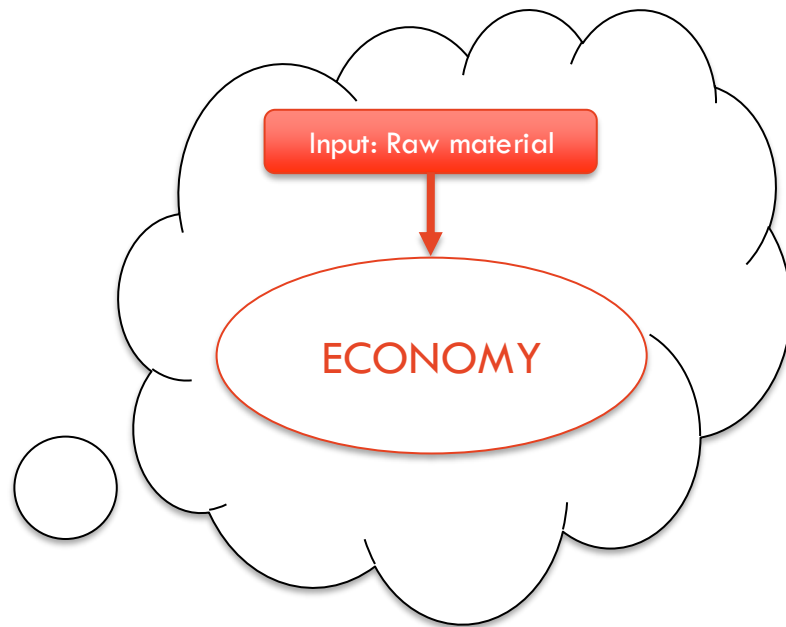
Economics, Computation and Turing Machines

Why is universal computation carried out by the economy?

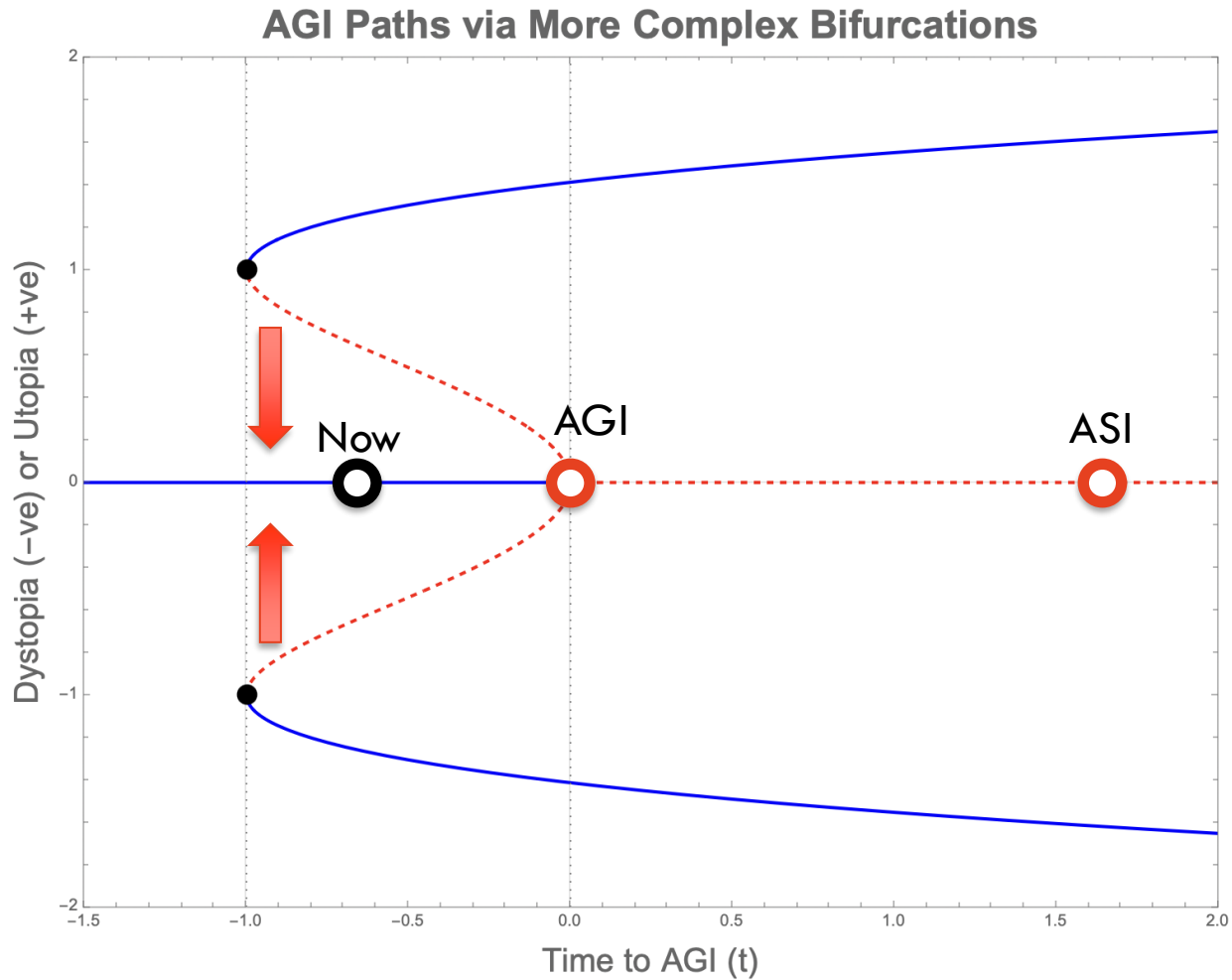


Only a UTM can produce a UTM as the output of its functioning

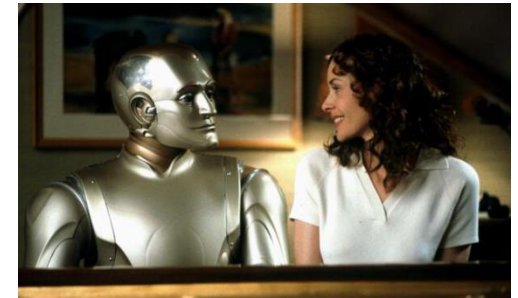
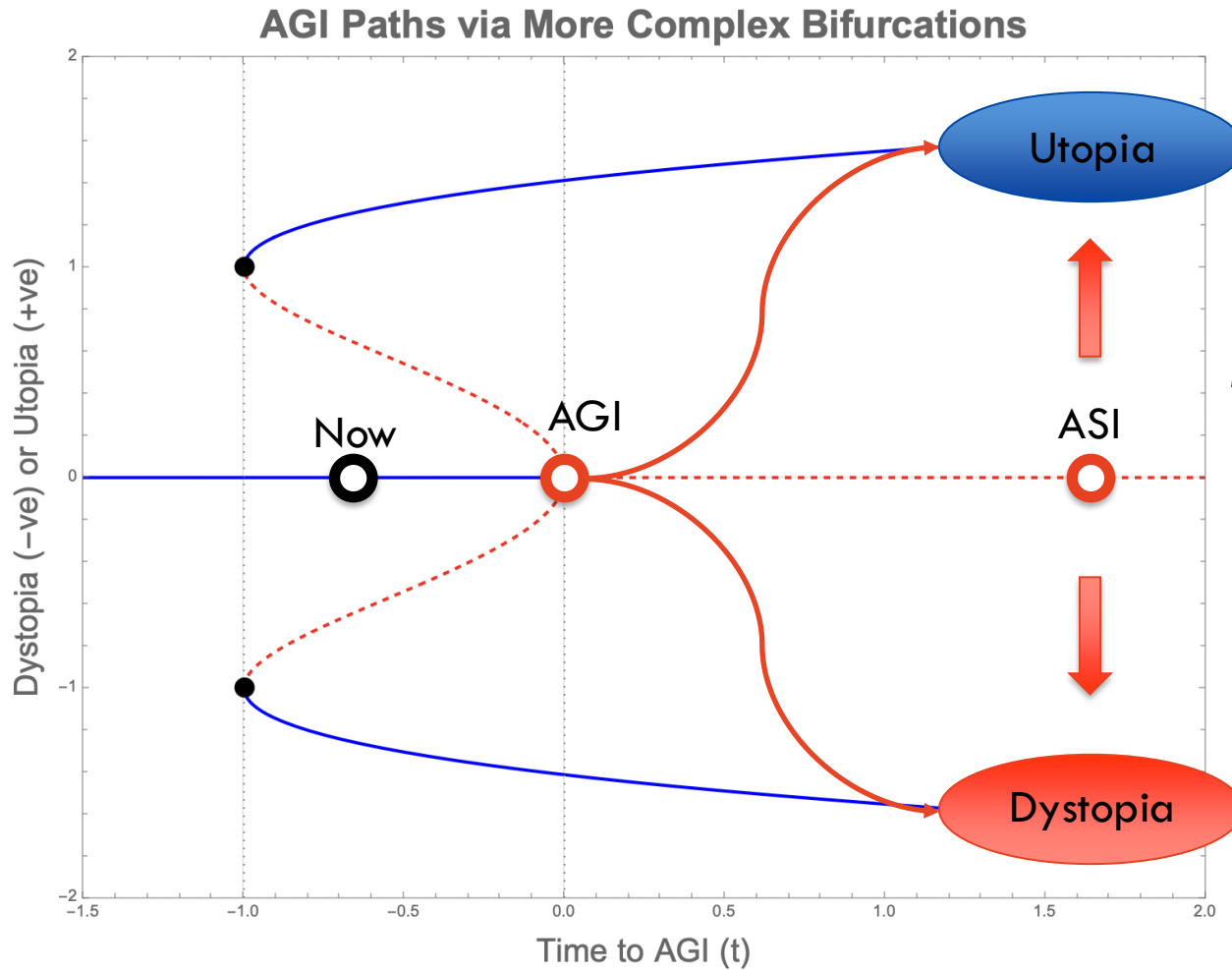
Consequently, an economy complex enough to produce a UTM can simulate itself.



