— 13 —

When the model moves the data

Models deployed at scale always influence future data, a phenomenon called performativity. Performativity breaks evaluation and creates the problem of data feedback loops. Dynamic benchmarks try to make a virtue out of it.

Source: The Emerging Science of Machine Learning Benchmarks. M. Hardt, 2025. URL: https://mlbenchmarks.org. Compiled on 2025-11-17.

It's about time that we questioned the *astronomical conception* that machine learning stands on. In the astronomical conception, populations are fixed probability distributions. You look for statistical regularities in populations in the same way that an astronomer would find patterns in the universe by pointing a telescope at the sky. Nothing the astronomer does at the observatory could possibly influence the stars. By analogy, in the traditional view of machine learning, the data-generating distribution is fixed and immutable. Nothing the model does could possibly influence the distribution. This chapter is about why this idea breaks down, what that means for evaluation, and how to cope.

Recall, the goal of prediction is to make an educated guess about an unavailable outcome y from available features x. If the outcome is binary $y \in \{0,1\}$, the best prediction given x is to choose the value $\hat{y} \in \{0,1\}$ such that $(x,\hat{y})$ is more likely to occur in the population than $(x, 1-\hat{y})$. If the outcome y is real-valued, pick its mean value among all instances with features x. The first rule minimizes the probability of a prediction error. The second minimizes the mean squared error of our prediction. More generally, optimal predictors minimize *risk*, the average of a loss function over the population.[1,2]

When this formalism came up in the 1930s, no one could've imagined just how successful prediction would turn out to be. Today, digital platforms deploy machine learning models—all built on the astronomical conception—to make billions of predictions that shape what we read, write, watch, buy, and eat. Every time you click on a YouTube video, for example, you act on a model predicting that you'd watch this video.

But the predictions of a content recommendation model on a digital platform outright defy the astronomical conception. If the model predicts that a visitor will watch a video ($\hat{y} = 1$), the platform will display the video prominently to the visitor, who is therefore more likely to click and watch the video ($y = 1$). If the platform doesn't recommend the video ($\hat{y} = 0$), the visitor is almost certainly not going to see it ($y = 0$). Clearly, the prediction $\hat{y}$ influences the outcome y, significantly so. In other words, there is no model-independent outcome. Outcomes fundamentally depend on what the model predicts.

In this example, predictions are a kind of *self-fulfilling prophecy*. Predicting high watch time increases actual watch time. Perversely, the model then looks more accurate than it actually is. Even if it was just making random guesses about watch time, it will look accurate—once deployed—due to how it influences actual watch time. The influence of a prediction on the outcome could go the other way, too. If a popular traffic app predicts low traffic on a specific route, drivers switch over, thus increasing traffic. In a kind of self-negating prophecy, traffic predictions tend to look worse than they are. In both cases, we'd be fooling ourselves if we evaluated the model on data that depends on the model's predictions.

What's troubling is that it's not even possible to express these problem in the standard language of prediction. Here, the data-generating distribution is immutable and cannot possibly depend on any predictions we choose based on that distribution. When we draw an instance (x, y) from a distribution we implicitly assume that the label y actualizes together with features x. Nothing we do with the features—like a making a prediction—could therefore change the label.

The problem we run into here is *performativity*. Performativity refers to a causal influence that a model and its predictions have on the target of prediction. It arises with all predictions that people react to: election polls, epidemiological forecasts, credit scores, online recommendations, digital ads, traffic estimates, chatbots, and AI assistants. In each case, people respond to a prediction in a way that may change the outcome, thus possibly reinforcing or invalidating the prediction. Online recommendations, as we saw, can be self-fulfilling. Traffic predictions, on the other hand, can be self-negating. The strength and direction of performativity varies from one case to the other.

Performativity implies that there is no model-independent ground truth. Outcomes necessarily depend on whatever model you deployed. But model-independent ground truth is the entire basis of evaluation and benchmarking. It's the essential ingredient we rely on when we set aside a test set for benchmarking purposes. Does performativity therefore turn prediction into guesswork with unforeseeable results—a kind of alchemy, profitable but obscure? This is precisely what the economist feared who first reckoned with performativity a century ago.

13.1 Morgenstern's prophecy about prediction

The first person to clearly recognize this conundrum central to prediction was the economist Oskar Morgenstern. Almost a century ago, Morgenstern studied the feasibility of economic forecasting. This was more than a decade before his work with von Neumann that founded the field of game theory.[3]

As statistics flourished in the 1920s, many of Morgenstern's contemporaries hoped to apply new statistical methods to the problem of charting the course of an economy. Morgenstern believed that this was a fool's errand. Economic forecasting, he argued in his century-old habilitation, was impossible with the tools of economic theory and statistics alone.[4]

There is a good reason for Morgenstern's pessimistic outlook on prediction. Any economic forecast, published with authority and reach, would necessarily cause economic activity that influences the data that the forecast derived from. This causal relationship between a prediction and its target, Morgenstern held, necessarily invalidated economic forecasts:

> A forecast somehow based on the [central limit theorem] thus saws off the branch it is sitting on and renders itself impossible after the first prediction, because in doing so it deprives itself of its empirical foundation.[4]

Predictions derive from frequencies in a population. But those frequencies aren't stable. A forecast can change the behavior of people, which in turn disrupts the statistical frequencies that the forecast stands on. As a consequence, a statistical model that is accurate in current conditions may not be accurate in the future conditions that result from *acting on* the model's predictions. Morgenstern's astute argument foreshadowed the 1970s Lucas critique that haunts macroeconomics to this day.

Although he didn't use the same term, Morgenstern accurately described the phenomenon of performativity, which he called one of the most central and difficult problems in the theory of prediction. Along the way, Morgenstern put his finger on the trouble with the astronomical conception:

> There is no causal connection between the prediction of astronomical events and their actual occurrence; in other words, the forecast can in no way exert any conceivable influence on the actual occurrence and unfolding of situations involving celestial bodies that follow their own mechanical laws.[4]

This is different when predicting things relating to people. The problem that clouds economic forecasts, Morgenstern held, is fundamental to predictions in the social world at large. The ubiquity of prediction today makes Morgenstern's critique more relevant now than ever before.

Simon's counterpoint

A first attempt at a formal counterpoint to Morgenstern's argument came thirty years later in a paper by Emile Grunberg and Franco Modigliani, and in a contemporaneous work by economist and AI pioneer Herbert Simon. Grunberg and Modigliani studied prices,[5] whereas Simon considered performativity in election forecasts.[6] In popular elections, a *bandwagon* effect is a kind of self-fulfilling prophecy that occurs when voters pick a candidate, because the candidate is strong in the polls. An *underdog* effect is a self-negating prophecy, where voters turn up in greater support for a candidate that polls poorly.

