(Over-Under)Confidence Effects in Knowledge Learning

Daniel Montero Rivas

Faculty of Economics,
University of Cambridge

February 15, 2026

Motivation

"...in the modern world the stupid are cocksure while the intelligent are full of doubt."

– Bertrand Russell

- When does social learning enhance collective knowledge vs. lead societies astray?
- How do metacognitive biases (overconfidence/underconfidence) shape influence?
- Why does misinformation persist despite widespread access to factual information?

Motivating examples

- **Football training**: coach explains a drill → players grasp fragments → confident (but mistaken) players influence unsure teammates → group converges to incorrect pattern.
- **Scottish dance**: non-Scottish participants misinterpret instructor → confident dancers propagate errors → group synchronizes on inaccurate choreography.
- **High-stakes settings**:
  - Diffusion of scientific claims through social media and influencers.
  - Economic policy conversations with sparse or unreliable feedback.
  - Judicial proceedings and hiring decisions.
  - Political discourse where misjudgments of knowledge distort influence.

Key tension: Confidence determines influence, not accuracy

- Overconfident "novices"[1] exert disproportionate influence.
- Knowledgeable but self-doubting agents fail to convince (and may be influentiable).
- Confidence evolves endogenously with learning (not a fixed trait).

[1] Those agents that are not very knowledgeable and skeptical ($\pi_i \approx 0$ for all unknown knowledge).

Key mechanism: Dunning-Kruger effect (Kruger and Dunning (1999))

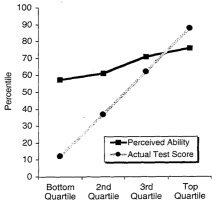


Figure 1. Perceived ability to recognize humor as a function of actual test performance (Study 1).

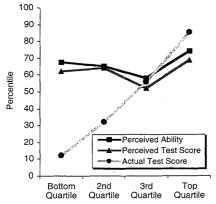


Figure 2. Perceived logical reasoning ability and test performance as a function of actual test performance (Study 2).

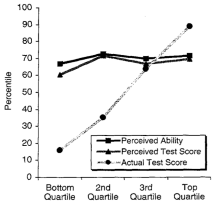


Figure 3. Perceived grammar ability and test performance as a function of actual test performance (Study 3).

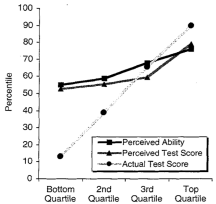


Figure 4. Perceived logical reasoning ability and test performance as a function of actual test performance (Study 4).

Empirical foundations

- **Dunning-Kruger effect**: Jansen et al. (2021), Kruger and Dunning (1999)
- **Non-monotonic pattern**: Knof et al. (2024), McIntosh et al. (2019), Pennycook et al. (2017), Crollen et al. (2011), Sakulku (2011), Clance and Imes (1978)
- **Social implications**: Sulphey and Senan (2025), Feld et al. (2017), Kennedy et al. (2011), Anderson and Kilduff (2009)
- **Overconfidence in economics**: Serra-García and Gneezy (2021), Benoît and Dubra (2011), Clark and Friesen (2009), Sandroni and Squintani (2007), García, Sangiorgi, and Urosevic (2007), Kószegi (2006), Menkhoff, Schmidt, and Brozynski (2006), Fang and Moscarini (2005), Hoelzl and Rustichini (2005), Malmendier and Tate (2005), Van den Steen (2004), Zábojní (2004), Noth and Weber (2003), Camerer and Lovallo (1999)

Theoretical position

- **Sequential social learning**: Bikhchandani et al. (2024), Lobel and Sadler (2015, 2016), Arieli and Mueller-Frank (2014), Goeree et al. (2006), Çelen and Kariv (2004), Smith and Sørensen, (2000), Banerjee (1992), Bikhchandani et al. (1992)
- **Misspecified models**: Arieli et al. (2025), Cho and Libgober (2025), He and Libgober (2025), Ba et al. (2024), Bohren and Hauser (2024), Bohren and Hauser (2021)
- **Statistical Mechanics**: Durlauf (2018), Blume et al. (2011), Durlauf and Ioannides (2010), Brock and Durlauf (2001), Topa (2001), Blume (1993), Ellison (1993), Liggett (1985)

The Model

Setup I

- Finite population of agents $i \in V = \{1, \ldots, N\}$.
- Continuous time $t \in [0, \infty)$.
- Unweighted and undirected network $G = (V, E)$.
- Poisson meetings at constant rate $\lambda_{i,j} > 0$.
- "Facts"/total knowledge indexed by $X \in [0, 1]$.
- Partition total knowledge into P equal-length intervals.
- Binary knowledge: $k_{i,t} : X \to \{0 = \text{ignorant}, 1 = \text{know}\}$.
- Universal state of the world is the truth function: $k^* : X \to \{0 = \text{false}, 1 = \text{true}\}$.