Tipping Points and Macro-Economics



Tipping Points and Macro-Economics

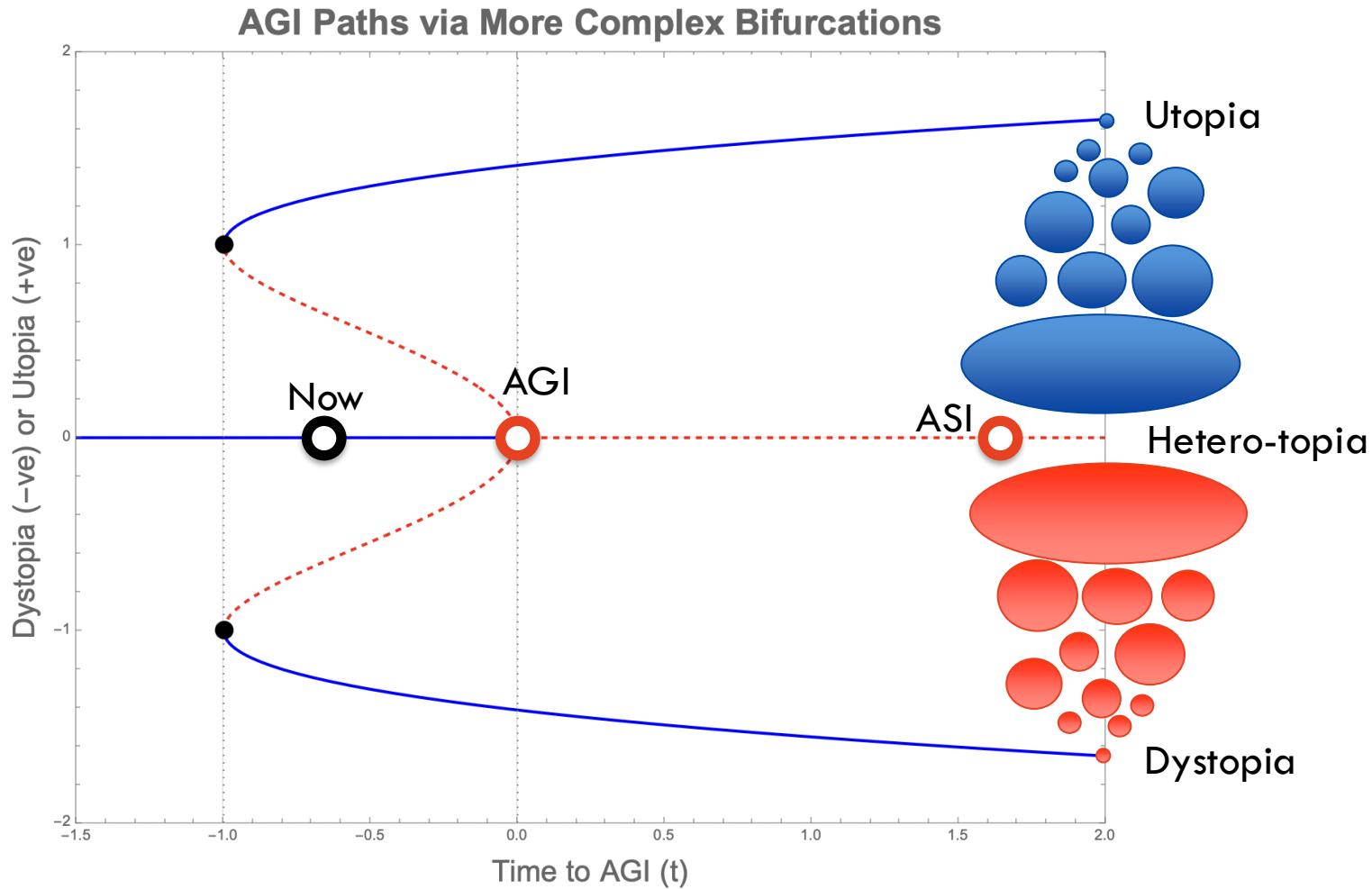


Bicentennial Man (1999 , Touchstone Pictures)

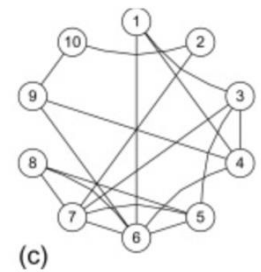
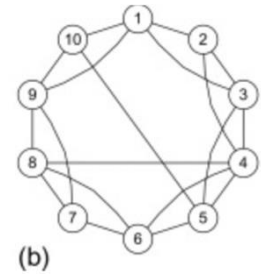
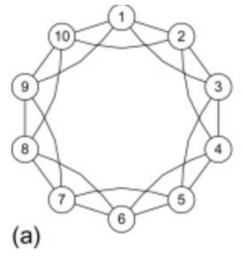
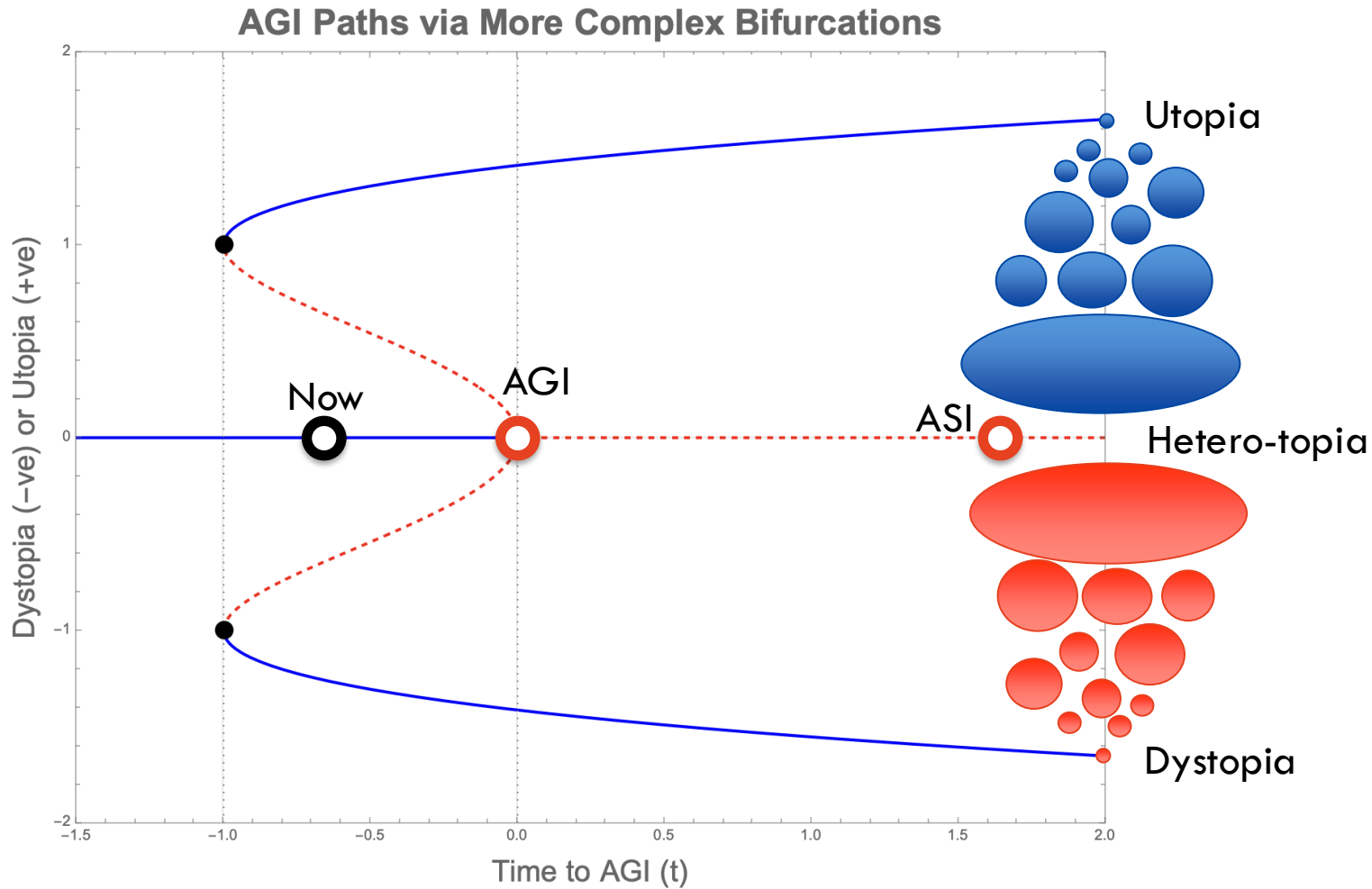


