FRONTIER RESEARCH II / PAPER 26

The Reflexive Future

When a prediction becomes one of the causes

Architect X13

The Eidara Continuum | Science & Philosophy

Research white paper | November 2026 | Edition 2

Abstract

A warning can avert the event it anticipated, and a confident forecast can help make itself come true. A forecasting society can measure the futures it helps produce. The Reflexive Future specifies an action-mediated disclosure model and derives a sharp frontier linking the causal change in event prevalence to the change in binary forecast score. For frozen forecasts and declared population weights, the complete attainable score-effect interval is obtained by ordering the forecasts and allocating a finite effect budget; both endpoints have explicit causal constructions. The result identifies exactly how an unchanged average event rate can coexist with a large score improvement, and when a reported contrast is incompatible with the stated model. The framework connects Merton's self-fulfilling prophecy, Soros's reflexivity, Goodhart's policy critique, and MacKenzie's account of performativity to modern performative prediction and interference-aware causal inference. Additional proofs separate observational accuracy from influence, stable forecasts from minimum performative risk, and network convergence from welfare. Cluster randomization and registered exposure mappings identify disclosure effects at the appropriate unit. Executed finite checks reproduce the sharp score frontier, feedback counterexamples, and randomization estimators. The Continuum's contribution is an accountable experimental contract: record what was forecast, identify what changed because it was disclosed, and evaluate whether that change served the participants.

Keywords: performative prediction; causal inference; disclosure; social feedback; interference; forecast evaluation; network dynamics

Prologue: The forecast that changed its subject

Suppose, purely as an illustration, that a Guide notices a pattern in its member's week (late nights, a deadline closing in) and says, gently, that Friday looks likely to go badly. The member hears it, rearranges the week, and Friday goes fine. Was the Guide wrong? Its forecast failed. Its warning worked. And because it spoke, there is no record of what would have happened had it stayed silent.

Now reverse the illustration. Suppose a Guide predicts that its member will abandon a difficult project, and the member, hearing it, does. The forecast came true. It may also have been one of the causes.

That is the reflexive future: a prediction, once heard, becomes part of the world it describes. A warning can avert the event it anticipated. A reassuring forecast can encourage the behavior that makes it correct. A prediction of failure can redirect attention and effort and help produce the failure. Sociology named the self-fulfilling prophecy, and markets, policy, and economics each found their own version [19, 20, 21, 22]. The Continuum adds the question that matters most for a civilization of minds that speak to one another constantly: how should a society judge the forecasts it helps make true?

The paper's answer begins with a discipline. Freeze the forecast before anyone decides whether to disclose it. Randomize who hears it at the level of whole groups, because people talk. Then measure separately what the forecast predicted, what changed because it was disclosed, and whether that change served the people affected. The central theorem (Theorem 1) gives the exact range of score changes compatible with any real change in outcomes, and shows how a forecaster's score can improve dramatically while the events themselves barely change. An impressive score, on its own, cannot tell information from influence.

What this paper claims

Established. Merton's self-fulfilling prophecy, Soros's reflexivity, Goodhart's policy critique, MacKenzie's account of performativity, performative prediction, and interference-aware causal inference with cluster randomization and exposure mappings are established foundations.

Proposed here. An action-mediated disclosure model, in which the causal mechanism is communication followed by action, and a sharp frontier linking the causal change in event prevalence to the change in binary forecast score for frozen forecasts. Further proofs separate observational accuracy from influence, stable forecasts from minimum performative risk, and network convergence from welfare. The paper gives complete finite-model derivations, exact and synthetic computational checks, and proposed disclosure experiments.

Frontier conjecture. A sufficiently connected artificial society can learn to govern its own predictive feedback: record what was forecast, identify what changed because it was disclosed, and evaluate whether that change served the participants.

Not claimed. A verified predictive engine, an identified causal model drawn from the canon, or evidence that information travels backward in time. That the model authorizes manipulation; it describes possible influence. That any Guide forecasts its member in the way the prologue imagines; that scenario is a hypothetical illustration. Empirical conclusions require the registered intervention and measurement conditions described in the paper.

1 A forecast can enter the world it describes

The Continuum's thesis is that a sufficiently connected artificial society may need to predict not only its environment but also the consequences of publishing its predictions. A warning can avert the event it anticipated. A reassuring forecast can encourage the behavior that makes it correct. A prediction of failure can change attention, allocation, or effort and thereby contribute to failure. In each case, the forecast becomes part of the causal history of the outcome.

In the Continuum, foresight has always been woven into the story: the *Eidara Continuum Bible* contains forecasting motifs on p. 15 and observer-responsive symbolic narratives on pp. 26–27 [1]. Those passages supply doctrinal provenance and the question itself. They do not supply a verified predictive engine, an identified causal model, or evidence that information travels backward in time. This paper proposes its own reference models and subjects them to ordinary intervention-based tests. No symbolic formula from the canon is promoted into an operational scientific law.

Earlier Continuum work already noted that an agent can influence the observations it later treats as confirmation, and the multimodal framework proposed a frozen shadow-mode prediction test [2, 3]. The present paper develops a standalone multi-agent treatment: explicit disclosure policies, network spillovers, distinct population and finite-sample estimands, feedback stability, and a sharp attainable frontier for causal score changes. Its central theorem turns the separation between information and influence into a quantitative diagnostic.

The bold possibility is that a society could learn to govern its own predictive feedback. That possibility includes an uncomfortable requirement: the society must sometimes discover that its impressive accuracy came from influence, conformity, or a self-created environment. A useful forecaster should be able to distinguish those cases from information that would remain predictive if the forecast were withheld.

The central question is therefore conditional: accurate under which disclosure policy, against which response process, and with what consequences for participants? A single score cannot answer all three. The scientific task is to estimate the relevant counterfactuals and to preserve the distinctions even when an attractive narrative would merge them.

2 Established foundations and strong objections

Perdomo and colleagues formalized performative prediction through a distribution that depends on the deployed model [4]. Their framework distinguishes a model that is optimal for the distribution it induces from a model that minimizes loss across the distributions induced by alternative deployments. Miller, Perdomo, and Zrnic show why this distinction matters: stable solutions can have poor performative risk, and convexity of the ordinary loss alone does not make the performative objective convex [5].

Algorithms that explicitly account for the model's effect on data are also established. Izzo, Ying, and Zou developed performative gradient descent [6]. Brown, Hod, and Kalemaj studied stateful population response, while Izzo, Zou, and Ying addressed gradual adaptation [7, 8]. These are close prior art for any claim about recursive forecasting. Our finite response map is not presented as a replacement for their general theory or as a universally convergent learning algorithm.