Grunberg and Modigliani distinguish between *private* and *public* predictions. Private predictions have no causal powers, whereas public predictions can alter the course of events. An economic forecast goes from private to public, once an agency communicates the forecast and people start acting on it. The notion of a public prediction raises a basic question: Under what conditions will a public prediction—even though it may change the course of events—still come true?

The three economists all came to the same verdict. It is, in principle, possible to make a public prediction that equals the outcome caused by the prediction. All that is needed, they argued, is the continuity of the function that relates predictions to outcomes. Such a prediction is a kind of *stable point*, where what we predict equals the outcome we get. That isn't true for *any* prediction under performativity, but at least we can find *one* where it is. This provides at least some way to cope with performativity.

To study this setting formally, consider a real-valued outcome $y \in [0,1]$. Denote the prediction $\hat{y}$. A *perfect* prediction, according to Grunberg and Modigliani, corresponds to the case that $\hat{y} = y$. Now express the relationship between the prediction and the outcome it causes through a response function

$$y = r(\hat{y}).$$

A perfect prediction then is any fixed point of the response function:

$$\hat{y} = r(\hat{y}).$$

Grunberg, Modigliani, and Simon observed that the continuity of the response function is sufficient to ensure that a fixed point—and therefore a perfect prediction—exists. Formally, their argument invokes Brouwer's fixed point theorem.[7] However, the one-dimensional case relevant here is just the intermediate value theorem from calculus.

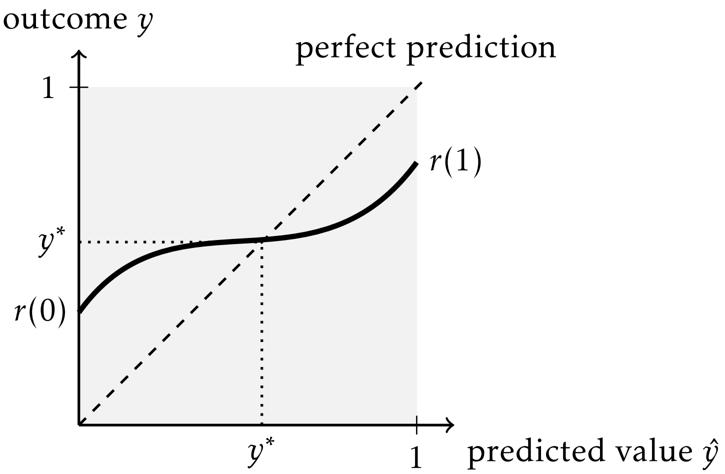

Figure 13.1: Herbert Simon's argument for the existence of stable points hinges on the continuity of the response function that relates predictions to outcomes.

First, draw a point anywhere on the y-axis, representing the realized outcome $r(0)$ when the prediction is $\hat{y} = 0$. Then, mark a second point on the vertical line $\hat{y} = 1$, representing the realized outcome $r(1)$. Under the constraint that your pencil may not leave the square it is impossible to connect the two points without touching the dashed line or lifting the pencil. Thus, for any continuous relationship between y and $\hat{y}$ there must be at least one point y^* for which $y^* = r(y^*)$. Thus, y^* comes true *after* being published.

13.2 Performative prediction

The standard theory of prediction starts from a fixed distribution $\mathcal{D}$ over instances of the prediction problem. To talk about performativity, we'll need to allow the distribution to depend on the predictive model. A *distribution map* expresses the dependence of the data-generating distribution on the predictive model. Formally, it is a mapping

$$\mathcal{D}\colon \Theta \to \Delta(\mathcal{X} \times \mathcal{Y}).$$

For every parameter vector $\theta \in \Theta$, the distribution $\mathcal{D}(\theta) \in \Delta(\mathcal{X} \times \mathcal{Y})$ describes the data-generating distribution that results from *deploying*—that is, taking action according to—the predictive model f_θ. Analogous to the idea of a *public* prediction, we consider a model *deployed* when it is used across a population to make predictions for multiple individuals.

Consider a couple of examples:

1. **Watch time prediction.** The features X describe both a user profile and a video. The prediction $f_\theta(X)$ represents the platform's predicted watch time. The outcome variable Y represents the time the user spends watching a video. The prediction influences watch time, since the platform ranks videos in descending order of predicted watch time. In a simple additive model, $Y = Y_0 + \beta f_\theta(X)$, where Y_0 is a model-independent distribution and β is the strength of performativity. The distribution map $\mathcal{D}(\theta) = (X, Y_0 + \beta f_\theta(X))$ captures the resulting joint distribution.
2. **Incentives.** Suppose a company uses a linear model $\langle\theta, X\rangle$ to award annual performance bonuses to employees. The marginal distribution X describes employee characteristics absent any incentives. An employee with features X receives a bonus proportional to $\langle\theta, X\rangle$. In response to the model θ, an employee with features X will strategically change their features to $X' = X + \beta\theta$ so as to increase their performance bonus. Suppose the target variable Y describes true employee skill and remains unaffected by the feature change. The resulting distribution map is $\mathcal{D}(\theta) = (X + \beta\theta, Y)$.

These examples describe two different mechanisms for how performativity can arise in machine learning applications. The distribution map gives a general way to describe how the distribution changes in response to model deployment. The formal setup is general and does not prescribe any specific mechanism behind the distribution map. By instantiating it in different ways, we can capture various aspects of performativity. In each case, the distribution map makes it explicit how the model influences the data-generating distribution.

In practice, a data-generating distribution might change over time for many different reasons. Performative prediction focuses on one and only one source of variation: the effect of model deployment on the data. In particular, the distribution map is stateless, meaning that deploying the same model at any point in time repeatedly leads to the same distribution. The simplicity of this formalism is intentional. It's the minimal change necessary to talk about model-dependent data-generating distributions.