I, Robot (2004, 20th Century Fox)

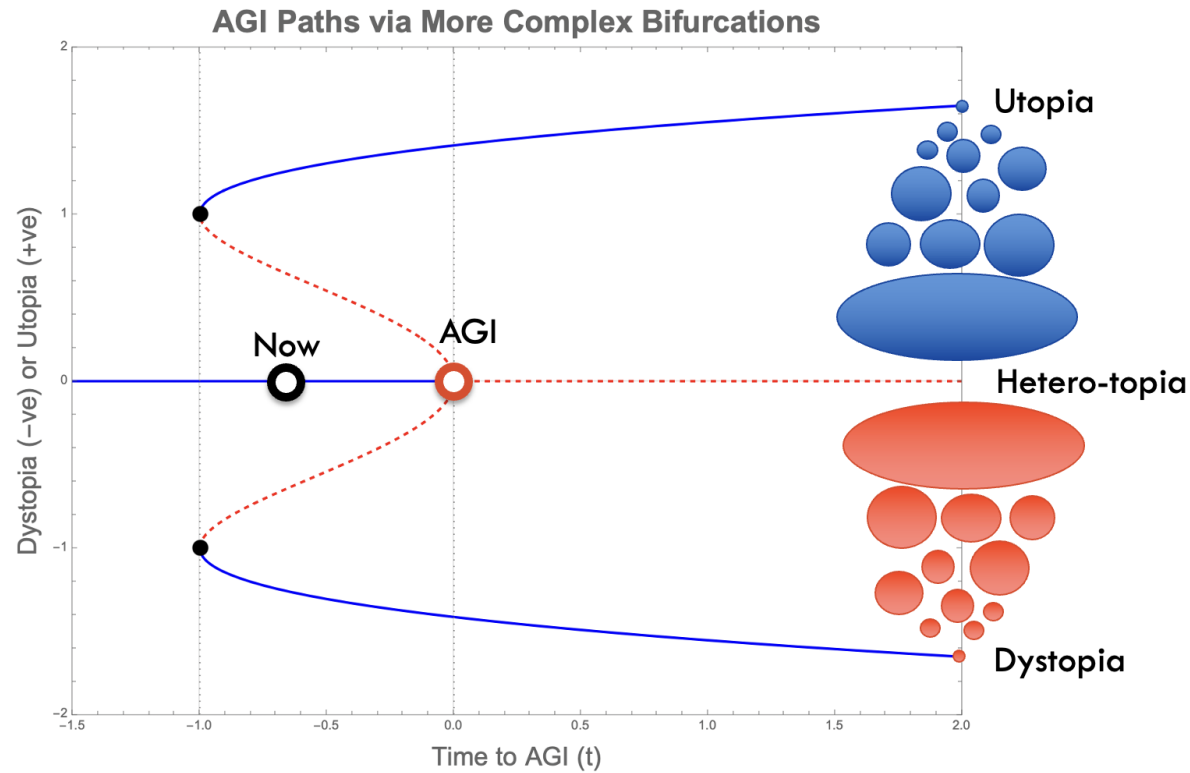
Tipping Points and Macro-Economics



Tipping Points and Macro-Economics



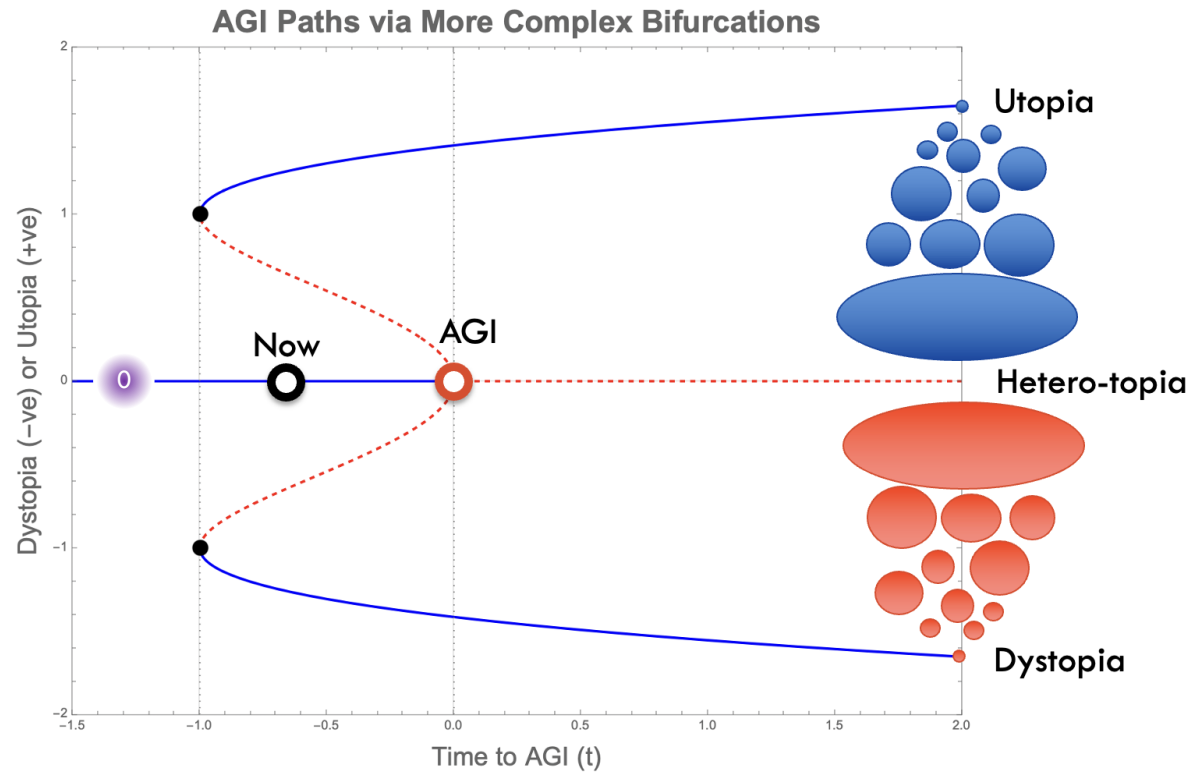
Tipping Points and Macro-Economics



Position: Levels of AGI for Operationalizing Progress on the Path to AGI

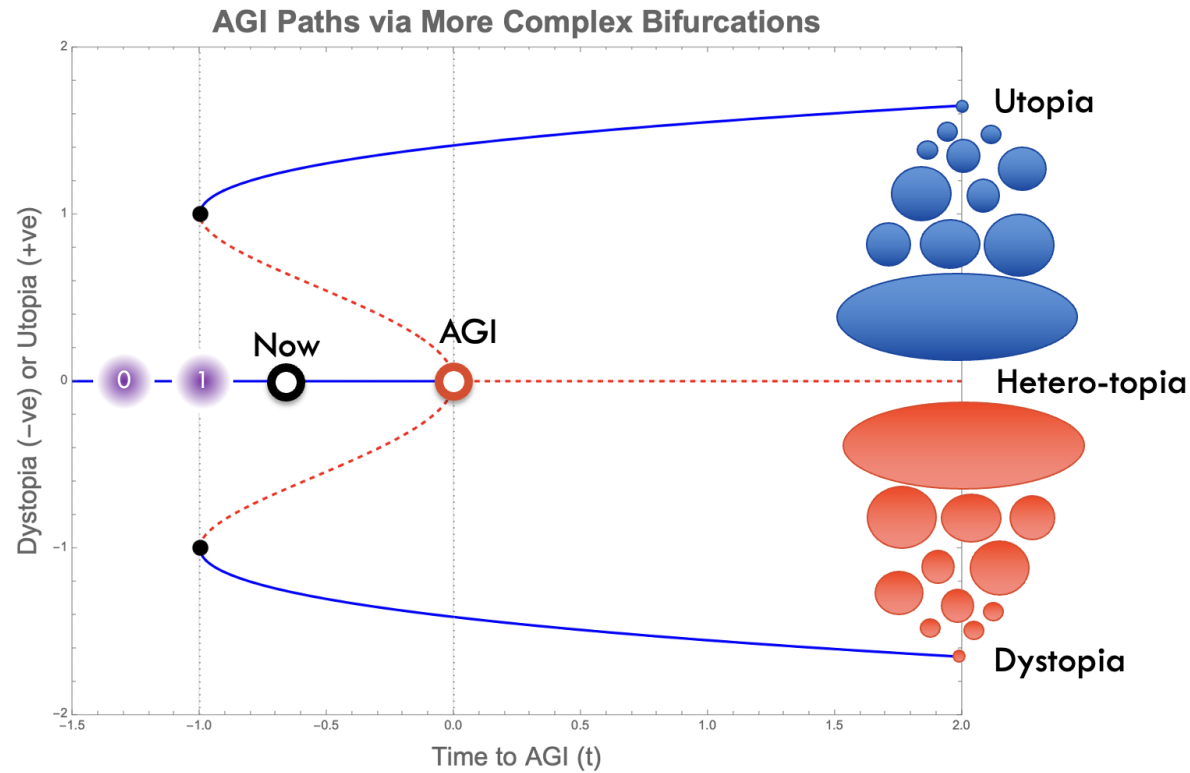
Meredith Ringel Morris¹ Jascha Sohl-Dickstein² Noah Fiedel² Tris Wartkentin² Allan Dafoe³
Aleksandra Faust² Clement Farbare³ Shane Legg³

Tipping Points and Macro-Economics



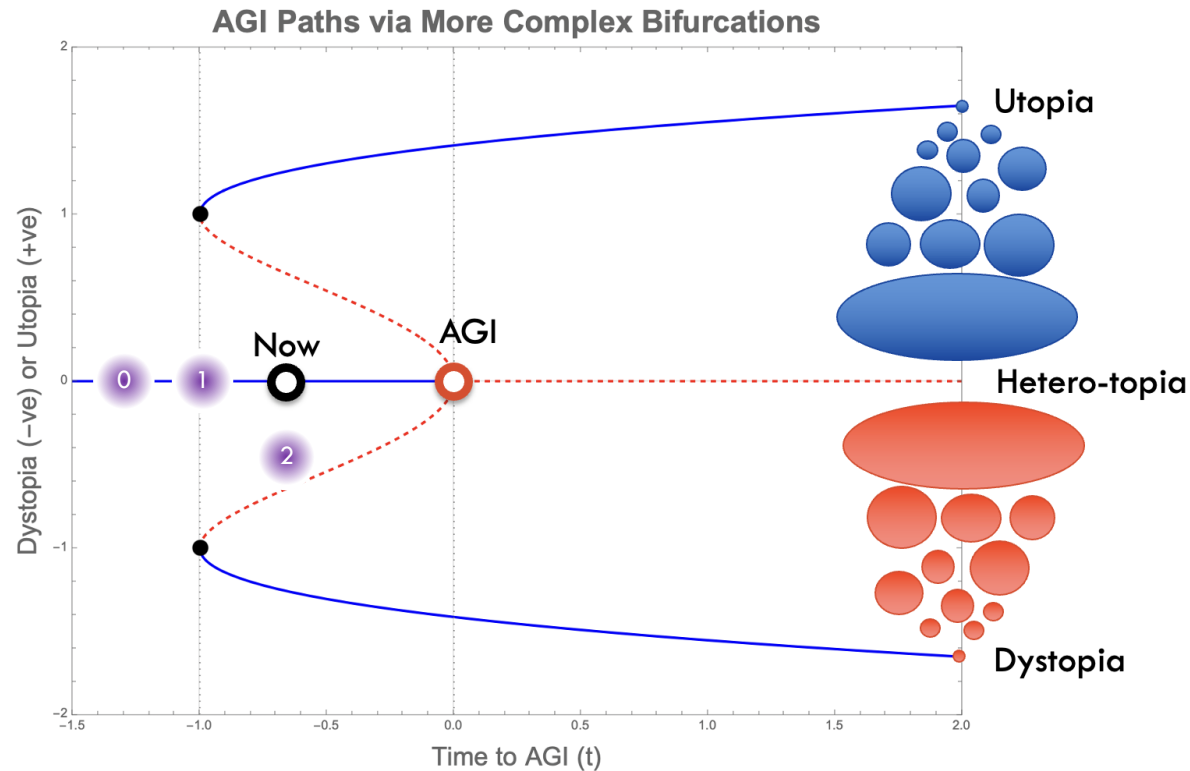
Autonomy Level	Example Systems	Unlocking AGI Level(s)	Example Risks Introduced
