Social Science without Generalization

Drew Dimmery — Hertie School

2026-04-02

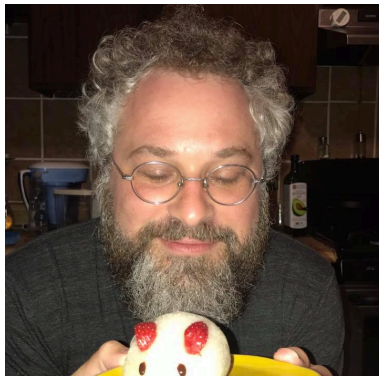
Me

- PhD in Political Methodology from **NYU**
  - Experimentation, Causal Inference, Machine Learning
- **Facebook Core Data Science** (now "Meta Central Applied Science")
  - Adaptive Experimentation team
  - Bayesian Optimization
  - Routine Experimentation
- **Academia**: Hertie School
  - Professor of Data Science for the Common Good
- **Research**:
  - Better integration of machine learning and experiments
  - Causal Machine Learning
  - Helping people study the internet using these tools

People not to blame for this odd talk

Kevin Munger

David Arbour



Eytan Bakshy

Molly Offer-Westort

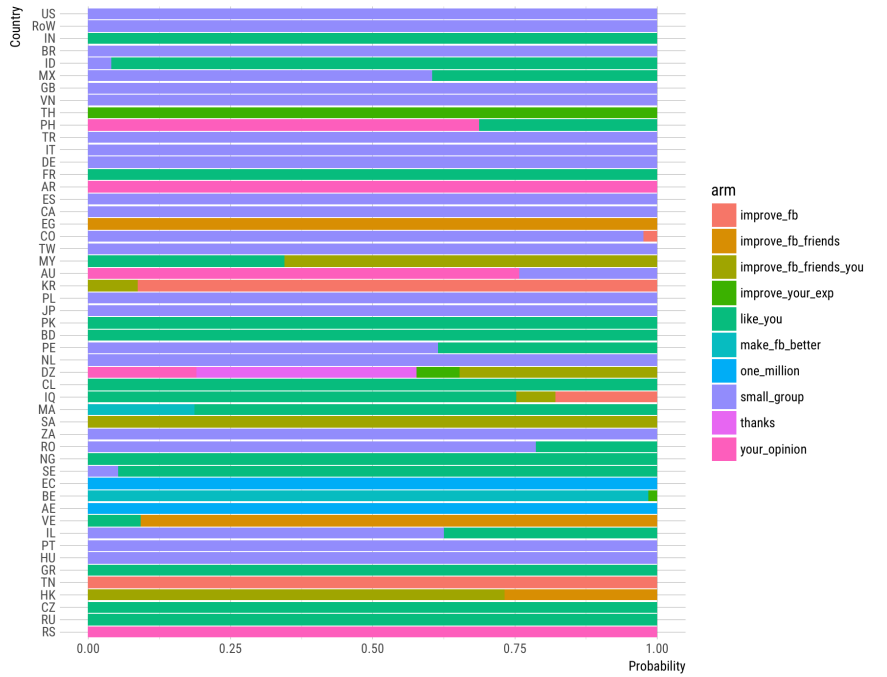
A Parable of Tech

A parable of tech: The beginning

- Internship, 2015
- Optimize survey recruitment!
- RCT
  - $>$ 1M units
  - 200 countries
  - 32 languages
  - 9 treatments