Stability, optimality, performative risk

Given the concept of a distribution map, the next step is to see how the notion changes risk minimization. Recall from Chapter 2, the risk of a predictor f_θ on a distribution $\mathcal{D} = (X, Y)$ is the expected loss

$$R(f_\theta, \mathcal{D}) = \mathbb{E}\,\ell(f(X), Y)\,.$$

Compared to the risk notation in Chapter 2, we add a distribution argument to the risk functional to make its dependence on the distribution explicit. Rather than evaluating a model on a fixed distribution, now we'll evaluate the model on the distribution that results from deploying the model:

$$R(f_\theta, \mathcal{D}(\theta)) = \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}(\theta)} \ell(f_\theta(X), Y)$$

As a risk minimization objective, this is a moving target. Suppose you're currently observing data from a distribution D_0. You solve the optimization problem to find the risk minimizer θ_1 on the distribution D_0. Once you deploy θ_1, however, the distribution changes to a new state $D_1 = \mathcal{D}(\theta_1)$. And so you continue:

$$D_0 \to \theta_1 \to D_1 \to \theta_2 \to D_2 \to \dots$$

This process of *repeated risk minimization*—or *retraining*—could go on forever, never settling on a fixed point. It might cycle or oscillate. Anything could happen, in principle. However, if the process ever converges, it must be at a fixed point θ_{PS} that satisfies:

$$\theta_{\mathrm{PS}} \in \operatorname*{argmin}_{\theta\in\Theta}\; R(\theta, \mathcal{D}(\theta_{\mathrm{PS}}))\,.$$

We call any such point θ_{PS} *performatively stable*. Performative stability is a natural equilibrium notion that requires the model to look optimal on the distribution it entails. This means that the data collected after deploying θ_{PS} gives us no empirical reason to deviate from the model. It is optimal on the static risk minimization problem defined by $\mathcal{D}(\theta_{\mathrm{PS}})$. That is not to say that there isn't some other distribution where models achieve smaller loss. It's just that on this particular distribution—its "own" distribution—the model is optimal.

Performative stability admits a simple empirical check:

1. Collect data in current conditions.
2. Solve a risk minimization problem on the data.
3. Check if the current model is at least as good as the risk minimizer.

If it is, then we're at a stable point and the data we see do not refute the optimality of the model. In this sense, an *echo chamber* is a good metaphor for a stable point: What we experience doesn't challenge our current beliefs.

Stable points can, in principle, have much larger loss than other solutions. An alternative solution concept therefore asks for the smallest possible risk *post deployment* globally among all models. More specifically, we say that a predictive model with parameters θ_{PO} is *performatively optimal* if it satisfies

$$\theta_{\mathrm{PO}} \in \operatorname*{argmin}_{\theta \in \Theta} \; R(\theta, \mathcal{D}(\theta)).$$

Performative stability is about optimality in *current* conditions. Performative optimality is about optimality in *future* conditions. It's about anticipating the consequences of model deployment.

We call the risk of a model on the distribution it entails the *performative risk*, defined as

$$\mathrm{PR}(\theta) := R(\theta, \mathcal{D}(\theta)).$$

Performatively optimal models minimize performative risk by definition and it always holds that $\mathrm{PR}(\theta_{\mathrm{PO}}) \le \mathrm{PR}(\theta_{\mathrm{PS}})$ for any performative optimum θ_{PO} and stable point θ_{PS}. In a static setting the two solution concepts, stability and optimality, both coincide with the classical supervised learning solution. In general, however, performatively stable points need not be optimal and optimal points need not be stable.

In contrast to stability, certifying optimality is harder, as the data available to us need not tell us anything about the performance of any other model post deployment. Moreover, performatively optimal points may be degenerate. For example, it could be that the distribution $\mathcal{D}(\theta_{\mathrm{PO}})$ is a point mass on a single instance where the model has smallest possible loss. Such a solution is performatively optimal, but far from desirable.

Learning versus steering. Performative prediction changes the rules of prediction. There are now two ways to be good at prediction. To appreciate this point, fix a model ϕ that we imagine provides the current data-generating distribution, and consider deploying another model θ. Observe that the model θ shows up in two places in the definition of the performative risk: in the first argument, reflecting the dependence of the loss $\ell(\theta; z)$ on θ, and in the second argument representing the dependence of the distribution $\mathcal{D}(\theta)$ on θ. Thus, we can decompose the performative risk $\mathrm{PR}(\theta)$ as:

$$\mathrm{PR}(\theta) = R(\theta, \mathcal{D}(\phi)) + (R(\theta, \mathcal{D}(\theta)) - R(\theta, \mathcal{D}(\phi)))$$

This tautology exposes two ways to achieve small performative risk. One is to optimize well in current conditions, that is, to minimize $R(\theta, \mathcal{D}(\phi))$. The other is to *steer* the data to a new distribution $\mathcal{D}(\theta)$ that permits smaller risk than $\mathcal{D}(\phi)$. The first mechanism—call it *learning*—is what machine learning is all about. Learning is the classical way that, for example, a platform can discover and target consumer preferences. Steering is original to performative prediction, allowing the platform to push consumption towards a distribution more favorable to its objectives. This second mechanism, steering, is key to understanding important questions about the impact of predictive models in applications.

What performativity means for model evaluation

Performativity is a ubiquitous phenomenon, arising with all predictions that humans act on. The phenomenon has several striking consequences for model evaluation:

1. **There is no model-independent ground truth.** All datasets we collect from live systems are model-dependent. Likewise, all metrics we compute on those datasets are model-dependent quantities.
2. **Performance on static benchmarks may not capture model performance on a live platform.** Once deployed in a live system, a model's performance is subject to performative effects not present in offline benchmarks.
3. **A decision maker—be it a machine, a human, or an institution—can appear more or less accurate than they are.** Even random guessing can look like non-trivial accuracy under positive performative effects.
4. **Direct comparisons between statistical models on static datasets and human decision makers *in situ* are often misleading.** A doctor can implore a patient to change their behavior, thus undermining her own prediction about the patient's future. A statistical model, trained and evaluated on offline data, cannot influence outcomes. Performativity disrupts the notion of a *human baseline*.
5. **Performativity is often a phenomenon of scale.** Performative effects may arise due to the networked interactions of individuals at scale over extended time horizons. As a result, small-scale experiments on short time horizons, such as randomized trials and A/B tests, may draw a misleading picture.

The point here is not that evaluation under performativity is futile. Rather, performativity should be one of the fundamental criteria that you apply to predictive systems. It's a lens on prediction that makes something important visible that's otherwise easy to overlook from a traditional perspective. In many cases, carefully designed experiments and tools from causal inference can give estimates of performative effects in practice. In other cases, theoretical models can help anticipate performative effects.

13.3 Repeated optimization converges to stable signals

Often performativity shows up in practice in the guise of *distribution shift*. Practitioners deploy a model and later observe that the distribution has changed. The standard solution is to train the model on the data and deploy it again. This process of *retraining* corresponds to the idea of repeated risk minimization mentioned earlier. Performatively stable points are fixed points of risk minimization. But under what conditions does repeated risk minimization converge to a stable point?