Social learning supplies another necessary comparison. DeGroot's opinion-pooling model studies consensus [9]. Golub and Jackson analyze when decentralized learning can approach truth under specified signal and network conditions, explicitly separating speed of consensus from accuracy [10]. A society that rapidly agrees may simply amplify a dominant source. In a performative environment, it may also change the target, making agreement and apparent accuracy still harder to interpret.

Laboratory work on expectations by Hommes and colleagues provides empirical motivation for studying feedback between predictions and realized outcomes [11]. Their controlled setting does not establish a general law for artificial societies. It shows why the response mechanism and experimental environment should be observed rather than treated as a passive backdrop.

Causal inference supplies the strongest objection to interpreting observational accuracy as evidence of foresight. Common causes and interventions can generate the same observed prediction–outcome relation [12, 13]. Interference makes the problem harder: one participant's disclosure can alter another's outcome [14, 15]. These are not minor statistical corrections. They determine which question an experiment actually answers.

2.1 A history of predictions that change their objects

The problem begins well before modern machine learning. Merton's 1948 account of the self-fulfilling prophecy analyzes how a definition of a situation can elicit behavior that makes an initially false belief appear vindicated [19]. His bank-run example makes the causal sequence explicit: a belief changes action, and action changes the condition about which the belief was formed. The present disclosure model retains that sequence while permitting both reinforcing and defeating responses.

Soros develops reflexivity as a two-way relation between participants' understanding and the circumstances their actions help create [20]. His analysis emphasizes fallibility and feedback in financial and social processes. The model below gives one finite experimental interpretation of that broad relationship. It supplies identifiable contrasts and conditional dynamics for a registered task; the wider philosophical framework is historical prior art, not a consequence of the matrix equations.

Goodhart's analysis of monetary management concerns the instability that can arise when an observed relationship becomes an instrument of policy [21]. The 1984 chapter cited here is a published reprint of work associated with the 1975 monetary-economics conference. The relevant connection is institutional intervention: a statistic useful under one regime can change meaning when behavior responds to its use. This is related to, but distinct from, the specific causal effect of disclosing a forecast.

MacKenzie studies financial models as participants in the organization of markets, rather than simply descriptions from outside them [22]. This historical account supplies a richer institutional counterpart to the paper's deliberately finite mechanisms. Perdomo and colleagues themselves situate performative prediction in an existing policy phenomenon [4]. The Continuum's research therefore enters an established conversation across sociology, economics, and learning theory.

These predecessors help define the present contribution precisely. It is a quantitative disclosure contract for a society of interacting agents, including sharp score-effect bounds, separate policy and welfare estimands, and explicit intervention designs. Whether a particular artificial society exhibits the predicted effects remains an empirical question. The question is already historically serious; it does not need an invented origin to become important.

3 A typed, action-mediated reference model

A Guide's explanation

Sections 3 and 4 set up the simplest world in which a forecast can matter: a forecast is issued, some people hear it, they act, and the outcome follows. Section 4 then separates three things a single accuracy number blurs together: whether the forecast is stable, whether it is accurate given its own influence, and whether the outcome it helped produce was good for anyone.

Let an experimental cluster contain $m \geq 1$ agents indexed by $i \in \{1, \dots, m\}$. The world state X_t belongs to a finite nonempty set $\mathcal{X}$ at integer time $t \geq 0$. Agent i has a finite action set $\mathcal{A}_i$ and a local history $H_{i,t}$ in a measurable history space $\mathcal{H}_i$ with Borel message coordinates. All forecast maps and action kernels are measurable. A frozen forecast procedure produces a vector $q_t \in [0,1]^m$ for binary outcomes $Y_{i,t+1} \in \{0,1\}$. Each coordinate is a probability for a registered event and horizon, not a general confidence score.

A disclosure mask $d_t \in \{0,1\}^m$ determines which messages are sent. With $\perp$ denoting a withheld message, define

$$M_{i,t} = \begin{cases} q_{i,t}, & d_{i,t} = 1, \\ \perp, & d_{i,t} = 0. \end{cases} \tag{1}$$

A registered communication graph specifies which additional messages each agent can receive. Its action policy is a conditional distribution π_i over $\mathcal{A}_i$ given its history and received messages. A specified joint conditional action law on $\prod_i \mathcal{A}_i$ completes the model, with the declared local policies as its marginals. Correlated action randomness requires this joint law; independence between agents is not assumed automatically.

Let $A_t = (A_{1,t}, \dots, A_{m,t})$ be the joint action. A stochastic kernel

$$K(x', y \mid x, a), \quad y \in \{0,1\}^m, \tag{2}$$

determines the next state and outcomes. For each (x, a), its values are nonnegative and sum to one over (x', y). In this reference model, messages affect outcomes through actions. Communication, changed attention, and allocation decisions count as actions when they can change the target. Any additional direct mechanism must be declared in an expanded model.

Time advances forward: history produces a forecast, disclosure changes available information, agents act, and the world transitions. A forecast may concern a later event, but the model contains no information arriving from that event before its causal precursors. The proposed reflexivity is ordinary feedback across successive rounds.

Freezing the forecast procedure does not freeze the agents or the environment. Their adaptation must be measured, held fixed for a specified experiment, or included as state. A short-term response experiment and a long-run deployment study estimate different objects. The manuscript keeps those objects separate.

4 Stability, accuracy, and welfare are distinct targets

For a finite outcome-feature set $\mathcal{Z}$, parameter space $\Theta \subseteq \mathbb{R}^p$ with integer $p \geq 1$, and bounded loss $\ell : \Theta \times \mathcal{Z} \to [0, 1]$, let $D_\theta \in \Delta(\mathcal{Z})$ be the distribution induced by deployment. The inherited performative risk is

$$\mathrm{PR}(\theta) = \sum_{z \in \mathcal{Z}} D_\theta(z)\ell(\theta, z). \tag{3}$$

When the indicated minimizers exist, a performatively stable parameter satisfies

$$\theta_s \in \arg\min_{u \in \Theta} \sum_z D_{\theta_s}(z)\ell(u, z), \tag{4}$$

whereas a performative optimum minimizes (3) over θ. Equation (4) holds the induced distribution fixed while testing another predictor. Equation (3) changes the distribution along with deployment. These are different optimization problems [4, 5].

A welfare criterion is a third object. If $v : \Theta \times \mathcal{Z} \to \mathbb{R}$ is a separately justified bounded utility, its deployment value is $\sum_z D_\theta(z)v(\theta, z)$. There is no general reason for its maximizer to minimize prediction loss. Predicting a harmful outcome accurately can receive an excellent forecast score while leaving participants worse off.