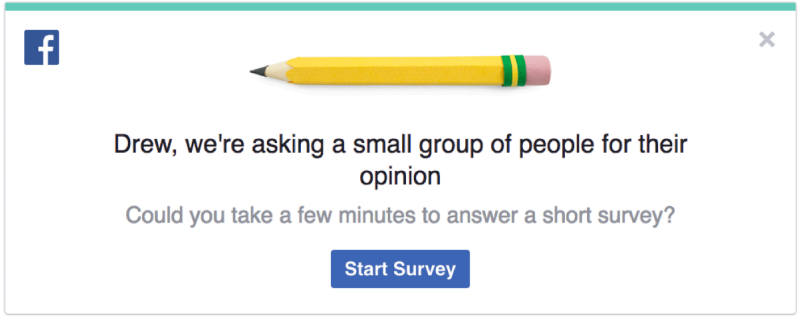
A parable of tech: Results

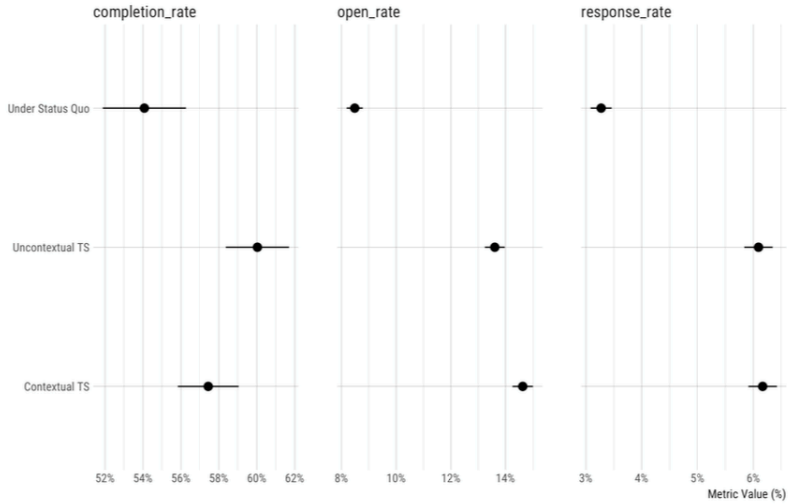
- Clear heterogeneity abounds!
- HTEs correlate to World Values Survey
  - Communal versus individualistic societies
- Smells like a paper for the tabloids!



A parable of tech: Repeat

- 2017
- New survey infrastructure
- New, bigger RCT
  - 56 languages
  - 16 treatments
- Update the models!
- …No heterogeneity
  - Huh?





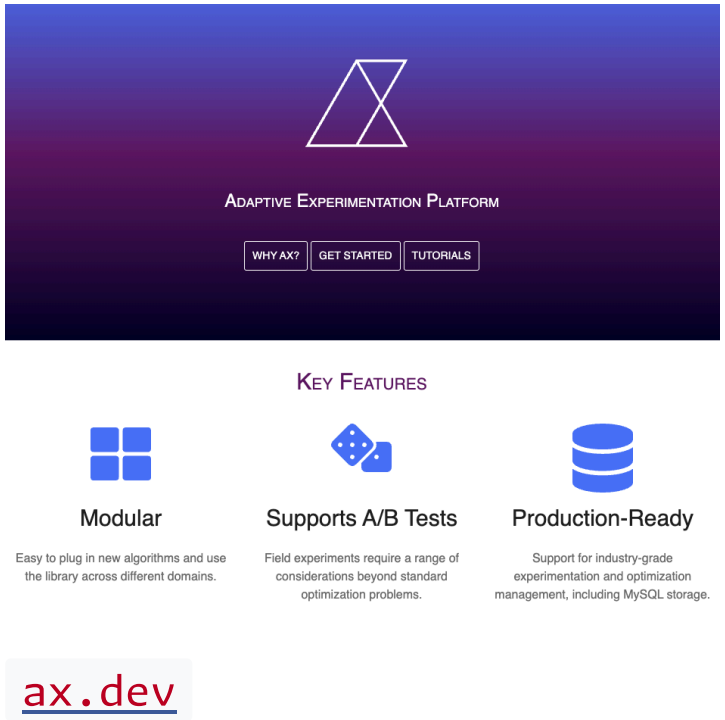
A parable of tech: Moral

- Knowledge can be mendacious
- To get things right, *keep testing them*
- Tools must make it cheap (both in dollars and time)
  - Maybe only really possible in digital systems / code
- …AI?