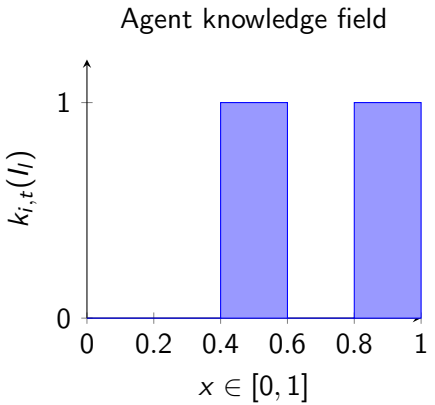
Setup II

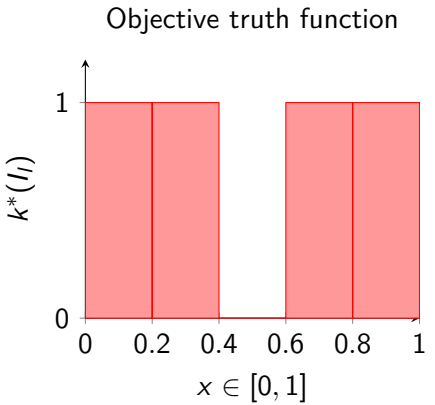
- $\mu_{i,t}(x) = \mathbb{P}_i(\text{fact } x \text{ is true} \mid \text{knowledge/info. at time } t)$ agent i's subjective belief.
- Main assumption: agent i believes everything she knows is true.
  - If $k_{i,t}(x) = 1$, agent believes $k^*(x) = 1$ w.p. 1 ($\mu_{i,t}(x) = 1$).
  - If $k_{i,t}(x) = 0$, $\mu_{i,t}(x) = \pi_i$ where $\pi_i \in [0,1]$: i's prior belief about unknown facts.
- Objective (average) total knowledge: $K_{i,t} = \int_0^1 k_{i,t}(x)\, dx \backsimeq \frac{1}{P}\sum_{l=1}^{P} k_{i,t}(I_l)$.
- Aggregate belief about own knowledge (measure of idiosyncratic understanding of knowledge domain):

$$\tilde{K}_{i,t} = \int_0^1 \mu_{i,t}(x)dx = \int_0^1 [k_{i,t}(x) + (1 - k_{i,t}(x))\pi_i]dx$$

$$\tilde{K}_{i,t} = \frac{1}{P}\sum_{l=1}^{P}[k_{i,t}(I_l) + (1 - k_{i,t}(I_l))\pi_i]$$

Representation





Example I: Addition in the reals

Fact ($x \in X$)	Truth k^*	$k_{i,t}$	$k_{j,t}$	Interpretation
$2+2=4$	1	1	1	Both know the true fact
$2+2=4$	1	1	0	i knows, j ignorant
$2+2=4$	1	0	1	j knows, i ignorant
$2+2=4$	1	0	0	Both ignorant of truth
$2+2=3$	0	1	1	Both hold false knowledge
$2+2=3$	0	1	0	i misinformed, j ignorant
$2+2=3$	0	0	1	j misinformed, i ignorant
$2+2=3$	0	0	0	Both ignorant of false claim

Note: $k^*(x) = 1$ for fact $x =$ "$2+2=4$" by definition; $k^*(x) = 0$ for fact $x =$ "$2+2=3$" by definition.

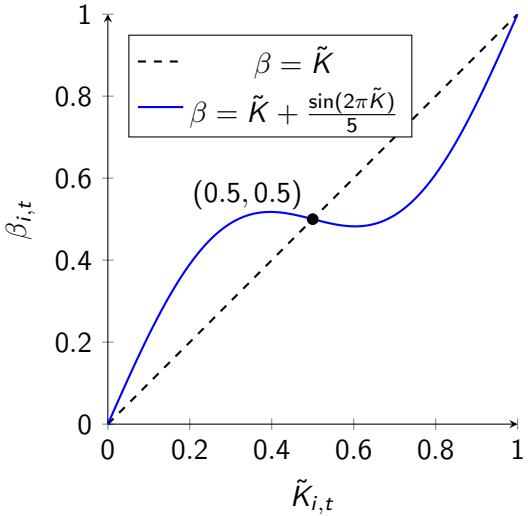
Example IIa: Partitions

- Partitions can be thought of stories or topics in knowledge.
- Even if a partition I_l is such that $k^*(I_l) = 0$, the individual pieces that conform this partition $x \in [\frac{l-1}{P}, \frac{l}{P})$ may be true (but not relevant these individual pieces).
- What people communicate and learn are intervals.