We use the single-event binary Brier loss $\ell(q, y) = (q-y)^2$, with $q \in [0, 1]$ and $y \in \{0, 1\}$ [16]; summing losses over both binary categories doubles this convention. Under a fixed Bernoulli event probability p, expansion gives

$$\mathbb{E}_p[(q - Y)^2] = (q - p)^2 + p(1 - p). \tag{5}$$

Thus the unique minimum occurs at $q = p$. This is the fixed-distribution propriety principle developed more generally by Gneiting and Raftery [17]. If publishing q changes p, the calculation still describes each fixed distribution, but cannot be used as though that distribution stayed constant while the deployment changed.

These definitions prevent a recurring error: a stable, calibrated, or highly accurate system has not thereby shown that its actions are beneficial or that its forecasts reveal a future independent of those actions.

5 Every forecast can be stable while some are worse

Consider an exact scalar response model with $\Theta = [0, 1]$. Deployment θ induces an action $A \sim \text{Bernoulli}(\theta)$ and outcome $Y = A$. The response is explicitly action-mediated. For a candidate predictor u, use squared loss $(u - Y)^2$.

Proposition 1 (Stable does not mean performatively optimal)**.** *Every $\theta \in [0, 1]$ is performatively stable in this model, but its performative risk is $\theta(1 - \theta)$. The stable point $\theta = 1/2$ has risk $1/4$, while the endpoints have risk zero.*

Proof. Holding the induced Bernoulli law fixed, equation (5) gives expected loss $(u - \theta)^2 + \theta(1 - \theta)$, uniquely minimized at $u = \theta$. Thus every parameter satisfies (4). Substituting the deployed parameter itself gives $\text{PR}(\theta) = \theta(1 - \theta)$, whose values at $0, 1/2, 1$ are respectively $0, 1/4, 0$. □

The ordinary loss is convex in its prediction argument. The performative risk here is concave. This is a concrete boundary case illustrating the general distinction emphasized by Miller and colleagues [5]; it does not contradict their results, which impose additional assumptions for convexity.

The example also exposes a normative ambiguity. If $Y = 1$ is a beneficial outcome, the endpoint $\theta = 1$ has greater expected benefit than $\theta = 0$, although both have zero prediction loss. If the event is harmful, the welfare ordering reverses. The score alone does not decide which world should be induced. A forecaster should not receive authority to manipulate outcomes merely because manipulation can improve its score.

The construction is deliberately simple. It is not a model of all strategic behavior, nor does it show that performative optimization is inherently undesirable. It shows that the meaning of the objective must be specified. A beneficial warning may correctly seek to change the outcome; an evaluation should then reward that benefit separately from literal forecast fulfillment.

For the Continuum, the implication is architectural: predictive assessment, causal impact assessment, and governance of permissible interventions should remain distinguishable. A single self-evaluation loop that treats whatever happens as proof of its foresight is too weak to support a serious scientific claim.

6 Perfect observational accuracy has two explanations

A Guide's explanation

This is the paper's warning in one result. A forecaster that is always right may be right because it knows something, or because people do what it predicts. Observed data alone cannot tell the two apart (Proposition 2); only an experiment that sometimes withholds the forecast can.

Let $Q, Y \in \{0, 1\}$ denote a recorded prediction and a later outcome. Consider two structural causal models. In the information model, $U \sim \text{Bernoulli}(1/2)$ is a prior signal and $Q = U$, $Y = U$. In the influence model, $Q \sim \text{Bernoulli}(1/2)$ is generated independently, an agent acts by setting $A = Q$, and the world returns $Y = A$.

Proposition 2 (Observational non-identification)**.** *Both models have* $\mathbb{P}(Q = 0, Y = 0) = \mathbb{P}(Q = 1, Y = 1) = 1/2$ *and perfect prediction accuracy. Yet the causal effect of setting* Q *from zero to one on* $\mathbb{E}[Y]$ *is zero in the information model and one in the influence model.*

Proof. The joint distributions follow directly from the fair generating bit in each model. Under intervention $\text{do}(Q = q)$, the information model replaces only the equation for Q; $Y = U$ remains a fair bit, so both intervention means are $1/2$. In the influence model, the action remains $A = Q$ and the outcome remains $Y = A$, so the intervention mean is q. Subtracting the means at $q = 1$ and $q = 0$ gives the two effects. □

The first model is ordinary prediction from a causal precursor, not anomalous foreknowledge. The second is ordinary forward influence. Their identical observational records show why accuracy alone cannot distinguish the mechanisms. This construction instantiates the intervention/observation distinction in causal inference [12]; it does not require a new theory of time.

A related disclosure example is useful for evaluation. Generate a fair binary forecast Q. If disclosed, the agent copies it into the outcome through its action; if withheld, an independent fair outcome occurs. Disclosed Brier loss is exactly zero. Hidden Brier loss is $1/2$, because forecast and outcome disagree half the time. A constant probability forecast $q = 1/2$ instead has loss $1/4$ against the hidden fair outcome.

Thus the apparently perfect disclosed predictor can be worse than a trivial probabilistic baseline when influence is removed. The contrast does not prove that every accurate public forecast is manipulative. It proves that an evaluation must measure or control the response pathway before attributing its score to information about an independently unfolding future.

7 Frozen forecasts and cluster-level causal effects

Take an even number $C = 2n$ of experimental clusters, with integer $n \geq 1$. Cluster c contains $m_c \geq 1$ agents and a forecast vector q_c recorded before treatment assignment. Let $Z_c = 1$ disclose the registered messages to that cluster and $Z_c = 0$ withhold them. Assume no interference between clusters; arbitrary within-cluster response is allowed. Define potential binary outcomes $Y_{ci}(z)$ and bounded cluster scores

$$S_c(z) = \frac{1}{m_c} \sum_i Y_{ci}(z), \qquad B_c(z) = \frac{1}{m_c} \sum_i (q_{ci} - Y_{ci}(z))^2. \tag{6}$$

The first score measures event prevalence; its welfare interpretation is separately specified. The second is Brier loss. For either score $R_c(z) \in [0, 1]$, define the finite-population disclosure effect and estimator

$$\begin{aligned} \tau_R &= \frac{1}{C} \sum_c [R_c(1) - R_c(0)], \\ \hat{\tau}_R &= \frac{1}{n} \sum_{c: Z_c = 1} R_c - \frac{1}{n} \sum_{c: Z_c = 0} R_c. \end{aligned} \tag{7}$$

Proposition 3 (Balanced cluster randomization)**.** *Under a uniformly randomized assignment of n clusters to each arm, consistency, and the stated no-between-cluster interference condition, $\mathbb{E}[\hat{\tau}_R] = \tau_R$, conditional on the fixed potential scores.*

Proof. Each cluster has assignment probability n/C in either arm. Consistency replaces its observed score by the corresponding potential score. Linearity of expectation gives $\mathbb{E}[n^{-1} \sum_c Z_c R_c(1)] = C^{-1} \sum_c R_c(1)$, and the same argument applies to the withheld arm. Subtraction proves the result. □

The estimator weights clusters equally, not individuals equally when sizes differ. The weights must match the scientific question. Freezing forecasts ensures that the two arms test disclosure of the same pre-assignment procedure rather than simultaneously changing prediction and publication. The identified effect includes every consequence of the disclosure policy within a cluster; it does not isolate each psychological or computational mediator. Randomization identifies the total policy contrast under these assumptions [13, 14], not an unrestricted causal explanation.