Dimmery, Bakshy and Sekhon (2019) *Shrinkage Estimators in Online Experiments* SIGKDD
Wu, Tan, Li, Garrard, Obeng, Dimmery, Singh, Wang, Jiang & Bakshy (2022) *Interpretable personalized experimentation* SIGKDD

A parable of tech: Moral

- Knowledge can be mendacious
- To get things right, *keep testing them*
- Tools must make it cheap (both in dollars and time)
  - Maybe only really possible in digital systems / code
- …AI?



ax.dev

Dimmery, Bakshy and Sekhon (2019) *Shrinkage Estimators in Online Experiments* SIGKDD
Wu, Tan, Li, Garrard, Obeng, Dimmery, Singh, Wang, Jiang & Bakshy (2022) *Interpretable personalized experimentation* SIGKDD

A Second Parable: US 2020

The Facebook/Instagram Election Study

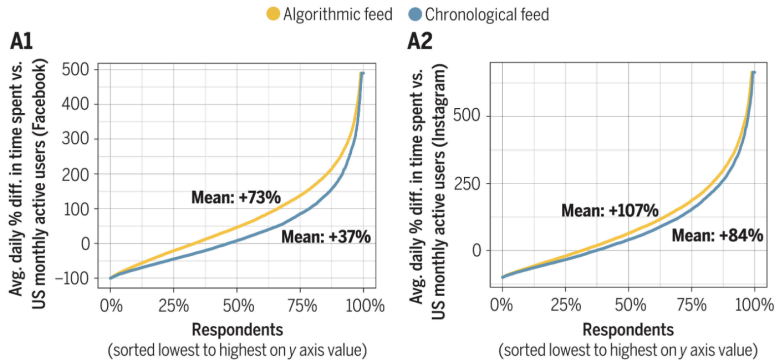
- The largest social media experiment ever conducted
- 2020 US Presidential Election
- Industry–academic collaboration
  - ~20 external researchers, ~10 internal
- **16** pre-registered studies
  - No pre-publication approval by Meta
  - Academic lead authors with **final control rights**
  - Independent rapporteur; data release for replication
- Manipulated core features of Facebook and Instagram
  - Feed algorithms, resharing, political content

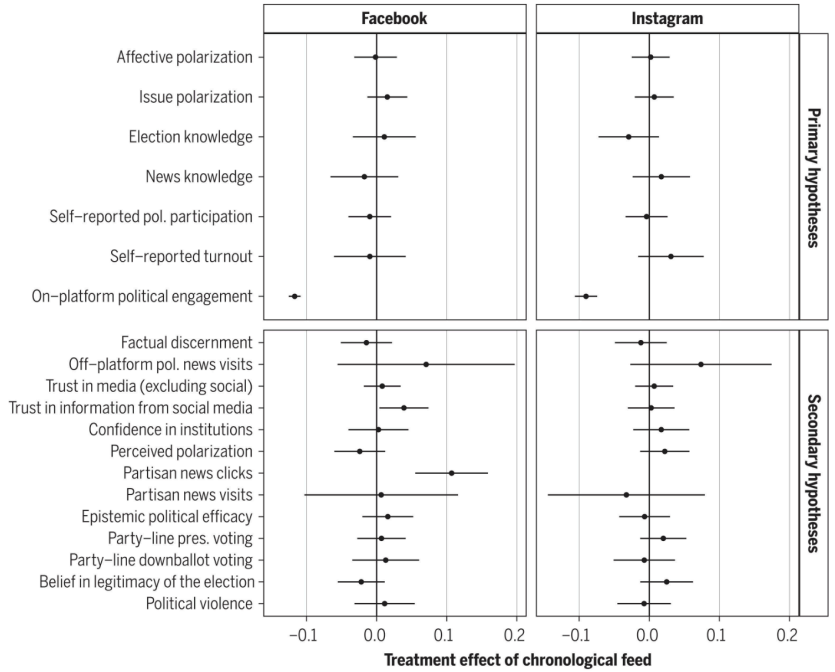
Guess et al. (2023) *How do social media feed algorithms affect attitudes and behavior in an election campaign?* Science
Nyhan et al. (2023) *Like-minded sources on Facebook are prevalent but not polarizing* Nature

What the study found

Treatment: **chronological** vs. **algorithmic** feed

- Changed on-platform behavior **dramatically**
  - Time spent, content exposure
- **Null effects** on attitudes and beliefs
  - Polarization, knowledge, democratic norms, vote choice