Example IIb: Partitions

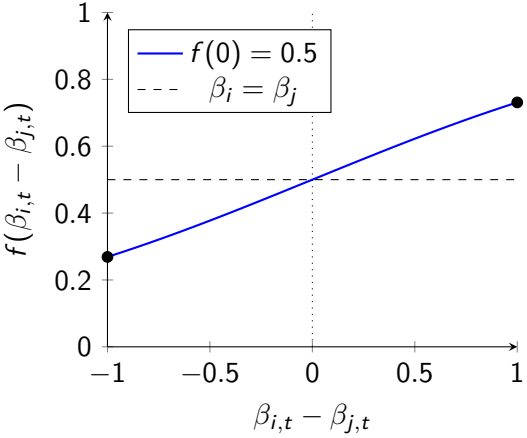
- I_l = "The Earth is flat and governments hide it" : $k^*(I_l) = 0$
  - Earth surface looks flat locally ✓
  - Long distances look flat ✓
  - Governments classify some information ✓
  - Photos can be manipulated ✓
  - Earth is flat ×
  - Space agencies fabricate all satellite imagery ×
- I_l = "String theory" : $k^*(I_l) = 0$
  - Quantum mechanics and general relativity are inconsistent at small scales ✓
  - Strings can mathematically describe particle spectra ✓
  - Extra spatial dimensions emerge in string models ✓
  - These extra dimensions physically exist ×
  - String theory uniquely predicts the Standard Model ×

(Endogenous) subjective perception of knowledge



- $\beta_{i,t} = \tilde{K}_{i,t} + \frac{1}{\alpha} \sin 2\pi \tilde{K}_{i,t}$
- Dunning-Kruger via $\beta_{i,t}$.
- $\alpha > 1.36$ controls miscalibration degree.
- Lower $\alpha \rightarrow$ stronger bias; $\alpha \rightarrow \infty \rightarrow$ perfect calibration.
- Agents unaware of own bias (not strategic, not a choice).
- Degree of bias: $\Delta_{i,t} \equiv |\beta_{i,t} - K_{i,t}|$.
- Agent i's $\beta_{i,t}$ is known by j (not $K_{i,t}$).

(Sigmoid) persuasion



- Probability of knowledge adoption.
- $f(\beta_{i,t} - \beta_{j,t}) = \frac{1}{1+e^{-(\beta_{i,t}-\beta_{j,t})}} \underset{\text{around 0}}{\overset{\text{Taylor approx}}{\approx}} 0.5 + 0.25(\beta_{i,t} - \beta_{j,t})$
- More confident agents are more persuasive.
- Main assumption: agent i takes agent j's subjective assessment of knowledge to be true.
- $f(0) = \frac{1}{2}$: indifference when equal confidence.

Persuasion dynamics: Communication protocol

- Edge (i, j) activates via Poisson.
- Interval I_l drawn uniformly at random.
- If $k_{i,t}(I_l) = 1$ and $k_{j,t}(I_l) = 0$, then

$$k_{j,t+1}(I_l) \leftarrow \begin{cases} 1, & \text{w.p. } f(\beta_{i,t} - \beta_{j,t}) \\ k_{j,t}(I_l), & \text{otherwise} \end{cases}$$

- Sincere transmission and agents interested in knowing.
- Key properties:
  - Irreversibility: $k_{i,t} = 1$ persists (no forgetting).
  - Asymmetry: persuasion only from $0 \to 1$.
  - Unidirectionality: persuasion either $i \to j$ or $j \to i$.

Initialization

- For each $i \in V$ and $l \in \{1, \ldots, P\}$, random initialization such that:

$$k_{i,0}(l_l) \sim \text{Bernouilli}(p_l), \quad \text{i.i.d. across agents}$$

- Assume homogeneous difficulty: $p_l = p$ for all l.
- Same initialization for the truth at $t = 0$.