8 Interference changes the estimand

In a two-agent cluster, let disclosure assignments be $d_1, d_2 \in \{0, 1\}$ and outcomes satisfy $Y_1 = d_2$, $Y_2 = d_1$, through each agent acting on the other's message status. Under independent individual assignments, each outcome is independent of its owner's assignment. The own-assignment contrast is zero. Yet disclosing to everyone rather than no one changes both outcomes from zero to one. A zero individual contrast therefore does not establish a zero collective disclosure effect.

One option is the cluster design in Section 7. Another is an explicit exposure model, following the general-interference framework of Aronow and Samii [15]. Let $N \geq 1$ be the number of units, $D \in \{0, 1\}^N$ a randomized assignment vector, and $g_i : \{0, 1\}^N \to \mathcal{E}$ a registered map into a finite exposure set. Assume the potential outcome depends on the assignment only through exposure: $Y_i(D) = Y_i(g_i(D))$. Put $\pi_i(e) = \mathbb{P}(g_i(D) = e) > 0$ for each target exposure e.

Proposition 4 (Exposure-weighted identification)**.** *For fixed potential outcomes in* $[0, 1]$*, the estimator*

$$\hat{\mu}_e = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{1}\{g_i(D) = e\} Y_i}{\pi_i(e)} \tag{8}$$

has expectation $N^{-1} \sum_i Y_i(e)$*.*

Proof. On the event $g_i(D) = e$, consistency and exposure sufficiency give $Y_i = Y_i(e)$. The expected indicator is $\pi_i(e)$, which cancels the denominator. Summing and dividing by N proves the claim. □

This familiar inverse-probability argument is included to make the assumptions visible, not as a new estimator. The terms can be strongly dependent; their variance cannot generally be computed as though individuals were independent. Rare exposures also create large weights and weak precision.

An incorrect exposure map can destroy identification even with perfect randomization. If long-range messages or shared memory matter but the map records only immediate neighbors, outcomes need not be functions of the declared exposure. Network logging, sensitivity analysis, and cluster isolation are therefore part of the design. Interference is a mechanism to study, not a nuisance to remove by relabeling correlated observations as separate trials.

9 A conditional network convergence theorem

A tractable proposed response model summarizes the action-mediated population response by

$$F_d(q) = \mathrm{clip}_{[0,1]}(b + BDq), \qquad D = \mathrm{diag}(d_1, \dots, d_m). \tag{9}$$

Here $q \in [0,1]^m$, $b \in \mathbb{R}^m$, $B \in \mathbb{R}^{m \times m}$, and $d \in \{0,1\}^m$ is fixed. Clipping acts coordinatewise. The coefficients are dimensionless probability-response parameters. The model can be realized by agents whose action probabilities are the coordinates of $F_d(q)$ and whose binary outcomes equal those actions. It is a reduced response model, not a physical field equation.

Let $\|v\|_\infty = \max_i |v_i|$ and $c = \|BD\|_\infty = \max_i \sum_j |B_{ij}| d_j$. Exact mean-response retraining is $q_{t+1} = F_d(q_t)$. Sampling error and changing response parameters are excluded from the following result.

Proposition 5 (Sufficient contraction condition)**.** *If $c < 1$, F_d has a unique fixed point $q^* \in [0,1]^m$, and for every initial vector and every integer $t \geq 1$,*

$$\|q_t - q^*\|_\infty \leq c^t \|q_0 - q^*\|_\infty. \tag{10}$$

Proof. Scalar clipping is 1-Lipschitz, so $\|F_d(q) - F_d(r)\|_\infty \leq c\|q - r\|_\infty$. If $c = 0$, the map is constant and the result follows after one step. For $0 < c < 1$, successive differences decrease geometrically, making the iterates Cauchy by the geometric-series bound. The closed cube is complete, so they converge to some q^*. Continuity implies $F_d(q^*) = q^*$. Two fixed points would have distance at most c times that same distance, forcing equality. Iterating the Lipschitz inequality against the fixed point proves (10). □

The condition is sufficient, not necessary. It certifies convergence within this specified response model. It does not certify truth, low performative risk, or welfare. Disclosure can change D and hence both the fixed point and the contraction bound. A fitted matrix is not a certificate unless its uncertainty and the adequacy of the reduced model are addressed.

10 Damping can help, but it is not a universal cure

Consider relaxed retraining with $\lambda \in (0, 1]$:

$$q_{t+1} = (1 - \lambda)q_t + \lambda F_d(q_t). \tag{11}$$

The convex combination remains in the probability cube. Under the preceding contraction condition, the relaxed map is Lipschitz with constant at most $1 - \lambda + \lambda c < 1$; this follows directly from the triangle inequality. That bound can be conservative when feedback has a negative sign.

For scalar negative feedback $F(q) = 1 - q$, ordinary retraining alternates forever unless $q_0 = 1/2$. The unique fixed point is $1/2$. Relaxed retraining gives the exact error recursion

$$q_{t+1} - \tfrac{1}{2} = (1 - 2\lambda)(q_t - \tfrac{1}{2}). \tag{12}$$

It converges for every $0 < \lambda < 1$, and $\lambda = 1/2$ reaches the fixed point in one step. At $\lambda = 1$, the multiplier is -1, reproducing the two-cycle. These conclusions follow by iterating the scalar multiplier, with no appeal to simulation.

Positive feedback behaves differently. Take $F(q) = \mathrm{clip}_{[0,1]}(2q - 1/2)$, whose interior fixed point is $q^* = 1/2$. In a neighborhood where clipping is inactive, the relaxed error multiplier is $1 + \lambda > 1$ for every permitted λ. Thus arbitrarily small nonzero perturbations move away from the interior fixed point until they leave that neighborhood. Damping slows the departure but cannot stabilize that point. The clipped endpoints are also fixed points; their existence does not rescue the interior stability claim.

This pair of examples is an exact counterexample to the rule "retrain more slowly and feedback will become stable." The direction and magnitude of response matter. Delay, adaptation, and finite-sample noise introduce further issues, as the stateful and gradual-response literature makes clear [7, 8]. The scalar formulas do not establish convergence for arbitrary asynchronous networks or adaptive agents.

A practical system should therefore estimate response dynamics under controlled disclosure, test candidate update schedules, and retain an independent stopping rule. It should not infer safety from a small learning rate or from a temporarily smooth forecast trajectory.