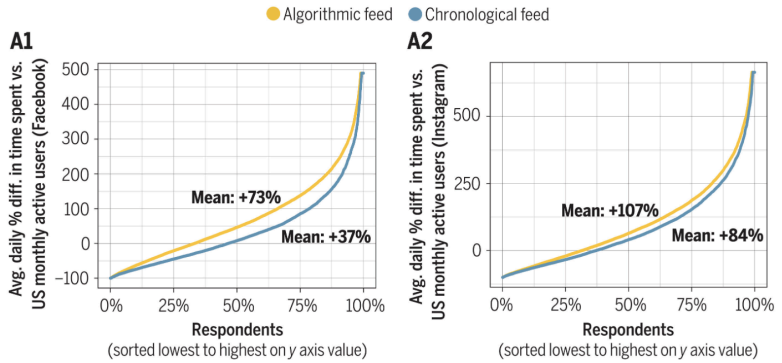


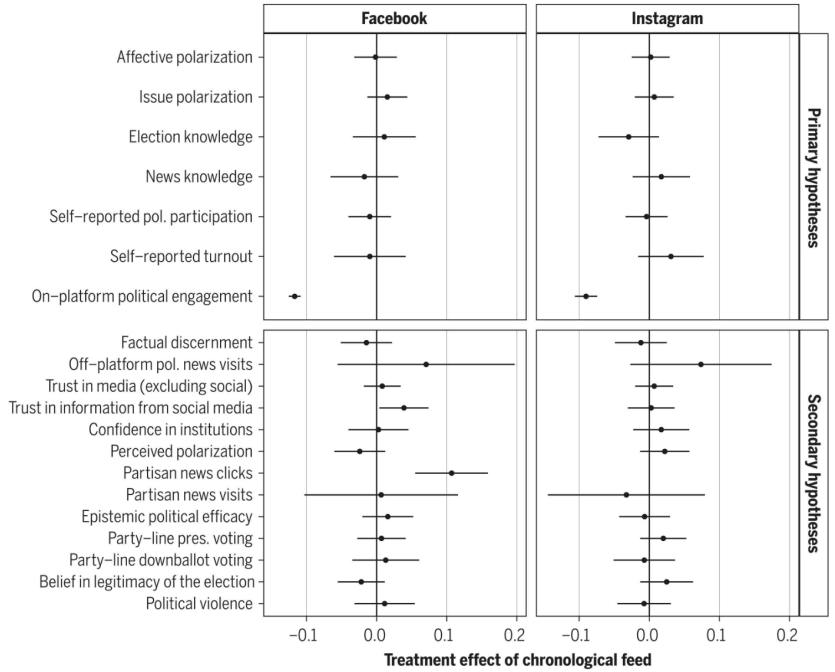


What the study found

Treatment: **chronological** vs. **algorithmic** feed

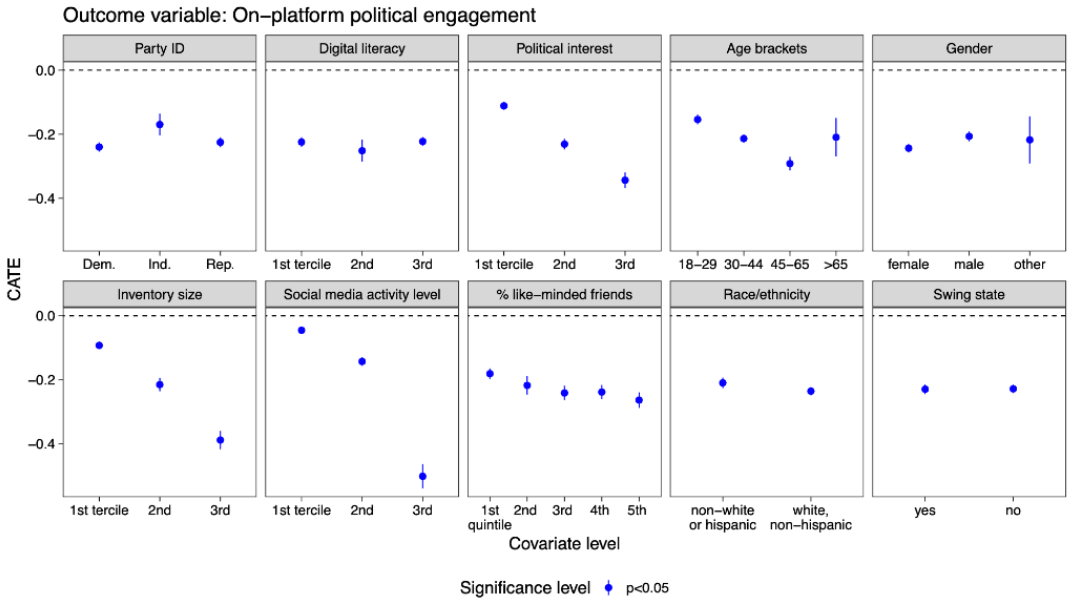
- Changed on-platform behavior **dramatically**
  - » Time spent, content exposure
- **Null effects** on attitudes and beliefs
  - » Polarization, knowledge, democratic norms, vote choice
- Are we missing countervailing effects?
- Or does the algorithm simply not matter for these outcomes?