In general, this is a difficult question. We'll derive an illustrative answer in a simple theoretical model of performative prediction: The marginal distribution X follows a fixed probability measure D_X, supported on a bounded domain $\mathcal{X}$, that does not change with the predictor. The outcome variable Y, however, does depend on the deployed predictor f via the equation:

$$Y = \alpha f^*(X) + \beta f(X)$$

The function f^* represents a *stable signal* that influences the outcome, but does *not* depend on the prediction. We think of the coefficient $\alpha > 0$ as the strength of the stable signal, whereas $\beta \geq 0$ is the strength of performativity. Let the distribution map $\mathcal{D}(f)$ describe the resulting joint distribution (X, Y). Since we'll be talking about *non-parametric* predictors given by a function f, we write the distribution map as a function of f.

Here are two typical examples of a stable signal:

1. At each step we mix original data and most recent data. The resulting distribution is a mixture distribution of original data and new data. The presence of the original data creates a stable signal, and the coefficient α represents the mixture weight of the original data.
2. In recommender systems, watch time might not only depend on ranking position but also on an intrinsic video quality. The intrinsic quality creates a stable signal.

An important aspect of this setting is that the prediction only affects the outcome, not the features. This is called *outcome performativity.* The example of watch time prediction fits nicely into this setting.

Repeated risk minimization

Starting from some fixed function $f_0\colon \mathbb{R}^d \to \mathbb{R}$, consider the retraining dynamic defined by the optimization problem:

$$f_{t+1} = \operatorname*{argmin}_{f\colon \mathcal{X}\to\mathbb{R}} \mathop{\mathbb{E}}_{\mathcal{D}(f_t)} (f(X) - Y)^2$$

At each step, we find an unconstrained risk minimizer f_{t+1} on the distribution $\mathcal{D}(f_t)$ under the squared loss. Under what conditions does repeated risk minimization converge to a performatively stable point?

On a technical note, we always work over the Hilbert space $L^2(D_X)$ of measurable, square-integrable functions with the inner product $\langle f, g\rangle = \mathbb{E}_{D_X} f(X)g(X)$ and the associated norm $\|f\| = \sqrt{\langle f, f\rangle}$. This is a fancy way of saying that you can treat the functions as vectors in an inner product space and do linear algebra with them. Given this inner product, we may always assume that the initial point f_0 is orthogonal to f^* and that both have unit norm. Ensuring so will only affect the coefficients α, β.

For $\beta = 1$, the problem is trivial, since $f(X) - Y = \alpha f^*(X)$ and so the risk does not depend on the predictor. We will therefore always assume $\beta \neq 1$. Since we're minimizing squared loss, the optimal solution to the unconstrained risk minimization problem for $\beta \neq 1$ gives us the relation:

$$f_{t+1} = \alpha f^* + \beta f_t$$

We take equality to hold pointwise on the domain $\mathcal{X}$ except on a measure zero set. The update rule has a unique performatively stable point, corresponding to a scaling of the stable signal.

Claim 1. *For $\beta \neq 1$, the objective has a unique performatively stable point*

$$f_{\mathrm{PS}} = \frac{\alpha}{1-\beta} f^* .$$

Moreover, f_{PS} is also performatively optimal.

Proof. The claim follows by inspecting the optimal solution and verifying that f_{PS} achieves 0 risk.

□

Given that there exists a unique performatively stable point, will repeated risk minimization converge to it? To answer this question, it will be helpful to analyze the tangent of the angle between two functions f, g, thought of as vectors with the inner product $\langle f, g\rangle$. The inner product allows to define the trigonometric functions

$$\tan(f, f') = \frac{\sin(f,g)}{\cos(f,g)}, \quad \cos(f,g) = \frac{\langle f, g\rangle}{\|f\|\|g\|}, \quad \sin(f,g) = \frac{\sqrt{\|f\|^2\|g\|^2 - \langle f, g\rangle^2}}{\|f\|\|g\|}.$$

Being a ratio of sine and cosine, the tangent is often convenient to work with. A simple argument gives a geometric—also called *linear*—convergence rate in terms of the tangent.

Claim 2. *Let $\beta \neq 1$. Then, for every $t \geq 0$,*

$$\tan(f_{t+1}, f^*) = \frac{\beta^t}{\alpha} \cdot \frac{1-\beta}{1-\beta^t}.$$

In addition, for $|\beta| < 1$, $\|f_t - f_{\mathrm{PS}}\| \to 0$ at a linear rate as $t \to \infty$.

Proof. The function f_{t+1} lies in the linear span of f_0 and f^*. Recall that we assume f_0 and f^* are orthonormal. For $\beta \neq 1$, we have $f_{t+1} = a_t f^* + b_t f_0$ with coefficients:

$$a_t = \alpha \cdot \frac{1-\beta^t}{1-\beta}$$
$$b_t = \beta^t$$

Therefore, $\tan(f_{t+1}, f^*) = b_t/a_t$. For $|\beta| < 1$, we have $\lim_{t\to\infty} a_t = \alpha/(1-\beta)$ and $\lim_{t\to\infty} b_t = 0$. The claim follows.

□

For $|\beta| > 1$ the coefficients of the solution grow out of bounds. However, when $\beta > 1$, the tangent $\tan(f_{t+1}, f^*)$ still converges to $(\beta - 1)/\alpha$. In other words, the angle between the iterate f_t and the stable signal f^* stabilizes at some fixed value.

For bounded negative performative effects $\beta \in (-1, 0)$, the iterates will oscillate around the stable signal f^* while decreasing the angle with the stable signal.