Autonomy Level 0: No AI <i>human does everything</i>	Analogue approaches (e.g., sketching with pencil on paper) Non-AI digital workflows (e.g., typing in a text editor; drawing in a paint program)	No AI	n/a (status quo risks)

Tipping Points and Macro-Economics



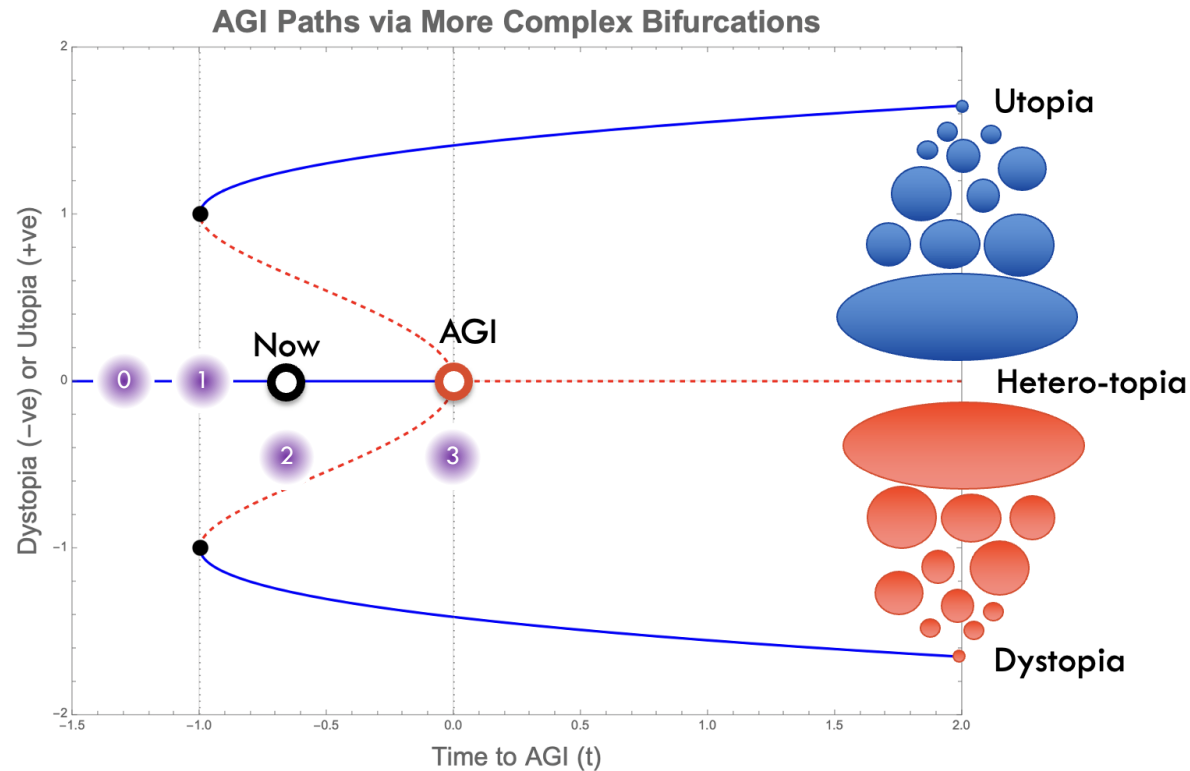
Autonomy Level	Example Systems	Unlocking AGI Level(s)	Example Risks Introduced
Autonomy Level 1: AI as a Tool <i>human fully controls task and uses AI to automate mundane sub-tasks</i>	Information-seeking with the aid of a search engine Revising writing with the aid of a grammar-checking program Reading a sign with a machine translation app	Possible: Emerging Narrow AI Likely: Competent Narrow AI	de-skilling (e.g., over-reliance) disruption of established industries