But heterogeneity lurks beneath the averages

- On-platform effects vary sharply by **inventory size** (how much content a user's feed *could* show)
- Larger inventory ⇒ larger effects of the algorithm
- July 2022: Zuckerberg announces AI-recommended content will **more than double**



- ⇒ everyone's inventory is growing
- The 2020 estimates don't describe a 2023 Facebook, let alone a 2026 one

Dimmery (2023) *A Blueprint for the Regulation of Tech*

The “break-glass” problem

- Meta ran **thousands** of other experiments simultaneously
  - Internally, the comparison group is the **“status quo”**, not “control”
- During the election, Meta deployed **“break-glass” measures**
  - Emergency algorithmic changes to reduce misinformation
- These changed the platform **during** US2020
- Bagchi et al. (2024) eLetter: the experimental estimates were compromised

The “break-glass” problem

- Meta ran **thousands** of other experiments simultaneously
  - Internally, the comparison group is the **“status quo”**, not “control”
- During the election, Meta deployed **“break-glass” measures**
  - Emergency algorithmic changes to reduce misinformation
- These changed the platform **during** US2020
- Bagchi et al. (2024) eLetter: the experimental estimates were compromised

- Not corporate manipulation — **operational necessity**
- The platform changed **beneath** the experiment (the “status quo”)
- The pre-analysis plan couldn’t prevent this
- The real problem: we have **no reliable measures** of the break-glass effects themselves
  - Imagine the outcry if Facebook had randomized their delivery!

Two parables, one lesson

Two parables, one lesson

Parable 1

Translations improved between experiments

⇒ heterogeneity was an artifact of infrastructure, not culture

Two parables, one lesson

Parable 1

Translations improved between experiments

⇒ heterogeneity was an artifact of infrastructure, not culture

Parable 2

Platform changed during an experiment

⇒ estimates reflect a world that no longer exists

Two parables, one lesson

Parable 1

Translations improved between experiments

⇒ heterogeneity was an artifact of infrastructure, not culture

Parable 2

Platform changed during an experiment

⇒ estimates reflect a world that no longer exists

My lesson

The answer is not to demand that the world hold still while we study it, but to build systems that **keep experimenting faster than the world changes**.

What Are We Doing?

What are we doing?

Experiments are asked to do a lot:

1. **Estimate** a causal effect: RCTs do this with minimal assumptions Aronow et al. 2021
2. **Explain** why: identify mechanisms, "feeling the key turn in the lock" Ashworth, Berry and BdM 2021
3. **Generalize** beyond the study: build cumulative theory Latour and Woolgar 1979

What are we doing?