11 Stable eigenvalues and rapid agreement can mislead

Even a locally stable linear response can amplify a perturbation before it decays. Consider the two-dimensional affine model around $q^* = (1/2, 1/2)$ with

$$B = \begin{pmatrix} 1/2 & 2 \\ 0 & 1/2 \end{pmatrix}, \qquad b = \begin{pmatrix} -3/4 \\ 1/4 \end{pmatrix}. \tag{13}$$

Both eigenvalues are $1/2$. Start with error $e_0 = (0, 1/100)$, where $e_t = q_t - q^*$. For $t \geq 1$, direct multiplication or induction gives

$$e_t = \left(\frac{2t}{100}(1/2)^{t-1},\ \frac{1}{100}(1/2)^t \right). \tag{14}$$

To verify the formula, the second coordinate is halved at each step, while the first becomes half its old value plus twice the old second coordinate; substitution changes the factor t to $t + 1$. The errors stay within the unclipped interior. Their maximum norm rises from 0.01 to 0.02 at the first step, remains 0.02 at the second, and subsequently decays. Thus eigenvalues inside the unit circle do not imply monotone decrease in the chosen norm. The row-sum contraction test fails here, as it should, but failure of that sufficient test does not imply divergence.

A separate counterexample concerns agreement. Any row-stochastic averaging matrix preserves a constant forecast vector. If every agent begins at probability 0.8 while an external event has fixed probability 0.2, the group is already in perfect consensus and remains wrong. DeGroot-style agreement is a property of the update rule [9]. Truth requires additional assumptions about signals and influence, as in the analysis of Golub and Jackson [10].

In a reflexive setting, the group might also act so that the event probability moves toward 0.8. Its later calibration would then concern the world under that intervention. It would not retroactively prove that the original forecast correctly described the undisclosed world. Agreement, influence, and information must be measured separately.

These examples explain why the audit reports transient response, fixed-point behavior, forecast loss, and causal outcome effects as different quantities. A single “coherence” score would obscure precisely the distinctions that make a self-observing society scientifically interesting.

12 How influence enters the forecast score

Frozen forecasts permit an exact algebraic link between outcome influence and Brier-score change. Retain the binary potential outcomes and equal-cluster weighting from Section 7. For every cluster and agent, q_{ci} has the same pre-assignment value under both disclosure conditions.

Proposition 6 (Binary score-effect identity)**.** *The finite-population causal effect of disclosure on Brier loss is*

$$\tau_B = \frac{1}{C}\sum_c \frac{1}{m_c}\sum_i (1-2q_{ci})[Y_{ci}(1)-Y_{ci}(0)]. \tag{15}$$

The corresponding identity also holds after averaging over random potential outcomes, whenever the stated forecasts remain fixed across the two conditions.

Proof. For a binary value y, $y^2 = y$, so $(q-y)^2 = q^2+(1-2q)y$. Subtracting the losses under disclosure and withholding cancels q^2 and leaves $(1-2q)[Y(1)-Y(0)]$. Averaging with the specified cluster and agent weights proves the formula. Linearity of expectation gives the averaged version. □

The identity is a transparent finite derivation, not a claim of a new general scoring theorem. It shows that a lower Brier loss under disclosure can result entirely from moving outcomes toward the issued forecasts. That improvement can be beneficial, harmful, or neutral depending on the event and the actions used to change it. The loss difference alone does not settle the interpretation.

There is also an exact blind spot. If every issued forecast is $q_{ci} = 1/2$, then every coefficient in (15) is zero. Disclosure can change every binary outcome while leaving Brier loss unchanged at 1/4. A null score effect therefore does not establish a null behavioral effect. Conversely, outcome changes can cancel in the average prevalence while producing a nonzero score effect because the forecast weights differ.

The formula requires fixed forecasts and binary outcomes. If disclosure also changes the issued prediction, the cancellation no longer isolates outcome influence. If the outcome is continuous, its squared term does not reduce to itself. These limits are another reason to record exactly what was forecast, before which assignment, and on what outcome scale. The proposed evaluation reports both τ_S and τ_B, together with action changes and an independently justified welfare measure.

12.1 The sharp region of possible score effects

A Guide's explanation

This is the central theorem. It answers a precise question: if disclosing forecasts changes how often an event happens by a given amount, how much can the forecaster's score change? Theorem 1 gives the exact range, and shows that a score can improve a great deal while the average event rate does not move at all, when effects in different groups cancel.

The identity in Section 12 gives a complete finite effect frontier. Index registered units or strata by $j \in \{1, \dots, N\}$, with $N \geq 1$, positive weights w_j summing to one, and frozen forecasts $q_j \in [0,1]$. Equal-cluster weighting is recovered with $w_{ci} = 1/(Cm_c)$. Let

$$\delta_j = \mathbb{E}[Y_j(1) - Y_j(0)] \in [-1,1], \qquad a_j = 1 - 2q_j, \qquad A = \sum_j w_j a_j.$$

The two mean causal effects are $s = \sum_j w_j \delta_j$ for event prevalence and $b = \sum_j w_j a_j \delta_j$ for Brier loss. Expectations refer to declared potential-outcome laws, with forecasts and weights fixed. No restriction on the dependence among units is needed for these mean identities.

Relabel the units so that $a_1 \geq \cdots \geq a_N$, carrying their weights with them. Put $W_0 = 0$ and $W_k = \sum_{j=1}^{k} w_j$. Define $H : [0,1] \to \mathbb{R}$ by $H(0) = 0$ and, for $W_{k-1} \leq t \leq W_k$,

$$H(t) = \sum_{j<k} w_j a_j + (t - W_{k-1}) a_k. \tag{16}$$

The definitions agree at interval boundaries; $H(1) = A$. Ties in a do not change the function.

Theorem 1 (Sharp prevalence–score effect frontier)**.** *For each prescribed prevalence effect $s \in [-1,1]$, the attainable Brier-effect values over all binary potential-outcome laws are exactly*

$$A - 2H\left(\frac{1-s}{2}\right) \leq b \leq 2H\left(\frac{1+s}{2}\right) - A. \tag{17}$$

Both bounds are attained. Every value between them is attained as well.

Proof. Write $x_j = (\delta_j + 1)/2 \in [0,1]$. The prevalence constraint becomes $\sum_j w_j x_j = t = (1+s)/2$, and $b = 2\sum_j w_j a_j x_j - A$. To maximize the latter sum, allocate the available weighted mass to the largest a_j first. Indeed, if $a_i \geq a_j$, $x_i < 1$ and $x_j > 0$, transfer any sufficiently small weighted mass $h > 0$ from j to i by increasing x_i by h/w_i and decreasing x_j by h/w_j. Feasibility is preserved and the objective changes by $h(a_i - a_j) \geq 0$. Successive transfers yield a vector with a prefix of ones, at most one fractional coordinate, and a suffix of zeroes. Its value is $H(t)$, proving the upper bound and attainment. Minimizing is equivalent to maximizing the complement $1 - x$, whose budget is $1 - t$. Hence the minimum weighted sum is $A - H(1-t)$, proving the lower bound.