Environments

- **Benchmark case: Full truth**
  - $k^*(x) = 1, \quad \forall_{x \in X}$
  - Since $K_{i,t} = \int_0^1 k_{i,t}(x)\, dx$ and $K^* = \int_0^1 k^*(x)\, dx$, then full knowledge $K_{i,t} = K^* = 1$ for all i.
- **Sparse truth**
  - $K^*(x) = \mathbf{1}_{x \in T}$ where $T \subseteq [0,1]$
- **Truth revelation: Correction mechanism**
  - w.p. $\varphi \in (0,1)$ Nature reveals $k^*(I_l)$ to one agent chosen uniformly at random from the Poisson-activated edge (i,j).
  - Private revelation.
  - Truth signal reveals only the binary state $k^*(I_l) \in \{0,1\}$, not the semantic content of I_l.
    - **Memory case**
    - **Memoryless case**

Application

Situation: Financial bubbles

In many or most economic applications, a person is not only a passive observer of her environment, but she also chooses actions based on her beliefs.

- Agents communicate and learn following the previous settings and protocols.
- Action is based on the subjective perception of knowledge/information they have:

$$a_i : \beta_i \to \beta_i$$

- In the limit, choose stocks to invest: $a_i^* \in [0,1] = \lim_{t\to\infty} \beta_{i,t} = \beta_{i,\infty}$
- If choices were made for each period: $a_{i,t}^*(\beta_{i,t}) = \beta_{i,t}$.
- The best action is the one that bests approximates the truth (assume agents take the correct action if they are perfectly informed; make full use of information).

Environments

Full truth

Proposition (Almost-sure convergence to full knowledge)

Suppose that

1. $k^*(x) = 1, \forall_{x \in X}$,
2. *Communication intensities satisfy:* $\lambda_{i,j} \geq \lambda > 0, \forall_{i,j,t}$,
3. *Initial knowledge is i.i.d., such that* $k_{i,0}(I_l) \sim$ *Bernoulli*(p) *with* $p > 0$.
4. *For all* $l \in \{1, \ldots, P\}$: $\exists i$ *such that* $k_{i,0}(I_l) = 1$.

Then with probability one, there exists $T < \infty$ *such that*

1. *(Convergence)*
$$K_{i,t} = 1 \quad \forall_{i \in V}, \forall_{t \geq T}.$$
2. *(Stability) the state* $\mathbf{k}^* = (k_{1,T}(x) = 1, \ldots, k_{N,T}(x) = 1, \quad \forall_x)$ *is absorbing.*

Full truth

Remark (Speed of convergence)

Fix the number of agents, communication intensities $\lambda_{i,j} \geq \lambda > 0, \forall_{i,j,t}$, and the initial knowledge distribution. Let T_P denote the (random) time to reach full knowledge when the knowledge domain is partitioned into P intervals. Then:

$$\mathbb{E}[T_P] \quad \text{is weakly increasing in } P$$

Sparse truth

Proposition (Full (correct) knowledge under sparse truth)

Suppose that $K^* < 1$ (sparse truth), and the network becomes connected almost surely. If the initial knowledge distribution satisfies

$$k_{i,0}(I_l) = 0 \quad \text{for all } i \in V \text{ and all } l \text{ such that } k^*(I_l) = 0,$$

then with probability one there exists $T < \infty$ such that

1. *(Convergence)*
$$\mathbb{P}[k_{i,T}(x) = k^*(x), \quad a.e.\ \forall_i] = 1$$
2. *(Stability) the state $\mathbf{k}^* = \{k_{i,T}(I_l) = k^*(I_l),\ \forall_{i,l}\}$ is absorbing.*

Sparse truth

Remark

False knowledge is an absorbing contamination. Under irreversible persuasion, any false belief present initially prevents convergence to full correct knowledge with probability one.