Handling large performative effects

We can go a step further and achieve convergence for any arbitrarily large performative effect $\beta > 0$. The only thing we need is to add a regularization

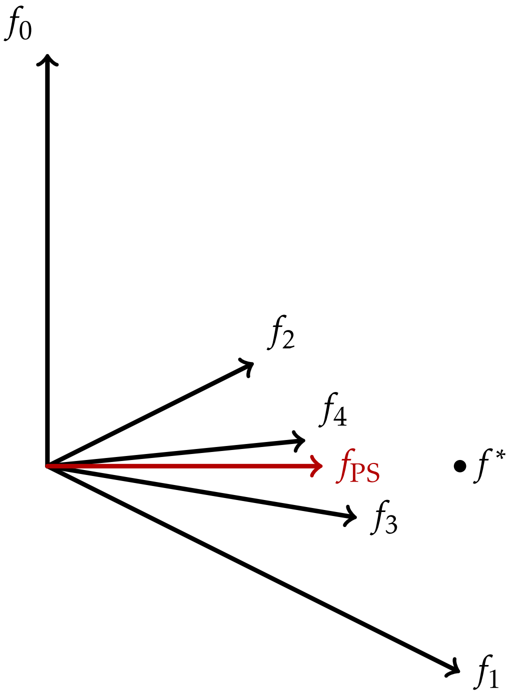

Figure 13.2: Oscillating convergence of repeated risk minimization under negative performative effects. Here, $\alpha = 1, \beta = -0.5$.

term to the objective:

$$f_{t+1} = \operatorname*{argmin}_{f\colon \mathcal{X}\to\mathbb{R}} \mathop{\mathbb{E}}_{D(f_t)} \left[(f(X)-Y)^2 + \delta f(X)^2\right]$$

The ℓ_2-regularized objective has the optimal solution (for $\delta \neq \beta - 1$):

$$f_{t+1} = \frac{\alpha}{1+\delta} f^* + \frac{\beta}{1+\delta} f_t$$

Provided that we choose $\delta > \beta - 1$, we have $\beta' = \beta/(1+\delta) < 1$ and so the previous analysis guarantees convergence to the performatively stable point

$$f_{\mathrm{PS}} = \frac{\alpha}{(1+\delta)(1-\beta')} f^* = \frac{\alpha}{1-\beta+\delta} f^*.$$

We can eliminate the regularization parameter δ by normalizing the function f_t to have unit norm after each update. This leads to the recurrence:

$$f_{t+1} = \alpha f^* + \beta \frac{f_t}{\|f_t\|}$$

Any fixed point $\bar{f}$ of this relation must be collinear with f^* so that $\bar{f} = r f^*$. Moreover, we must have $r = \alpha + \beta \mathrm{sign}(r)$. This implies:

- For $\beta \leq -\alpha$, there are no fixed points.
- For $-\alpha < \beta \leq \alpha$, there is a unique fixed point $(\alpha+\beta)f^*$.
- For $\beta > \alpha$, there are two fixed points $(\alpha+\beta)f^*$ and $(\alpha-\beta)f^*$.

Next we show that repeated risk minimization with the normalization step converges to a performatively stable point at a linear rate for any arbitrarily large performative effect $\beta \geq 0$ so long as $\alpha > 0$.

Claim 3. *Let $\alpha > 0$ and $\beta \geq 0$. For $t \geq 1$, we have*

$$\tan(f_{t+1}, f^*) \leq \frac{\beta}{\alpha}\left(\frac{\beta}{\alpha+\beta}\right)^t.$$

Moreover, f_t converges to the performatively stable point $(\alpha+\beta)f^*$.

Proof. Without loss of generality, we may take $f^* = \begin{pmatrix}1\\0\end{pmatrix}$ and $f_0 = \begin{pmatrix}0\\1\end{pmatrix}$. By definition, $f_1 = \begin{pmatrix}\alpha\\\beta\end{pmatrix}$. Therefore $\tan(f_1, f^*) = \beta/\alpha$. Now, let $\frac{f_t}{\|f_t\|} = \begin{pmatrix}a_t\\b_t\end{pmatrix}$ so that $\tan(f_t, f^*) = \frac{b_t}{a_t}$. Then, for $t \geq 1$, we have

$$\tan(f_{t+1}, f^*) = \frac{\beta b_t}{\alpha + \beta a_t} = \frac{\beta a_t}{\alpha+\beta a_t} \cdot \tan(f_t, f^*) \leq \frac{\beta}{\alpha+\beta} \cdot \tan(f_t, f^*),$$

since $0 < a_t \leq 1$, noting that $a_t > 0$ because $\alpha > 0$. In particular, the argument shows that $f_t/\|f_t\|$ converges to f^*. Therefore, it must be the case that f_t converges to $(\alpha+\beta)f^*$.

□

What these results show is that repeated risk minimization can converge even if the performative effect is arbitrarily large. What's needed, though, is the existence of a stable signal that remains invariant under model deployment. Eventually, repeated risk minimization must settle on the stable signal—as if for lack of alternatives. At least, that's the lesson here.

Above we considered *non-parametric* risk minimization, optimizing over all functions. An analogous analysis applies to linear models where $f^*(X) = \langle\theta^*, X\rangle$ and $f_t(X) = \langle\theta_t, X\rangle$. Provided the marginal distribution on X is isotropic, $\mathbb{E}XX^T = I$, we get the same recurrence relationship in terms of the parameter vectors θ_t and θ^*.

13.4 *Data feedback loops*

Repeated risk minimization is a mathematically precise instance of a *data feedback loop*. In a data feedback loop, data and model evolve in tandem over time. The model generates, augments, labels, or otherwise modifies a data source. In turn, the data are continually used to train or update the model. Performativity is the key characteristic of all data feedback loops: the data depends on the model.

Chatbots have fueled academic interest in data feedback loops in recent years. If ChatGPT really loves *em dashes*—and it sure looks like it—we can expect people to use more of them in all their trade books about AI. Trained on those books, newer versions of ChatGPT might then recommend even more em dashes. Soon, em dashes will crowd out all whitespace and punctuation on the internet. Seriously though, language models as chatbots and assistants do create a feedback loop between the model and the data on the internet. There's been much speculation, and some fear, about what this feedback loop might do.

It's inherently difficult to experiment with data feedback loops. Performativity is often a phenomenon of scale. It results from applying models with sufficient reach over sufficiently long time horizons. There's no simple experiment that tells us the long run outcome of a data feedback loop. As a result, researchers have proposed several mathematical models of data feedback loops.

Stories about data feedback loops

Mathematically, data feedback loops are dynamical systems maintaining a state consisting of data and a model. Both evolve over time according to update rules that depend on the application.

Polarization, echo chambers, and filter bubbles. Recommender systems always create a data feedback loop. Engagement signals—such as clicks, likes, or shares—always depend on the model that recommends content to platform users. The most basic feedback loop of algorithmic content curation is that people click on what the platform recommends and platforms recommend what people click on.[8] Many scholars have feared that this self-reinforcing feedback loop leads to some kind of content homogenization or collapse in content diversity. *Polarization*, *echo chambers*, and *filter bubbles* are some of the concerns that are often attributed to feedback loop on digital platforms.[9] But there are also feedback loops between the platform and the *supply side*, the content creators.[10] Decisions about algorithmic curation create specific incentives for content creators, thus influencing future data.