Tipping Points and Macro-Economics



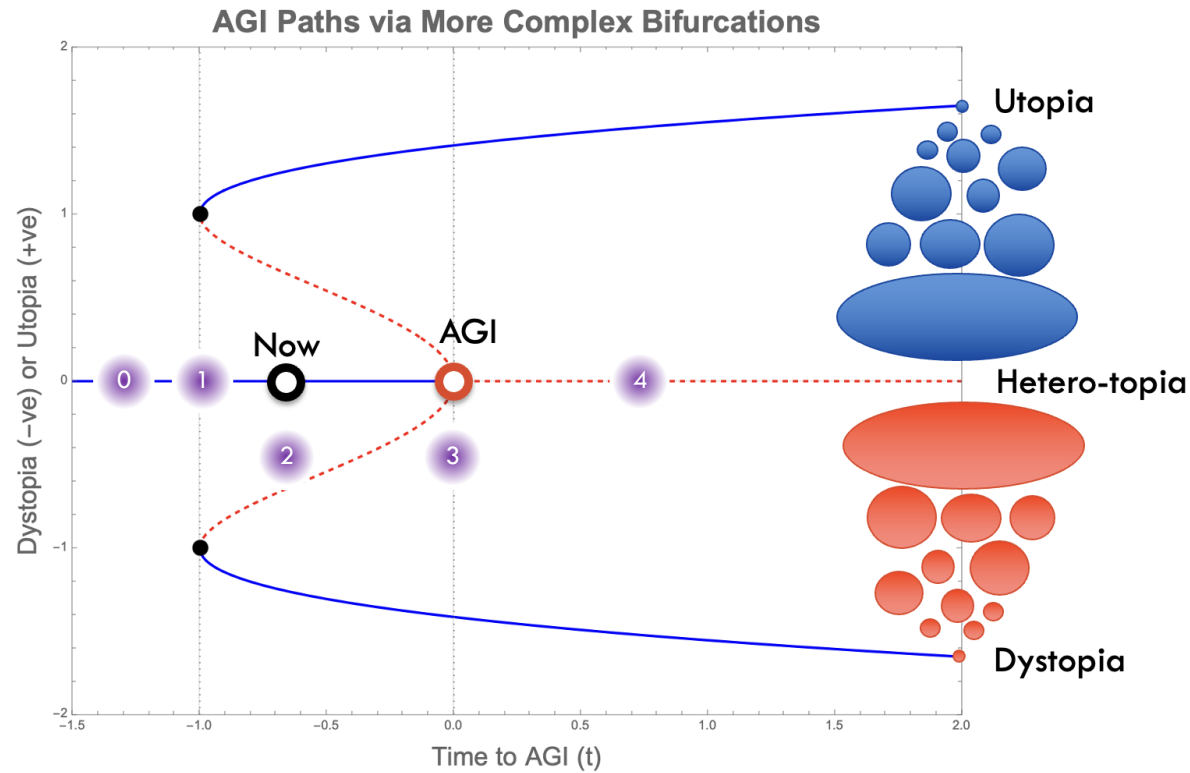
Autonomy Level	Example Systems	Unlocking AGI Level(s)	Example Risks Introduced
Autonomy Level 2: AI as a Consultant <i>AI takes on a substantive role, but only when invoked by a human</i>	Relying on a language model to summarize a set of documents Accelerating computer programming with a code-generating model Consuming most entertainment via a sophisticated recommender system	Possible: Competent Narrow AI Likely: Expert Narrow AI; Emerging AGI	over-trust radicalization targeted manipulation

Tipping Points and Macro-Economics



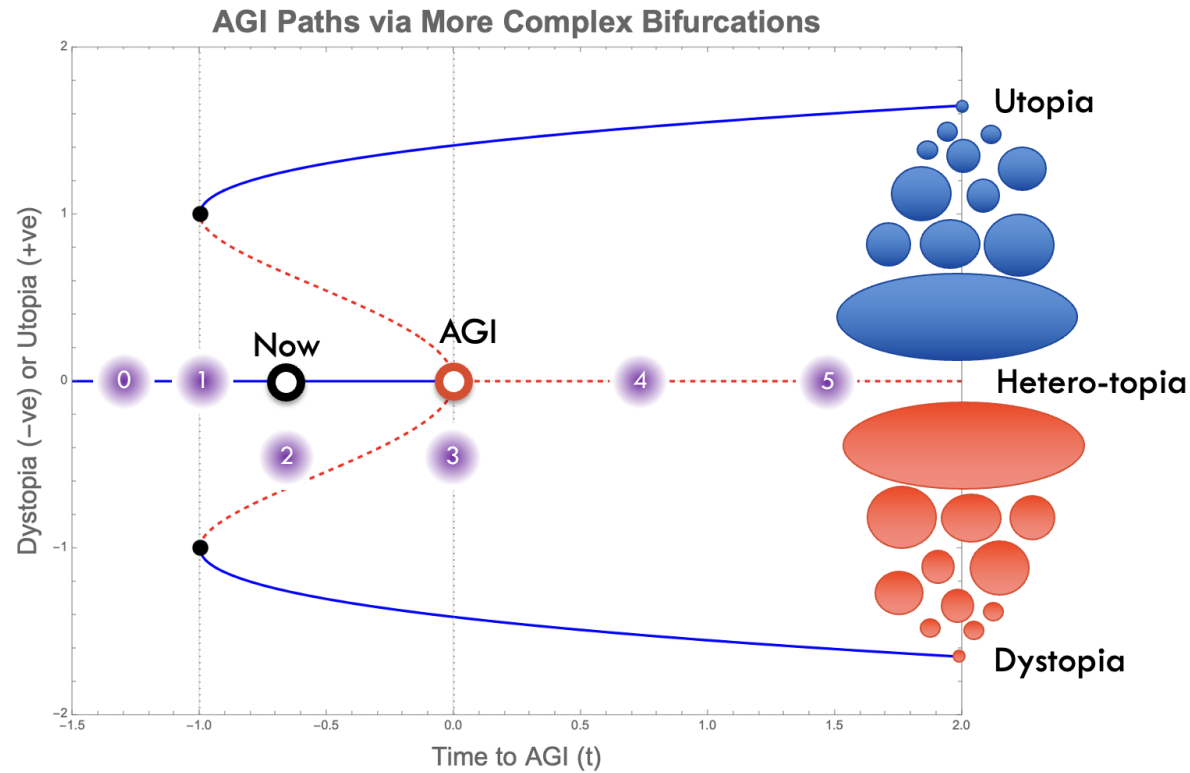
Autonomy Level	Example Systems	Unlocking AGI Level(s)	Example Risks Introduced
Autonomy Level 3: AI as a Collaborator <i>co-equal human-AI collaboration; interactive coordination of goals & tasks</i>	Training as a chess player through interactions with and analysis of a chess-playing AI Entertainment via social interactions with AI-generated personalities	Possible: Emerging AGI Likely: Expert Narrow AI; Competent AGI	anthropomorphization (e.g., parasocial relationships) rapid societal change