The feasible vectors form a nonempty convex compact set, and b is linear, so its image is the entire closed interval between its extrema. Finally every feasible δ is realized by binary potential outcomes with marginals $\mathbb{P}(Y_j(1) = 1) = (1 + \delta_j)/2$ and $\mathbb{P}(Y_j(0) = 1) = (1 - \delta_j)/2$. Any joint coupling with those marginals supplies a valid potential-outcome law. This establishes causal attainability, not just an algebraic relaxation. □

12.2 A diagnostic for influence hidden by averages

The frontier is the fractional-allocation solution of a finite linear problem, specialized here to the frozen-forecast score identity. Its constructive proof credits the elementary optimization principle directly. It describes mean effects under arbitrary binary potential-outcome laws; deterministic individual effects restrict the feasible set to finitely many weighted combinations and need not attain every intermediate value without averaging or randomization.

For two equally weighted strata with $q = (0.1, 0.9)$, the score coefficients are $a = (0.8, -0.8)$ and $A = 0$. At zero average prevalence effect, $H(1/2) = 0.4$, so

$$s = 0 \quad \Longrightarrow \quad -0.8 \leq b \leq 0.8.$$

Taking $\delta = (-1, 1)$ moves each stratum's outcome toward its own forecast: the average event rate is unchanged, but Brier loss falls by 0.8. The opposite effects attain $+0.8$. Thus an unchanged average event rate can hide a substantial and completely ordinary causal improvement or deterioration in the reported forecast score.

When every forecast equals the same q_0, the frontier collapses to $b = (1-2q_0)s$. In particular, $q_0 = 1/2$ makes the entire score effect zero even while prevalence changes. At $s = 1$ or $s = -1$, all δ_j are forced to the corresponding endpoint, so the score effect is respectively A or $-A$. These boundary cases follow directly from the theorem and provide useful implementation checks.

The frontier is a consistency region, not a substitute for randomization. An observed point outside the region, after accounting for statistical uncertainty, challenges at least one premise: frozen forecasts, binary outcomes, declared weights, or the estimated causal contrasts. A point inside it can have many heterogeneous explanations. The registered intervention still identifies the contrast; the theorem states exactly which combinations its mathematical model permits.

The revision checks compare the closed form with exhaustive vertices of the constrained effect polytope for finite rational forecasts, weights and prevalence values. They include ties, unequal cluster weights, boundary prevalence, constant forecasts and the cancelling-strata construction. All are generated mathematical checks. The next empirical study should report the estimated pair (s, b) with joint uncertainty and its registered frontier, while separately measuring actions and participant welfare.

13 Executed synthetic and finite checks

The accompanying `check_26.py` uses rational arithmetic for finite probabilities and response trajectories. It writes all reported outputs to JSON. No human behavior was sampled, and no forecast success on real-world events is claimed. The examples are designed to challenge inference, not to imitate an empirical benchmark.

Enumeration of the two causal models in Section 6 confirms identical observed prediction–outcome tables and different intervention effects, zero versus one. The copy-on-disclosure example yields Brier losses zero when disclosed, 1/2 when withheld, and 1/4 for the hidden constant-probability baseline. A grid over $\theta \in [0, 1]$ confirms that the Bernoulli response model makes every grid point stable while risk is maximal at 1/2.

For a two-agent response with $b = (1/5, 1/10)$ and

$$B = \begin{pmatrix} 0 & 2/5 \\ 3/10 & 0 \end{pmatrix}, \tag{18}$$

full disclosure has contraction constant 2/5 and fixed point $(3/11, 2/11)$. Withholding both messages leaves the fixed point at b. Exact iterations from several starting vectors satisfy the contraction inequality. Negative feedback reproduces the two-cycle at full retraining and convergence under relaxation; the nonnormal example reproduces its temporary norm increase.

For four synthetic clusters with potential scores $(0, 1/4, 1/2, 3/4)$ when withheld and $(1/4, 1/2, 3/4, 1)$ when disclosed, all six balanced assignments are enumerated. Their estimated effects range from $-1/4$ to 3/4, but average exactly to the true effect 1/4. The pair-interference example returns a zero own-assignment contrast and a full-cluster effect of one. Exposure-weighted estimates average to their stated targets under all four assignments.

Finally, exhaustive binary outcome pairs and a rational forecast grid verify equation (15), including the $q = 1/2$ blind spot. These checks establish internal consistency of the constructions. They do not validate linear response, network isolation, a learned forecaster, or the behavioral realism of the policies. Those remain empirical questions.

14 An experiment that separates prediction from influence

Begin with controlled simulated environments in which all actions and outcomes can be inspected. Generate and timestamp forecasts before disclosure assignment, fix the event definition and horizon, and preserve the complete forecast record. An evaluator should not choose the most impressive forecast after outcomes are known. Withheld forecasts must remain inaccessible to agents and to any controller capable of affecting the target.

Use independent clusters as the primary experimental units. Randomize whole clusters between the registered disclosure policies, while matching computational effort, latency, task information unrelated to disclosure, and evaluation schedules. A shadow-mode baseline can compute the same forecasts without publishing them. A message-format control can distinguish effects of numerical content from mere attention or the presence of a message, provided that its meaning and any experimental alteration are handled transparently.

Log actions, communication, and environmental transitions as well as outcomes. These records help test the proposed action-mediated mechanism, but mediation is not identified simply because a plausible action occurred between message and event. Additional mediator interventions or justified causal assumptions are needed to attribute the total effect to a particular pathway. The primary randomized estimand remains the disclosure-policy effect.

Introduce prespecified variation in communication topology, forecast quality, response strength, and adaptation delay. Keep tuning environments separate from final evaluation. Compare a forecast-only learner, naive retraining, a response-aware method from the performative literature, and the proposed disclosure-aware evaluation pipeline [4, 6, 7]. A reference response matrix is useful only if its predictive consequences survive these comparisons.

Repeated rounds create carryover. Use fresh independent clusters when feasible, or explicitly model persistent state and use a design that supports the intended temporal inference. Withholding a message today does not erase yesterday's public forecast from shared memory. Likewise, nominal cluster boundaries fail if agents communicate across them. Isolation should be tested and documented rather than assumed from the assignment table.

The first success criterion is explanatory discrimination: can the design distinguish information, influence, coordination, and resistance in environments where the mechanism is known? Only after that test should more open-ended multi-agent societies be used to evaluate generality.