Feedback loops are also inherent to the digital advertising ecosystem.[11,12] An advertising platform never sees clicks for ads it hasn't shown, a problem that can partially be addressed with experiments, A/B tests, and causal inference.[11,13] But there are many effects that small-scale randomized experiments can't anticipate. Suppose a platform experiments with showing three ads, instead of two, on a page in a small randomized experiment. Not knowing about this experiment, advertisers will still bid as if there were only two ads per page. And so the revenue for the platform will likely increase with three ads in the experimental group. Were the change rolled out to the whole platform, the market equilibrium would change. Advertisers will then likely bid lower, reflecting the fact that attention is now split across one more ad, thus diminishing the value of an ad.

Fairness and welfare dynamics. Suppose a governments predicts criminal activity in a neighborhood from the number of arrests made in that neighborhood in the last month. *Predictive policing* is a controversial policy that assigns additional police force based on high predicted criminal activity. Assigning additional officers to a neighborhood, however, will likely result in additional arrests made. The increase in police force therefore also leads to an increase in arrests, which in turn leads to higher predicted criminal activity in the future. Predictive policing therefore creates a positive feedback loop between police force allocation and observed criminal activity.[14] This dynamic amplifies initial differences in arrest rates between different neighborhoods. The welfare implication is that some neighborhoods end up being overpoliced and criminalized relative to other neighborhoods. In addition, the self-reinforcing nature of the feedback loop can give proponents of predictive policing a false sense of accuracy of the solution.

The problem isn't specific to predictive policing. Feedback loops are common in lending, for example. If a bank estimates that a customer has a high default risk, they will assign a higher interest rate to any loan, thereby increasing the financial burden on the borrower. More broadly, research on algorithmic fairness provides different *criteria* for evaluating the fairness properties of a predictor. The meaning and interpretation of these fairness criteria changes significantly once you take performative effects into account.[15]

Model collapse? What happens when ChatGPT trains on *synthetic data* generated by its former selves? In 2023, a group of researchers issued a stark warning: *AI models collapse when trained on recursively generated data.*[16] The proclamation is the upshot of a simple theoretical analysis. Train a generative model on a distribution, generate new data, train another model on that data, and repeat indefinitely. If you take a simple generative model and a diverse initial distribution, the model will forget different parts of the initial distribution and ultimately "collapse" onto one or a few modes the distribution.

Collapse in this setting is largely a consequence of two factors. First, the model's limited capacity prevents it from faithfully representing the training distribution. In the more recent data-constrained compute regime (see Chapter 10), where compute resources exceed data resources, it's less clear that model capacity would be such a strong bottleneck. Second, the model is trained exclusively on synthetically generated data. Mixing in sufficient data from previous iterations turns out to prevent model collapse.[17]

Bias amplification in data labeling. In a similar vein, we can ask what happens if a model trains on data *annotated* by the model itself? To lift the capacity constraint on the model, it makes sense to look at what happens for Bayes optimal predictors. Since we're talking about labeling only, we assume the marginal distribution X is fixed throughout. The initial distribution $\mathcal{D}_0 = (X, Y)$ consists of accurately labeled data.

Consider the following dynamic, repeated for each step $t \geq 1$:

- Find the Bayes optimal predictor f_t on the distribution $\mathcal{D}_{t-1}$.
- Add a batch of new samples drawn from the distribution (X, Y_t) where $Y_t = f_t(X)$ is the prediction of model f_t.

To make this more concrete, suppose the initial distribution ranges over labeled images. There are images of women wearing scrubs in a hospital environment labeled either *nurse* or *doctor*. If the model f_t is an accuracy-maximizing classifier, it will pick the majority label, hypothetically, *nurse*. This implies that all additional labels will be *nurse* and the percentage of such images labeled *nurse* will steadily increase in each iteration. If the model f_t, however, is Bayes optimal under, say, the cross entropy loss, it will faithfully represent the label proportion of the original distribution. Of course, optimality is not necessary. It's sufficient for the model to be calibrated in the sense that it preserves the correct label frequencies on average.[18]

Cautionary tale about cautionary tales

Stories about data feedback loops are interesting and helpful as cautionary tales. They illustrate hypothesized aspects of model deployment in the real world through stylized theoretical models. The worst-case scenarios suggested by some of these dynamical models, however, are likely overblown. Recommendation engines and digital advertising, to give two examples, always operate in data feedback loops. Clicks are model-dependent. It's a real issue that engineers grapple with. But there isn't strong evidence for model collapse in recommender systems or the advertising ecosystem. This may in part be due to randomness in the ecosystem.[19]

Mathematically, the state of dynamical models typically either explodes or collapses as you iterate the dynamics. This is a more complicated version of the basic fact that the exponential a^t either goes to zero or infinity as you grow the exponent t, except in special cases. It's no surprise to see extreme outcomes of stagnation or collapse in dynamical models. As illustrative models of social processes, dynamical system models are often one-trick ponies. They demonstrate an extreme effect as a kind of cautionary tale.

This much is, in fact, a forgotten lesson of the 1970s, the heyday of the once flourishing discipline of *system dynamics*. Like the models we've just seen, system dynamics modeled social processes as formal dynamical systems. Forecasts and policy recommendations derived from numerical simulations of these systems. The system modelers were ambitious, taking on industrial organizations, cities sprawl, and ultimately the whole world. The most well-known model of the era was *World3*, a dynamic model of global industrial output, population growth, food production, and climate impacts.[20]

Although the world may still be on track for its predicted collapse between 2030 and 2050, system dynamics were plagued by numerous problems, scientific validity central among them. In particular, it was difficult to scientifically justify the transitions, weights, initial conditions, and various modeling choices. System modelers instead typically *winged it* with intuition and confidence. The historian of science Kevin Baker found that system dynamics founder Jay Forrester had, in fact, drawn the precursor to World3 on a queen-sized bed sheet:

> Emblazoned on the sheet was an assortment of rectangles, circles, and other shapes, all connected by a chaotic latticework of lines. All these symbols had been hand-drafted in black marker and labeled with the neat, gothic lettering distinctively used by engineers.[21]

Bed sheets aside, Forrester was known for his disregard for domain expertise that he did not keep a secret.[21] Like World3, system dynamics would often show the same qualitative behavior regardless of initial conditions, transitions, or weights. The lack of scientific foundations and unstable dynamics, in turn, led to dubious policy recommendations.[21,22]

13.5 Dynamic benchmarks

Where data feedback loops point at *vicious* cycles, dynamic benchmarks try to create *virtuous* cycles. The basic idea is simple: Start from an initial dataset, let model builders compete, then add failure cases of the best models to the data pool, and continue. The dataset in a dynamic benchmark is model-dependent, as data and models evolve in tandem. Any dynamic benchmark is therefore, by definition, a data feedback loop. The hope is that as models get better, the benchmark becomes more challenging, thus leading to continual improvement.