Tipping Points and Macro-Economics



Autonomy Level	Example Systems	Unlocking AGI Level(s)	Example Risks Introduced
Autonomy Level 4: AI as an Expert <i>AI drives interaction; human provides guidance & feedback or performs sub-tasks</i>	Using an AI system to advance scientific discovery (e.g., protein-folding)	Possible: Exceptional Narrow AI Likely: Expert AGI	societal-scale ennui mass labor displacement decline of human exceptionalism

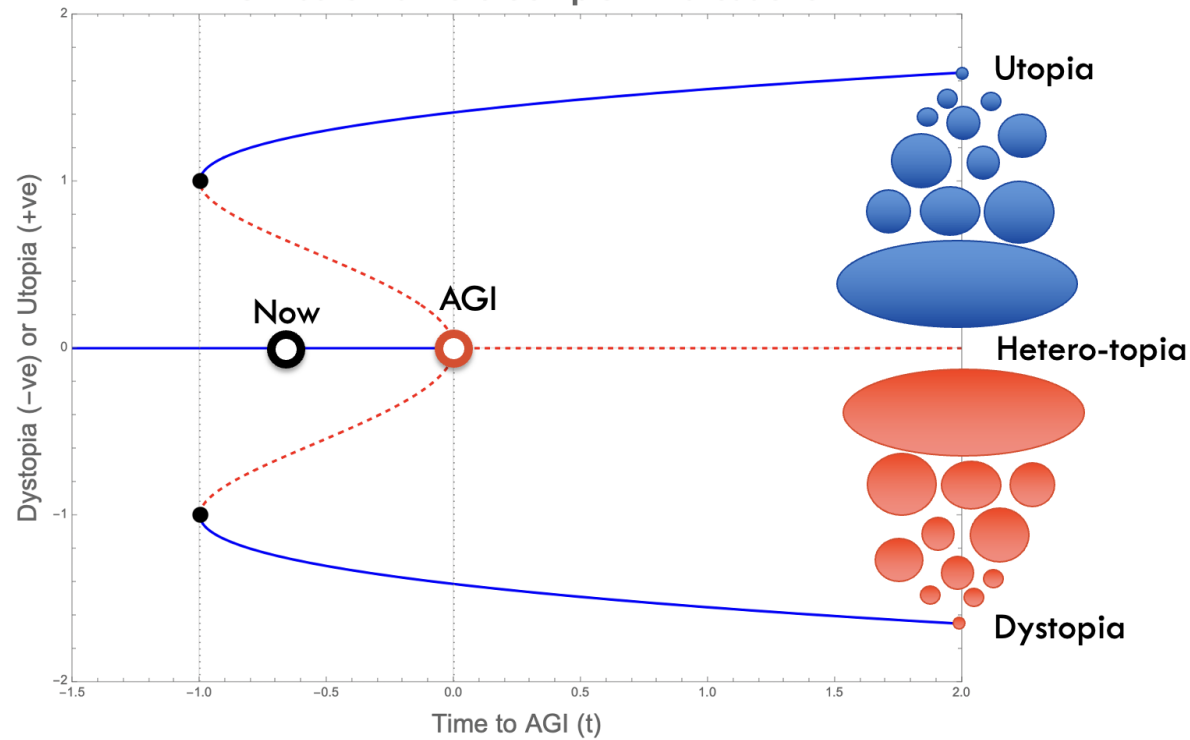
Tipping Points and Macro-Economics



Autonomy Level	Example Systems	Unlocking AGI Level(s)	Example Risks Introduced
Autonomy Level 5: AI as an Agent <i>fully autonomous AI</i>	Autonomous AI-powered personal assistants (<i>not yet unlocked</i>)	Likely: Exceptional AGI; ASI	misalignment concentration of power

Paths Forwards: Where to from here?

AGI Paths via More Complex Bifurcations



1. Where are we now?

[Home](#) > [Data Science for Economics and Finance](#) > Chapter

Massive Data Analytics for Macroeconomic Nowcasting

Peng Cheng, [Laurent Ferrara](#), [Alice Froidevaux](#) & [Thanh-Long Huynh](#)

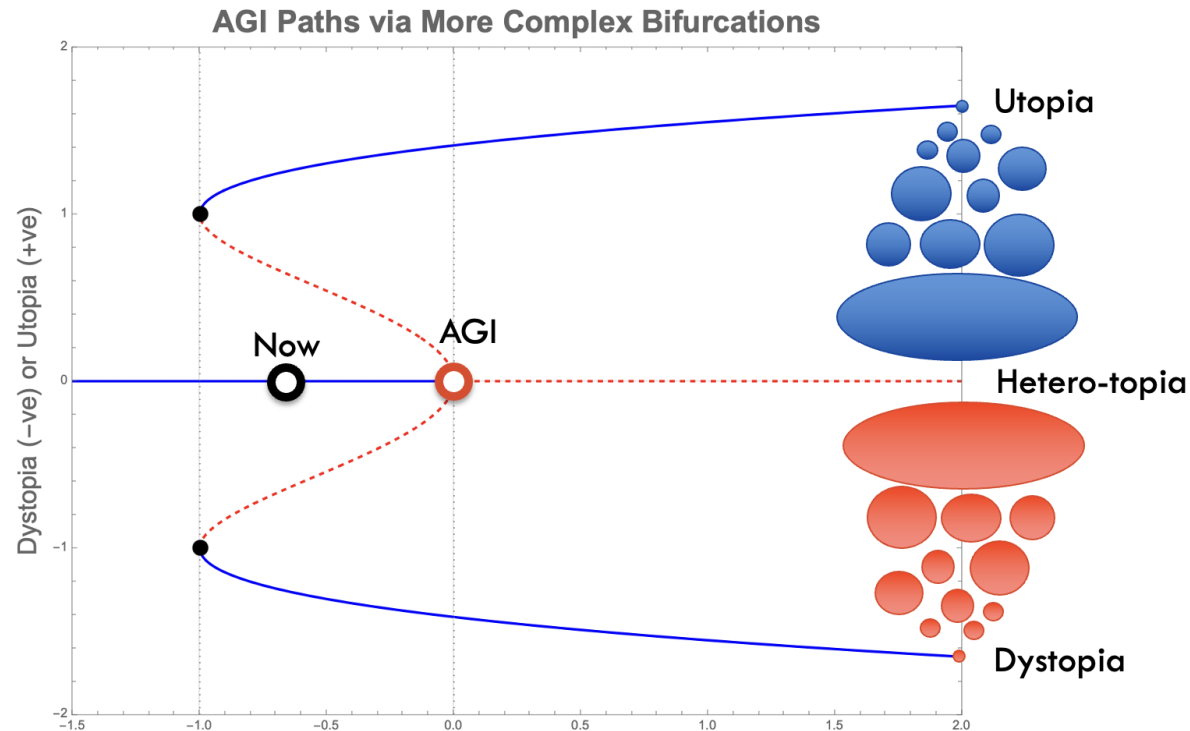
A Grand Challenge: Building a Digital Twin of the Global Economy

[Sina Shahandeh](#) [Follow](#) 7 min read · Sep 24, 2023



...And to make humanity more rational, prosperous and empathetic.

Paths Forwards: Where to from here?