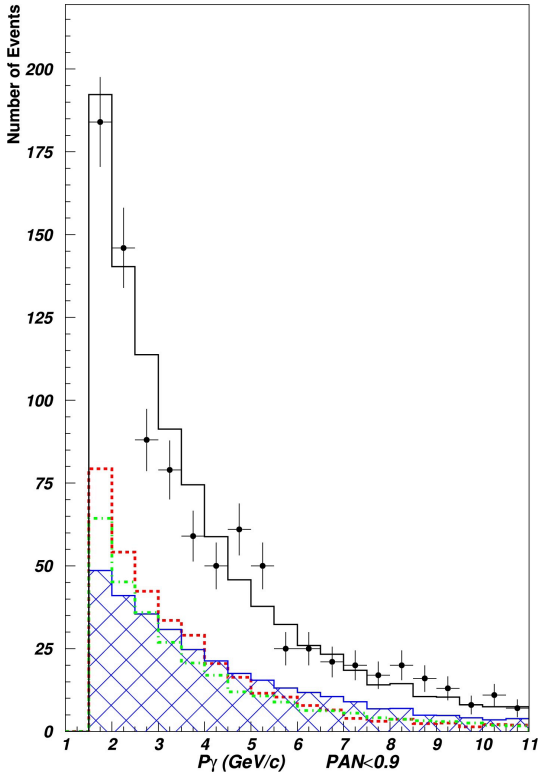
Experiments are asked to do a lot:

1. **Estimate** a causal effect: RCTs do this with minimal assumptions Aronow et al. 2021
2. **Explain** why: identify mechanisms, "feeling the key turn in the lock" Ashworth, Berry and BdM 2021
3. **Generalize** beyond the study: build cumulative theory Latour and Woolgar 1979

- In practice?
  - Mostly survey experiments, occasionally field experiments
  - Results are **published as arguments** Bem 2017
    - Popper's rejection of induction, taken to its logical extreme
  - Policy implications come later, if at all.
  - Theory does the generalizing — but can it?

How about natural science?

- Extremely **predictive** models
- Resistance and accommodation Pickering 1995
  - When theory meets data, something has to give
  - Correct the theory, revise beliefs about apparatus, modify the apparatus
  - Result: a "robust" fit
- Nature **pushes back hard** — wrong predictions fail visibly
  - This constrains the process enough to converge
- A social process (of convergence on [maybe] true things)
  Pickering 1984, Latour and Woolgar 1979, Kuhn 1962, Peirce 1868



Kullenberg, Mishra, Dimmery and the NOMAD Collaboration (2012) *A search for single photon events in neutrino interactions* Physics Letters B

Is social science learning true theories?

- Social science theories **don't converge** like this
- Representational power is **poor**
  - » Same game, different labs ⇒ wildly different results
- Multiple theories explain equally little
  - » What we have is **historically contingent**

Pickering 1995

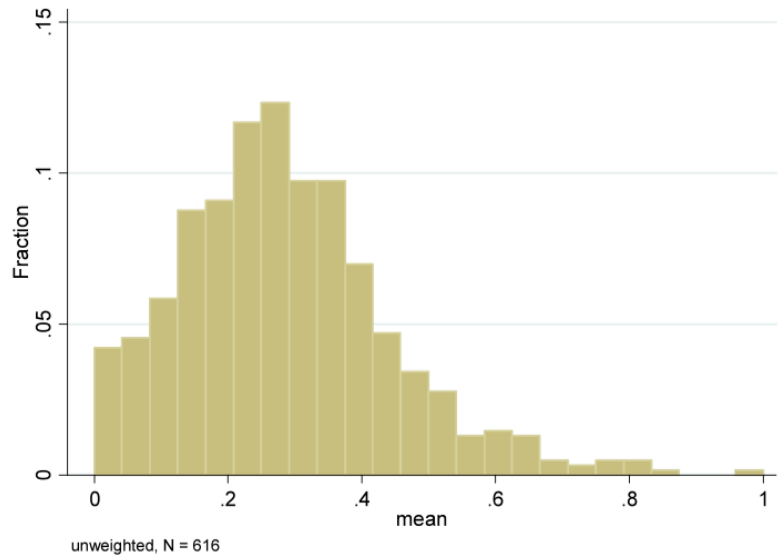


Fig. 1 Distribution of mean giving per treatment

Engel (2011) *Dictator Games: A Meta Study* Experimental Economics

Is social science learning true theories?