In 2021, Meta launched *Dynabench*,[23] a platform to create and host dynamic benchmarks:

> We believe the time is ripe to radically rethink the way we do benchmarking. With Dynabench, we can collect human-in-the-loop data dynamically, against the current state-of-the-art, in a way that more accurately measures progress, that cannot saturate and that can automatically fix annotation artifacts and other biases over time.[24]

Similar ideas also come up in AI safety under the idea of *red teaming*. In red teaming, companies pay data workers to find failure cases of a model. These failure cases are then added in some form to the post-training pipeline.

Do dynamic benchmarks work? Empirical results on Dynabench show that indeed dynamic benchmarks can lead to higher accuracy than a static dataset in some cases.[25] But not everyone has been optimistic about the general idea. Researchers Bowman and Dahl caution:

> Given a source of examples and a model, adversarial-filtering-style approaches build a benchmark based on samples from that source for which the model fails. Adversarial filtering can remove examples that are easy due to trivial artifacts, but it does not ensure that the resulting dataset supports a valid test of model ability, and it can systematically eliminate coverage of linguistic phenomena or skills that are necessary for the task but already well-solved by the adversary model. This mode-seeking (as opposed to mass covering) behavior by adversarial filtering, if left unchecked, tends to reduce dataset diversity and thus make validity harder to achieve.[26]

In other words, dynamic benchmarks might lead to the kind of mode collapse we know from data feedback loops. Rather than covering a target domain broadly, dynamic benchmarks might end up putting too much weight on a few modes of the distribution.

A formal model of dynamic benchmarks

To better understand the promise of dynamic benchmarks, it makes sense to first define formally what a dynamic benchmark is and what exactly it's trying to achieve. The standard dynamic benchmark alternates model building and adversarial data collection. We can generalize this idea a fair bit by allowing a dynamic benchmark to be any directed acyclic graph in which each node represents one of the following operations:

- **Model building:** Given a data distribution, the community finds a model that minimizes risk on the current distribution up to an $\epsilon > 0$ error. Here, $\epsilon > 0$ is fixed parameter, say, 0.2.

- **Model ensembling:** Given a collection of models, we can ensemble them whichever way we want.
- **Adversarial Data collection:** Given a model, data workers sample the uniform distribution over the failure cases (error set) of the model.
- **Data pooling:** Given multiple data sources, we can combine them with arbitrary weights.

Defining the goal of a dynamic benchmark is a bit subtle. We already assumed that we can invoke model builders that'll find a risk minimizer up to error $\epsilon > 0$ on any distribution. So, the goal has to be more ambitious. Assume there is a target distribution $\mathcal{D}^*$ that contains all instances of interest. If we could learn $\mathcal{D}^*$ perfectly with no error, we'd have a perfect model as far as we're concern. Our goal is to set up a dynamic benchmark that produces a sequence of models $f_0, f_1, \ldots$, so that the error of f_t on $\mathcal{D}^*$ quickly converges to 0 as t grows.

Focusing on binary classification and classification error as the loss function, the standard design of a dynamic benchmark achieves error $O(\epsilon^2)$ after three rounds.

Proposition 1. *Let f_0, f_1, f_2 be the sequence of models produced by a standard dynamic benchmark with three rounds. Then, the the ensemble classifier*

$$f = \mathrm{Majority}(f_0, f_1, f_2)$$

achieves error $O(\epsilon^2)$.

Three rounds of dynamic benchmarking can get the error down from ϵ to ϵ^2. Unfortunately, in the worst-case this is where things stall.

Proposition 2. *For any number $t > 2$ of rounds, there is a problem instance such that the ensemble classifier*

$$f = \mathrm{Majority}(f_0, ..., f_{t-1})$$

resulting from a t-round standard dynamic benchmark has error at least $\Omega(\epsilon^2)$.

That is, progress in a standard dynamic benchmark can stall after three rounds in the worst case! More delicate designs that use the generality of the definition can guarantee strictly more progress.[27] But these benchmark designs may also be hard to implement in practice.

Notes

Aimed at a general audience, Jacob Ward's *The Loop* gives an excellent account of feedback loops and performativity in digital platforms.[28]

The framework of performative prediction is due to Perdomo et al.,[29] who also gave conditions under which repeated risk minimization converges to a stable point. The conditions require the loss to be smooth, strongly convex, and the distribution map to be sufficiently insensitive to changes in model parameters. In particular, this implies that the performative effects are small. For an entry point to performative prediction, see the primer by Hardt and Mendler-Dünner.[30] The result on repeated risk minimization in this chapter is from joint work with Jiduan Wu.[31]

The material about Morgenstern comes from his 1928 habilitation written in German.[4] Fifty years later, Morgenstern remarked: "The problems touched in that book never left me."[32]

Data feedback loops, system dynamics, dynamic benchmarks Ensign et al. studied runaway feedback loops in predictive policing.[14] Jiang et al. study a model of degenerate feedback loops in recommender systems.[9] Cheng et al. study a dynamic model of recommender systems and show how to identify performative effects from exogenous variation without explicit randomization.[8] What exactly YouTube's recommendation model does today is not publicly known. The example of watch time prediction stems from a 2016 Google paper about applying deep learning to YouTube recommendations.[33] Turning to home price prediction, Malik and Manzoor ask if machine learning can amplify pricing errors in the housing market?[34] There has been much recent work on data feedback loops motivated by large language models.[16,17,35] The setting of bias amplification in data labeling is due to Taori and Hashimoto.[18]

Forrester published extensively on system dynamics, notably, the sequence *Industrial Dynamics*[36], *Urban Dynamics*[37], and *World Dynamics*[38], imagining a future of another fifty years of system dynamics in 2007.[39] Although largely forgotten by computer scientists, system dynamics was hugely influential. Nobel Prize-winning economist William Nordhaus, for example, published extensively in response to Forrester, proposing a dynamic climate model, called DICE, based on empirical evidence.[40] System dynamics also lives on as *systems thinking*, popular in business and management circles.[41] Finally, Will Wright coded up *Urban Dynamics* when creating the video game *Sim*

City in 1985.