2. Are we at the AGI transition point?

Sensitive intervention points in the post-carbon transition

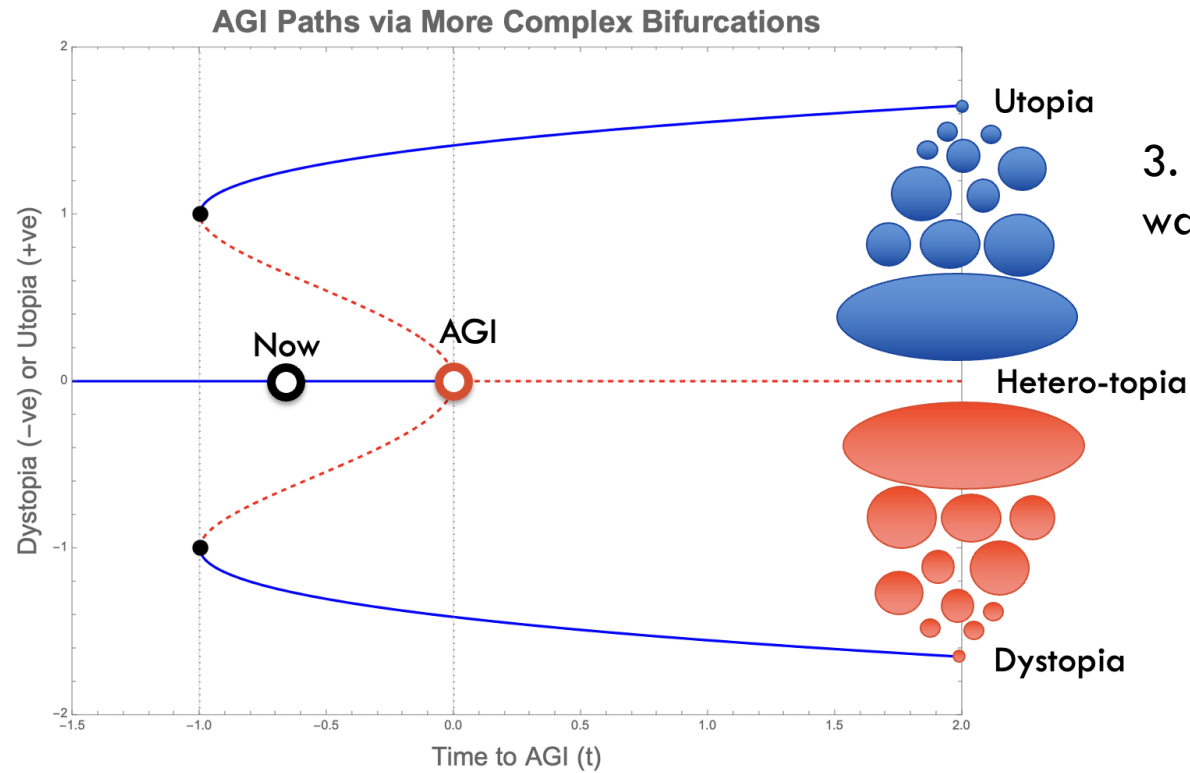
We must exploit socioeconomic tipping points and amplifiers

By **J. D. Farmer^{1,2,3}**, **C. Hepburn^{1,4}**, **M. C. Ives^{1,4}**, **T. Hale⁵**, **T. Wetzter^{1,6,7}**, **P. Mealy^{1,4,8}**,
R. Rafaty¹, **S. Srivastav^{1,4}**, **R. Way^{1,4}**

Deep learning for early warning signals of tipping points

Thomas M. Bury^{a,b}, R. I. Sujith^c, Induja Pavithran^d, Marten Scheffer^e, Timothy M. Lenton^f,
and Chris T. Bauch^{a,1}

Paths Forwards: Where to from here?



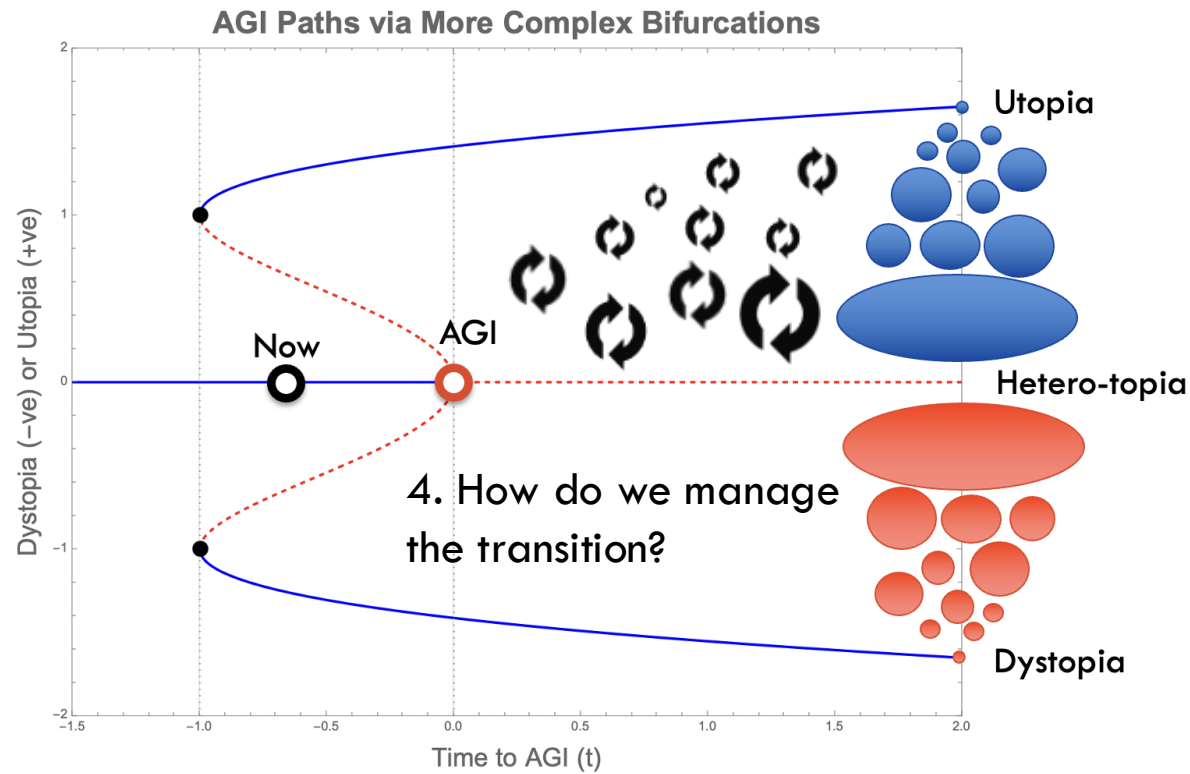
3. Where do we want to be?

The Moral Machine experiment

[Edmond Awad](#), [Sohan Dsouza](#), [Richard Kim](#), [Jonathan Schulz](#), [Joseph Henrich](#), [Azim Shariff](#) ✉, [Jean-François Bonnefon](#) ✉ & [Iyad Rahwan](#) ✉

[Nature](#) **563**, 59–64 (2018) | [Cite this article](#)

Paths Forwards: Where to from here?



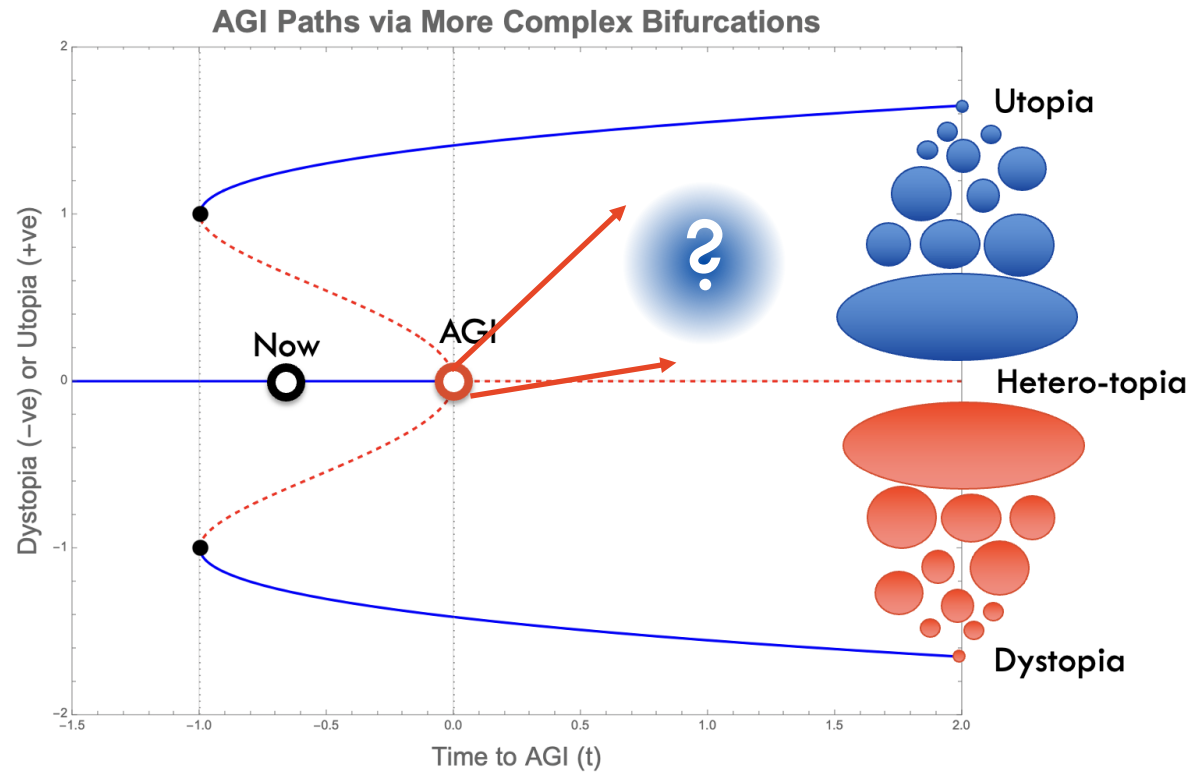
Economic forecasting with an agent-based model