- Social science theories **don't converge** like this
- Representational power is **poor**
  - » Same game, different labs ⇒ wildly different results
- Multiple theories explain equally little
  - » What we have is **historically contingent**
    Pickering 1995

⇒ Theory is a **source of hypotheses**, not a map of reality

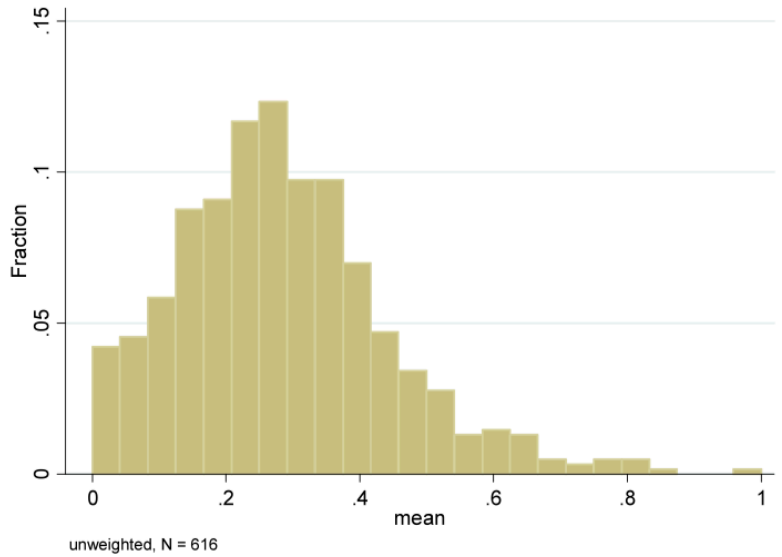


Fig. 1 Distribution of mean giving per treatment

Engel (2011) *Dictator Games: A Meta Study* Experimental Economics

Is social science learning true theories?

- Social science theories **don't converge** like this
- Representational power is **poor**
  - » Same game, different labs ⇒ wildly different results
- Multiple theories explain equally little
  - » What we have is **historically contingent**
    Pickering 1995

⇒ Theory is a **source of hypotheses**, not a map of reality

Alas, theory will not generalize for us!

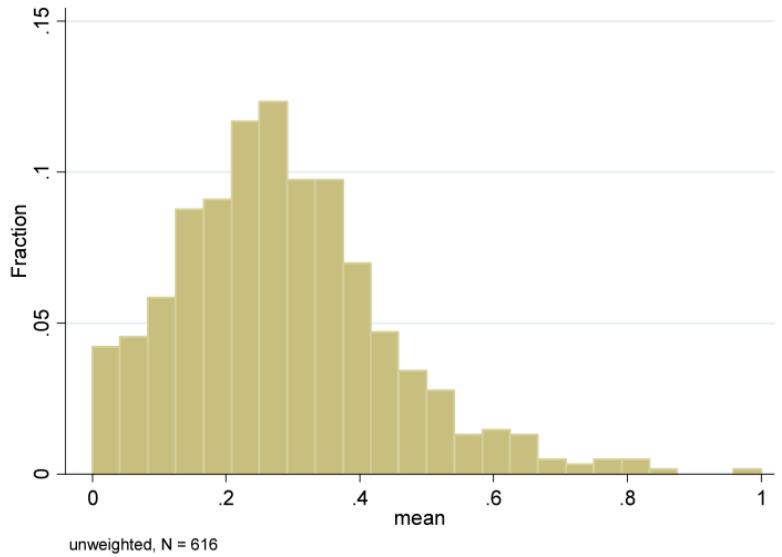


Fig. 1 Distribution of mean giving per treatment

Engel (2011) *Dictator Games: A Meta Study* Experimental Economics

Limits to Generalization

What does an experiment actually identify?

An RCT identifies a treatment effect for a specific:

$$\mathbb{E}[\tau(S, P, R, T) \mid S = s, P = p, R = r, T = t]$$

Dimmery and Munger (2025) *Enough?* Observational Studies

What does an experiment actually identify?

An RCT identifies a treatment effect for a specific:

$$\mathbb{E}[\tau(S, P, R, T) \mid S = s, P = p, R = r, T = t]$$

- S = **sample** (which units were studied)
- P = **population / site** (where the study took place)
- R = **realization** (how the treatment was operationalized)
- T = **time** (when the study was conducted)

Dimmery and Munger (2025) *Enough?* Observational Studies

What does an experiment actually identify?