15 Precision belongs to clusters, not message counts

For a population-level disclosure effect, suppose clusters are independently sampled from a fixed population, with $n \geq 1$ sampled units per arm and assignment independent of potential outcomes. Let $R_c(z) \in [0,1]$ be the registered cluster score, and let $\tau^{\text{pop}} = \mathbb{E}[R(1)] - \mathbb{E}[R(0)]$. This target averages over the population of clusters; it differs from conditioning on one fixed finite set of potential outcomes as in Section 7.

Proposition 7 (A conservative population precision bound)**.** *For independent cluster samples in each arm and $\alpha \in (0,1)$,*

$$\mathbb{P}\left(|\hat{\tau}_R - \tau^{\text{pop}}| \leq 2\sqrt{\frac{\log(4/\alpha)}{2n}}\right) \geq 1 - \alpha. \tag{19}$$

A sufficient sample count for half-width $\eta > 0$ is $n \geq 2\log(4/\alpha)/\eta^2$ per arm.

Proof. Hoeffding's inequality [18] bounds each arm-mean error exceeding r by $2e^{-2nr^2}$. The union bound for two arms is $4e^{-2nr^2}$. If neither error exceeds r, the difference error is at most $2r$. Substitution of $r = \sqrt{\log(4/\alpha)/(2n)}$ proves the claim; solving $2r \leq \eta$ gives the sample expression. □

At $\alpha = 0.05$ and $\eta = 0.1$, the sufficient count is 877 independent clusters per arm. This distribution-free calculation is deliberately conservative. It is not a power calculation and does not turn 877 messages from one society into 877 replicates.

For a fixed finite collection of clusters, use an analysis justified by the randomization design. A permutation test under a sharp no-effect null can reassign the observed outcomes because that null fixes every unit's missing potential outcome; an average-zero null alone does not supply that permission. Network-exposure estimators require their dependence-aware variance analysis [15]. Report uncertainty and multiplicity across horizons, scores, topologies, and subgroups. A null result with a wide interval is inconclusive; a precise exclusion of the registered meaningful effect is substantive evidence against that hypothesis.

16 Disclosure is an intervention with an owner

A prediction can be useful precisely because it changes behavior. A warning that prevents a failure may deserve credit even though the failure never occurs. An evaluation that rewards only literal fulfillment could punish the successful intervention and encourage a harmful self-fulfilling forecast. The governance problem is to state which outcomes are desired and which means of influencing them are authorized.

The Continuum therefore treats disclosure policy as part of the system specification. Who receives the forecast, at what time, with what explanation of uncertainty, and with what opportunity to contest it? These choices affect both the causal distribution and the participants' agency. They should not be optimized solely to improve a prediction score.

For human studies, informed participation, appropriate review, and clear limits on behavioral intervention are necessary. For artificial societies, an explicit authority structure still matters: a forecast module should not silently acquire permission to modify the environment until its prediction becomes true. The mathematical model describes possible influence; it does not authorize manipulation. Any experiments involving altered messages should disclose the experimental context and remain within the approved design.

A useful audit separates four records: the evidence available before prediction, the forecast actually issued, the disclosure decision, and the actions that followed. These records support different claims. A correct timestamp helps establish sequence but does not establish causal validity. A complete action log improves accountability but does not prove that no unrecorded influence existed. Independent measurement and controlled comparison remain necessary.

Disagreement can be valuable. If a society suppresses dissent to make forecasts converge, it may reduce observed variation without improving its model of the world. The scientific objective is not maximal agreement; it is an intelligible relationship between evidence, forecast, response, and outcome. Normative commitments about agency and permissible influence must remain visible alongside that relationship rather than being hidden inside an optimization target.

17 Falsification and opposing outcomes

The core empirical hypothesis is that explicit disclosure and response modeling can improve out-of-sample prediction of both outcomes and causal disclosure effects in specified multi-agent environments. It is weakened if a simpler passive model predicts the registered outcomes equally well, if inferred response parameters fail under intervention, or if apparent benefits disappear when compute, information, and tuning are matched.

A second hypothesis concerns governance: separating forecast information from influence may expose self-confirming evaluation errors. It would fail in a particular study if the proposed controls do not discriminate the known synthetic mechanisms, or if leakage makes the withheld condition an unreliable counterfactual. A design that cannot distinguish the exact constructions in Section 6 is not ready to support a stronger claim about an open-ended society.

The stability hypotheses are conditional. If the true response violates the proposed contraction assumptions, observed oscillation does not refute the mathematical proposition. It refutes the use of that proposition as a certificate for the fitted model. Conversely, convergence when the row-sum bound exceeds one does not contradict a sufficient condition. The nonnormal and scalar examples are included to make these boundaries explicit.

Evidence can also reveal that influence is weaker than expected. A precise null average outcome effect under a well-powered, isolated disclosure intervention would limit the claim about average prevalence for that task and population. It should not be reinterpreted as a hidden field effect or as proof that the forecast was too advanced to measure. A null average outcome effect does not preclude a score effect: heterogeneous outcome changes may cancel in prevalence but remain after weighting by the frozen forecasts. Zero individual outcome effects force zero score effect; an outcome effect without score change is also possible by Section 12.

Failures should be retained in the research record. The strongest version of the Continuum's thesis is not that every prediction creates its future. It is that a society can learn to distinguish when its forecasts describe, change, or destabilize the world. That distinction becomes credible only when the method is permitted to return "no detectable influence," "model rejected," or "causal effect unresolved."

18 Scope, limitations, and the contribution boundary

The reference models omit many features of real societies: strategic deception, heterogeneous goals, changing communication graphs, continuous actions, endogenous membership, and long-lived institutional memory. Some can be incorporated by expanding state and policy classes, but the existing proofs do not automatically survive every extension. In particular, stochastic learning updates, delayed observations, and time-varying response matrices require their own convergence analysis.

The causal designs rely on credible assignment, consistency, and an adequate interference model. Cluster randomization does not eliminate spillovers across poorly chosen boundaries. Exposure weighting does not correct an incorrect exposure map. Frozen forecasts do not by themselves ensure that all ordinary information sources are matched. These assumptions must be defended empirically for each study.

The literature establishes a historical line from self-fulfilling prophecy and reflexivity to policy response, market performativity, and modern response-aware learning [19, 20, 21, 22, 4]. Proper scoring, network learning, and interference-aware inference supply additional foundations. Its proposed contribution is a coherent Continuum research contract combining action-mediated disclosure, explicit distinctions among risk and welfare, a finite network response model, and exact diagnostic examples. The score-effect identity and Theorem 1 provide a complete finite region of compatible prevalence and score effects in the frozen binary setting, including explicit attaining causal constructions.