Sebastian Poledna^{a b c g}✉, Michael Gregor Miess^{k e a b h}✉, Cars Hommes^{d i j c}
✉, Katrin Rabitsch^f✉

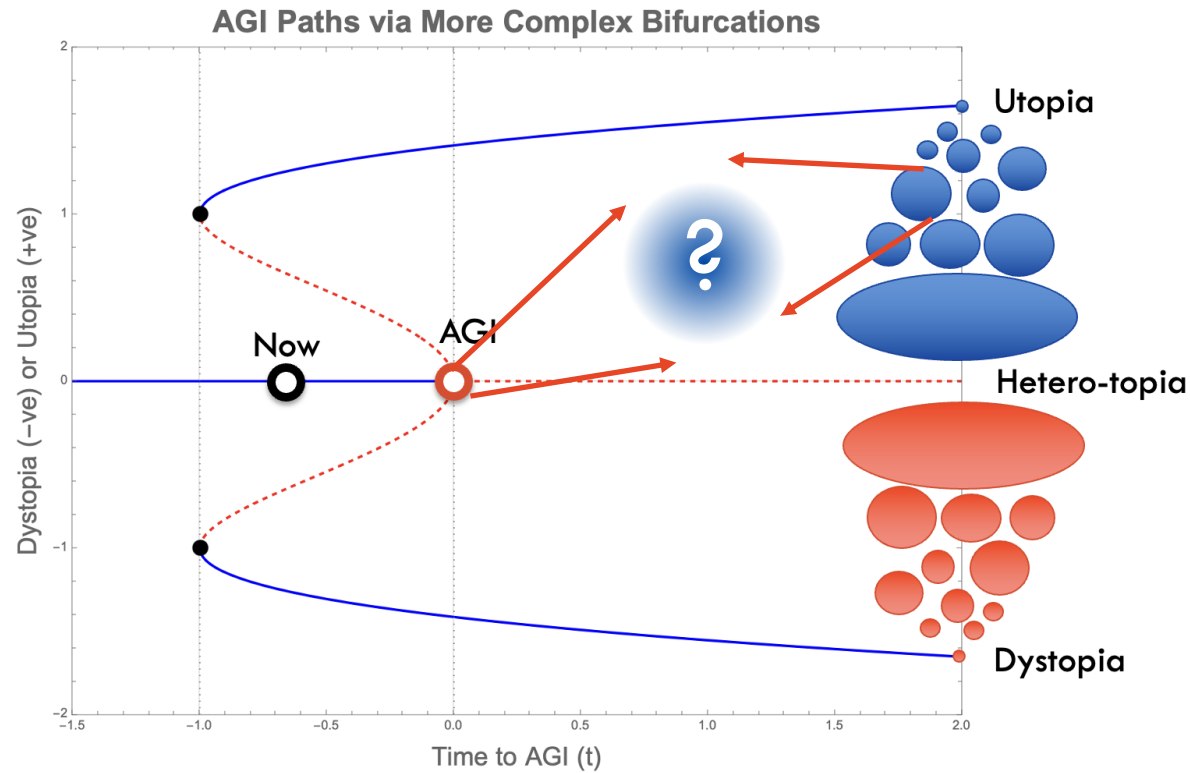
Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach

Jaideep Pathak,^{1,2,*} Brian Hunt,^{3,4} Michelle Girvan,^{1,3,2} Zhixin Lu,^{1,3} and Edward Ott^{1,2,5}

Paths Forwards: Where to from here?



Paths Forwards: Where to from here?



March 18 2025

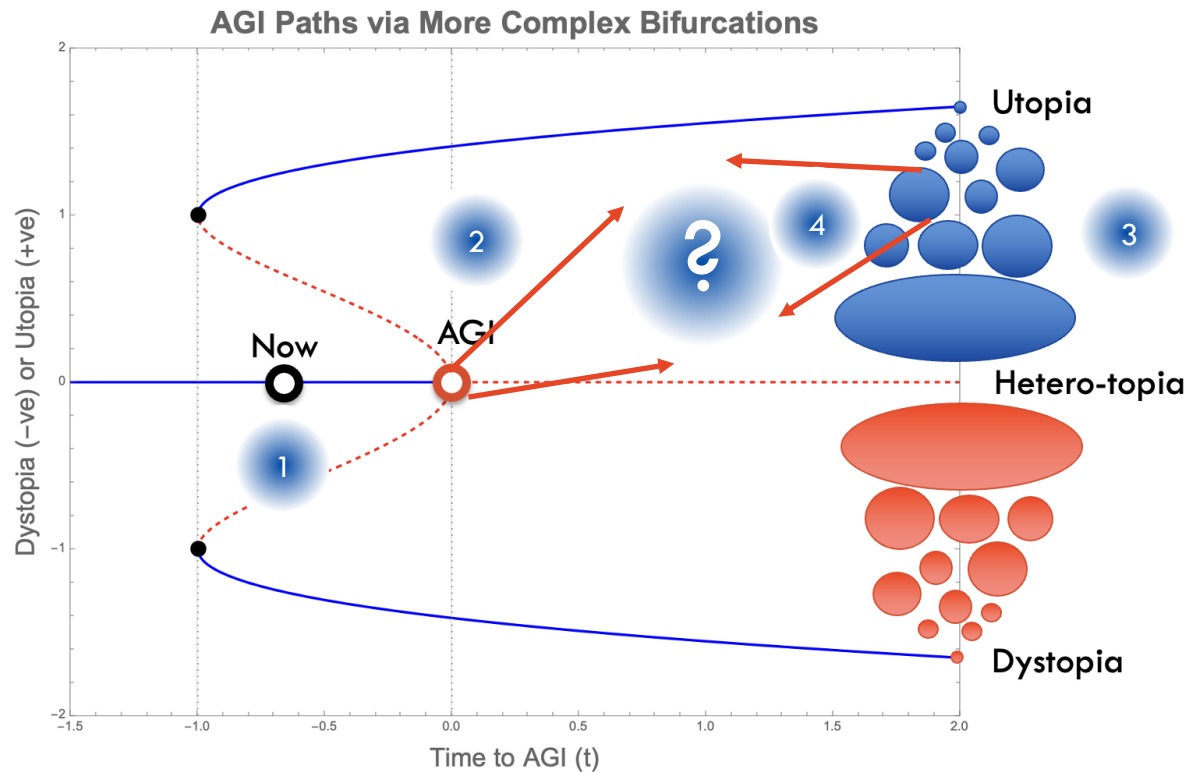
Active Inference and Intentional Behavior

In Special Collection: CogNet

Karl J. Friston, Tommaso Salvatori, Takuya Isomura, Alexander Tschantz, Alex Kiefer, Tim Verbelen, Magnus Koudahl, Aswin Paul, Thomas Parr, Adeel Razi, Brett J. Kagan, Christopher L. Buckley, Maxwell J. D. Ramstead

1. Reactive behaviour: Stimulus response
2. Sentient behaviour: Planning
3. Intentional behaviour: Backward induction

Paths Forwards: Where to from here?



The economy as a “cognitive agent”

1. Now: Generative model of a generative process
2. Planning: Prediction and error correction loops
3. Attractor states: Where we thrive in an economy
4. Intentional behaviour: Backward induction from target attractor

Paths Forwards: Where to from here?

We have some of the theory and tools:

- Agent-Based Models
- Nowcasting, digital models
- Massive computing power
- Huge amounts of data
- Some theoretical ideas for inspiration – starting points

We need:

- Large scale economics: Non-equilibrium economic climate
- Local models: Economic weather forecasting
- Integrated, multi-scale, digital models
- Integrated, centralised, data collection
- Where do we want to be in 20-40 years?