The model and results about dynamic benchmarks are from work by Shirali et al.[27]

Bibliography

1. Wald, A. Contributions to the Theory of Statistical Estimation and Testing Hypotheses. *Annals of Mathematical Statistics* **10,** 299–326 (1939) (↑ 2).
2. Wald, A. *Statistical Decision Functions* (John Wiley & Sons, New York, 1950) (↑ 2).
3. Von Neumann, J. & Morgenstern, O. *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, NJ, 1944) (↑ 3).
4. Morgenstern, O. *Wirtschaftsprognose: Eine Untersuchung ihrer Voraussetzungen und Möglichkeiten* (Springer, 1928) (↑ 4, 23).
5. Grunberg, E. & Modigliani, F. The Predictability of Social Events. *Journal of Political Economy* **62,** 465–478 (1954) (↑ 5).
6. Simon, H. A. Bandwagon and Underdog Effects and the Possibility of Election Predictions. *The Public Opinion Quarterly* **18,** 245–253 (1954) (↑ 5).
7. Brouwer, L. E. J. Über Abbildung von Mannigfaltigkeiten. *Mathematische Annalen* **71,** 97–115 (1911) (↑ 5).
8. Cheng, G., Hardt, M. & Mendler-Dünner, C. *Causal inference out of control: estimating performativity without treatment randomization* in *International Conference on Machine Learning (ICML)* (2024) (↑ 17, 23).
9. Jiang, R., Chiappa, S., Lattimore, T., György, A. & Kohli, P. *Degenerate feedback loops in recommender systems* in *AAAI/ACM Conference on AI, Ethics, and Society (AIES)* (2019), 383–390 (↑ 17, 23).
10. Jagadeesan, M., Garg, N. & Steinhardt, J. *Supply-side equilibria in recommender systems* in *Neural Information Processing Systems (NeurIPS)* (2023), 14597–14608 (↑ 17).
11. Bottou, L. *et al.* Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research* **14,** 3207–3260 (2013) (↑ 17).
12. Hwang, T. *Subprime attention crisis: Advertising and the time bomb at the heart of the Internet* (FSG originals, 2020) (↑ 17).
13. Kohavi, R., Tang, D. & Xu, Y. *Trustworthy online controlled experiments: A practical guide to a/b testing* (Cambridge University Press, 2020) (↑ 17).
14. Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C. & Venkatasubramanian, S. *Runaway feedback loops in predictive policing* in *Conference on fairness, accountability and transparency* (2018), 160–171 (↑ 18, 23).
15. Liu, L. T., Dean, S., Rolf, E., Simchowitz, M. & Hardt, M. *Delayed impact of fair machine learning* in *International Conference on Machine Learning (ICML)* (2018), 3150–3158 (↑ 18).

16. Shumailov, I. *et al.* AI models collapse when trained on recursively generated data. *Nature* **631,** 755–759 (2024) (↑ 18, 23).
17. Gerstgrasser, M. *et al.* Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv:2404.01413* (2024) (↑ 18, 23).
18. Taori, R. & Hashimoto, T. *Data feedback loops: Model-driven amplification of dataset biases* in *International Conference on Machine Learning (ICML)* (2023), 33883–33920 (↑ 19, 23).
19. Hummel, P. & McAfee, R. P. Machine learning in an auction environment. *Journal of Machine Learning Research* **17,** 1–37 (2016) (↑ 19).
20. Meadows, D. & Randers, J. *The limits to growth: the 30-year update* (Routledge, 2012) (↑ 20).
21. Baker, K. T. *World Processors Computer Modeling, the Limits to Growth, and the Birth of Sustainable Development* PhD thesis (Northwestern University, 2019) (↑ 20).
22. Baker, K. T. Model Metropolis. *Logic(s) Magazine.* Accessed: 2025-10-25. https://logicmag.io/play/model-metropolis/ (6 Jan. 2019) (↑ 20).
23. Kiela, D. *et al.* Dynabench: Rethinking benchmarking in NLP. *arXiv:2104.14337* (2021) (↑ 20).
24. Dynabench. *About* https://dynabench.org/about. Accessed: 2025-10-25 (↑ 21).
25. Wallace, E., Williams, A., Jia, R. & Kiela, D. Analyzing dynamic adversarial training data in the limit. *arXiv:2110.08514* (2021) (↑ 21).
26. Bowman, S. R. & Dahl, G. E. What will it take to fix benchmarking in natural language understanding? *arXiv:2104.02145* (2021) (↑ 21).
27. Shirali, A., Abebe, R. & Hardt, M. *A Theory of Dynamic Benchmarks* in *International Conference on Learning Representations (ICLR)* (2023) (↑ 22, 24).
28. Ward, J. *The loop: How technology is creating a world without choices and how to fight back* (Grand Central Publishing, 2022) (↑ 23).
29. Perdomo, J. C., Zrnic, T., Mendler-Dünner, C. & Hardt, M. *Performative Prediction* in *International Conference on Machine Learning (ICML)* (2020) (↑ 23).
30. Hardt, M. & Mendler-Dünner, C. Performative prediction: Past and future. *arXiv:2310.16608* (2023) (↑ 23).
31. Hardt, M. & Wu, J. *Repeated risk minimization converges to stable signals* Manuscript. 2025 (↑ 23).
32. Morgenstern, O. The collaboration between Oskar Morgenstern and John von Neumann on the theory of games. *Journal of Economic Literature* **14,** 805–816 (1976) (↑ 23).
33. Covington, P., Adams, J. & Sargin, E. *Deep neural networks for youtube recommendations* in *ACM Conference on Recommender Systems (RecSys)* (2016), 191–198 (↑ 23).
34. Malik, N. & Manzoor, E. Can Machine Learning Amplify Pricing Errors in Housing Market?: Economics of ML Feedback Loops. *Economics of ML Feedback Loops (September 18, 2020)* (2020) (↑ 23).
35. Dohmatob, E., Feng, Y., Subramonian, A. & Kempe, J. Strong model collapse. *arXiv:2410.04840* (2024) (↑ 23).
36. Forrester, J. W. Industrial dynamics. *Journal of the Operational Research Society* **48,** 1037–1041 (1997) (↑ 23).
37. Forrester, J. W. Urban dynamics. *IMR; Industrial Management Review (pre-1986)* **11,** 67 (1970) (↑ 23).
38. Forrester, J. W. *World dynamics* (Wright-Allen Press, 1971) (↑ 23).
39. Forrester, J. W. System dynamics—the next fifty years. *System Dynamics Review: The Journal of the System Dynamics Society* **23,** 359–370 (2007) (↑ 23).

40. Nordhaus, W. *A question of balance: Weighing the options on global warming policies* (Yale University Press, 2008) (↑ 23).

41. Meadows, D. *Thinking in systems: International bestseller* (Chelsea Green Publishing, 2008) (↑ 23).