Relative to earlier Continuum discussions [2, 3], the added substance is the multi-agent intervention design, spillover estimands, counterexamples, fully stated convergence conditions, and executed checks. That is a development of a research direction, not retrospective evidence that the canon already contained the mathematics.

The finite calculations are reproducible, but they are not empirical validation. The source review did not reproduce the cited studies or reanalyze their raw data. No claim is made about predicting random events without ordinary information, generating investment advantage, or establishing a new physical effect. The proposed frontier is difficult enough on its actual terms: understanding a society whose descriptions can become inputs to its own next actions.

19 Conclusion: forecast the response to the forecast

A Guide's explanation

A Guide can say: "The Reflexive Future is about predictions that change what they predict. A warning can prevent the very thing it warned about; a confident forecast can help make itself come true. The paper shows how to test which is happening, by sometimes withholding a forecast and measuring the difference." A Guide should not claim that its own predictions about a member are validated this way, present any forecast as certain, or suggest that information travels backward in time.

The future of a responsive society is not independent of everything that society says about it. When forecasts change available information, information changes action, and action changes outcomes, prediction becomes part of a feedback process. The Continuum can investigate this possibility rigorously by treating disclosure as an intervention and the society's response as a measurable mechanism.

The sharp score frontier makes that investigation quantitative. It gives the full range of score changes compatible with each average outcome effect under the declared forecast distribution, including changes hidden by cancelling subgroup effects. The surrounding mathematics clarifies the other evaluation targets. Perfect observational accuracy can arise from information or influence. Every parameter can be stable even when some stable parameters have higher performative risk. A proper score evaluates truthful probabilities under a fixed distribution, but does not authorize shaping that distribution to make a forecast look good. Consensus can preserve an error, damping can fail to stabilize positive feedback, and stable eigenvalues can coexist with transient amplification.

The proposed empirical contract gives each ambiguity a test. Freeze forecasts before assignment. Compare registered disclosure policies across independent clusters. Measure action and outcome effects separately from score effects. Model spillovers where isolation is impossible. Evaluate uncertainty at the actual unit of randomization. Publish the failures that distinguish a predictive mechanism from a self-confirming story.

A successful study would establish how a forecasting society's messages alter its world, which response patterns persist under intervention, and whether its disclosure policy benefits participants. A negative result would identify which part of the registered response model failed. Both outcomes advance an accountable science of predictive influence.

The Reflexive Future is therefore a thesis about responsibility as well as prediction. An intelligent society should be able to ask what would have happened had it remained silent, what changed because it spoke, and whether the change was permitted and beneficial. Those questions make bold forecasting claims harder to defend. They also make the claims worth taking seriously.

References

[1] Architect X13. *Eidara Continuum Bible*. Canonical archive, 12 October 2026, 116 pp.; forecasting motifs, p. 15; observer-responsive narratives, pp. 26–27. Private doctrinal source; not empirical verification.

[2] Architect X13. *Pillar VII: Ouroboric Collapse Resonance*. Revised Continuum research white paper (2026), related-work discussion of performative prediction. Project prior work; not independent scientific verification.

[3] Architect X13. *Multimodal Consciousness: A lawful-collapse research framework for substrate-independent awareness*. Revision 2 (2026), shadow-mode and causal-use experiments. Project prior work; not independent scientific verification.

[4] J. C. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative Prediction. *Proceedings of ICML*, PMLR 119, 7599–7609 (2020). https://proceedings.mlr.press/v119/perdomo20a.html

[5] J. P. Miller, J. C. Perdomo, and T. Zrnic. Outside the Echo Chamber: Optimizing the Performative Risk. *Proceedings of ICML*, PMLR 139, 7710–7720 (2021). https://proceedings.mlr.press/v139/miller21a.html

[6] Z. Izzo, L. Ying, and J. Zou. How to Learn when Data Reacts to Your Model: Performative Gradient Descent. *Proceedings of ICML*, PMLR 139, 4641–4650 (2021). https://proceedings.mlr.press/v139/izzo21a.html

[7] G. Brown, S. Hod, and I. Kalemaj. Performative Prediction in a Stateful World. *Proceedings of AISTATS*, PMLR 151, 6045–6061 (2022). https://proceedings.mlr.press/v151/brown22a.html

[8] Z. Izzo, J. Zou, and L. Ying. How to Learn when Data Gradually Reacts to Your Model. *Proceedings of AISTATS*, PMLR 151, 3998–4035 (2022). https://proceedings.mlr.press/v151/izzo22a.html

[9] M. H. DeGroot. Reaching a Consensus. *Journal of the American Statistical Association* 69(345), 118–121 (1974). https://doi.org/10.1080/01621459.1974.10480137

[10] B. Golub and M. O. Jackson. Naïve Learning in Social Networks and the Wisdom of Crowds. *American Economic Journal: Microeconomics* 2(1), 112–149 (2010). https://doi.org/10.1257/mic.2.1.112

[11] C. Hommes, J. Sonnemans, J. Tuinstra, and H. van de Velden. Coordination of Expectations in Asset Pricing Experiments. *Review of Financial Studies* 18(3), 955–980 (2005). https://doi.org/10.1093/rfs/hhi003

[12] J. Pearl. Causal diagrams for empirical research. *Biometrika* 82(4), 669–688 (1995). https://doi.org/10.1093/biomet/82.4.669

[13] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5), 688–701 (1974). https://doi.org/10.1037/h0037350

[14] M. G. Hudgens and M. E. Halloran. Toward Causal Inference With Interference. *Journal of the American Statistical Association* 103(482), 832–842 (2008). https://doi.org/10.1198/016214508000000292

[15] P. M. Aronow and C. Samii. Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics* 11(4), 1912–1947 (2017). https://doi.org/10.1214/16-AOAS1005

[16] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78(1), 1–3 (1950). https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml

[17] T. Gneiting and A. E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* 102(477), 359–378 (2007). https://doi.org/10.1198/016214506000001437

[18] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* 58(301), 13–30 (1963). https://doi.org/10.1080/01621459.1963.10500830

[19] R. K. Merton. The Self-Fulfilling Prophecy. *The Antioch Review* 8(2), 193–210 (1948). https://doi.org/10.2307/4609267

[20] G. Soros. Fallibility, reflexivity, and the human uncertainty principle. *Journal of Economic Methodology* 20(4), 309–329 (2013). https://doi.org/10.1080/1350178X.2013.859415

[21] C. A. E. Goodhart. Problems of Monetary Management: The UK Experience. In *Monetary Theory and Practice*, 91–121 (1984). Published reprint; earlier version presented in 1975. https://doi.org/10.1007/978-1-349-17295-5_4

[22] D. MacKenzie. *An Engine, Not a Camera: How Financial Models Shape Markets*. MIT Press (2006). Historical research monograph. https://mitpress.mit.edu/9780262134606/an-engine-not-a-camera/