Sparse truth

Theorem (Misinformation Spread Under Overconfidence)

Suppose $K^* < 1$, the network becomes connected a.s., and:

1. *Initial false beliefs: $\exists l \in \{1, \ldots, P\}$ with $k^*(I_l) = 0$ and $\exists i$ with $k_{i,0}(I_l) = 1$.*
2. *Overconfidence concentration: $|O(0)| \geq \theta N$ for some $\theta \in (0, 1)$*
3. *False belief holders are overconfident: If $k_{i,0}(I_l) = 1$ and $k^*(I_l) = 0$, then $i \in O(0)$.*

Then:

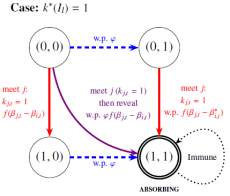
1. ***(Non-convergence)**: $P[\text{system reaches full correct knowledge}] = 0$.*
2. ***(Scaling with partitions)**: The expected number of intervals with false consensus is at least:*

$$\mathbb{E}\left[\sum_{l=1}^{P} \mathbf{1}\{k_i(I_l) \neq k^*(I_l) \text{ for some } i\}\right] \geq (1-\rho)Ph(\theta)$$

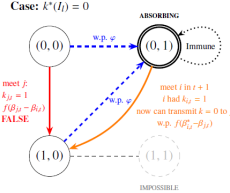
where ρ is truth density and $h(\theta)$ is increasing in θ.

3. ***(Speed of contamination)**: If $\theta > 1/2$, false beliefs spread faster than true beliefs: the expected time for false consensus on an interval with $k^*(I_l) = 0$ is shorter than for true consensus on an interval with $k^*(I_l) = 1$.*

Correction mechanism: Memory



(a) *Absorbing state is* $(1, 1)$



(b) *Absorbing state is* $(0, 1)$

- State space for each agent i and interval l be: $(k_{i,t}(I_l), v_{i,t}(I_l)) \in \{0, 1\}^2$.
- $K^T_{i,t} \equiv \frac{1}{P} \sum_{l=1}^{P} |\{l : k_{i,t}(I_l) = k^*(I_l)\}|$
- $K^E_{i,t} \equiv \frac{1}{P} \sum_{l=1}^{P} \max\{k_{i,t}(I_l), v_{i,t}(I_l)\}$
- Now: $\beta_{i,t} = g(K^E_{i,t})$
- $K^E_{i,t} \sim$ posterior belief (update from knowledge acquisition and verification).
- Revelations $v_{i,t}(I_l)$ are only from Nature; agents do not transmit verification, only knowledge.

Correction mechanism: Memory

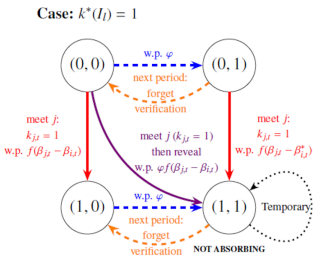
Proposition (Almost-Sure Convergence with Truth Revelation and Memory)

Suppose truth revelation occurs with probability $\varphi > 0$, agents have perfect memory, and the network becomes connected a.s. Then:

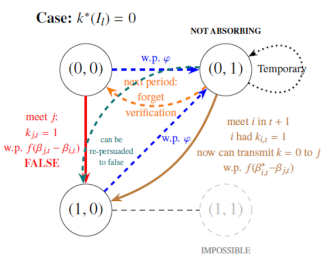
$$\mathbb{P}\left[\exists T < \infty : (k_i(I_l), v_i(I_l)) = \begin{cases} (1,1) & \text{if } k^*(I_l) = 1 \\ (0,1) & \text{if } k^*(I_l) = 0 \end{cases} \quad \forall_{i,l} \text{ and } \forall_{t \geq T}\right] = 1$$

That is, with probability one, the system converges to full correct knowledge.

Correction mechanism: Memoryless



(a) *No absorbing state (verification forgotten)*



(b) *No absorbing state (can be re-persuaded)*

Correction mechanism: Memoryless

Proposition (Failure of Convergence without Memory)

Suppose truth revelation occurs with probability $\varphi > 0$, but agents are memoryless. Then full correct knowledge is not reached almost surely. Instead, the system converges to a stationary distribution π where:

$$\pi(\textit{all agents correct on all intervals}) < 1$$

Extensions

Next steps

- **Misspecified models** (not misspecification of the environment, BUT misspecification of ones own epistemic state):
  - True Data Generating Process (DGP):
  - Utility function:
  - Belief distribution:
  - Self-confirming equilibrium: beliefs are confirmed by data along the equilibrium path
- **Statistical Mechanics**:
  - Lattice network
  - Infinitesimal generator: knowledge diffusion process $\{\eta_t\}_t \geq 0$ is a continuous-time Markov chain on X such that

$$\Omega\phi(\eta) = \sum_{\eta' \in X} c(\eta, \eta')[\phi(\eta') - \phi(\eta)]$$

Thank you for listening !

Daniel Montero Rivas