An RCT identifies a treatment effect for a specific:

$$\mathbb{E}[\tau(S, P, R, T) \mid S = s, P = p, R = r, T = t]$$

- S = **sample** (which units were studied)
- P = **population / site** (where the study took place)
- R = **realization** (how the treatment was operationalized)
- T = **time** (when the study was conducted)

Design-based generalization requires randomizing **all four** — but we cannot for T:

> We cannot decide to run an experiment before it is conceived.

Dimmery and Munger (2025) *Enough?* Observational Studies

What does an experiment actually identify?

An RCT identifies a treatment effect for a specific:

$$\mathbb{E}[\tau(S, P, R, T) \mid S = s, P = p, R = r, T = t]$$

- S = **sample** (which units were studied)
- P = **population / site** (where the study took place)
- R = **realization** (how the treatment was operationalized)
- T = **time** (when the study was conducted)

Design-based generalization requires randomizing **all four** — but we cannot for T:

> We cannot decide to run an experiment before it is conceived.

The field is already retreating:

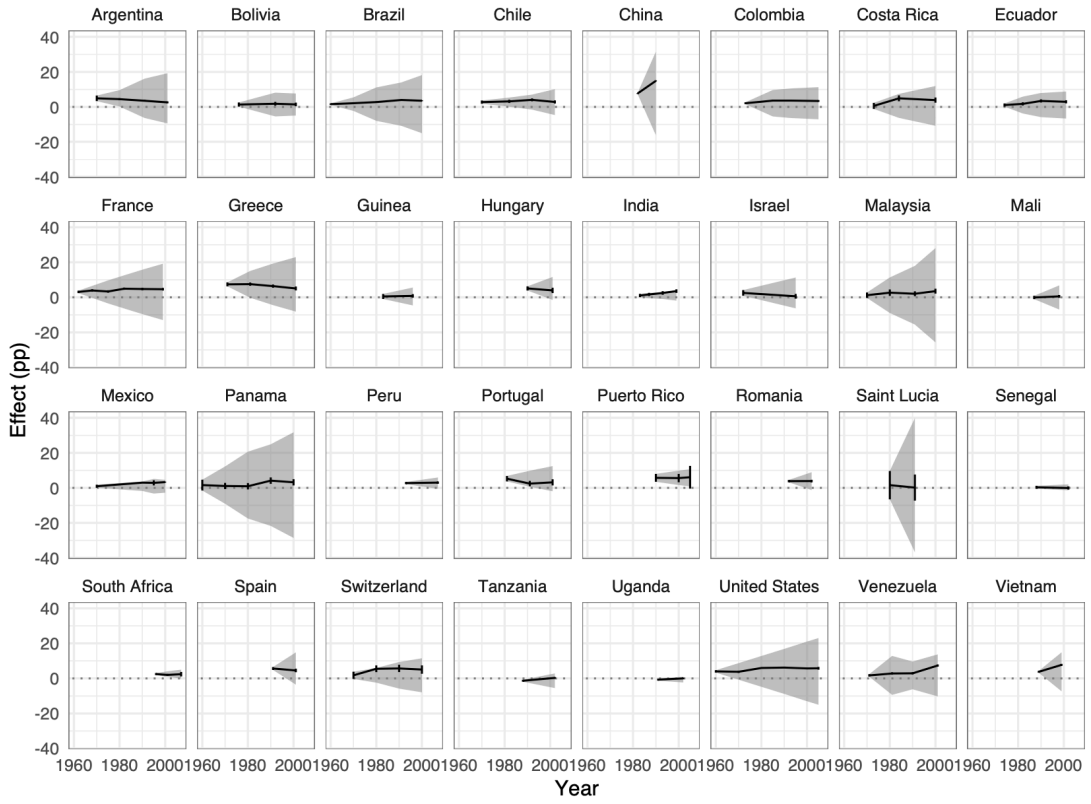
- we want to generalize over too much (Yarkoni 2020)
- so we settle for **"sign generalizability"** (Egami & Hartman 2023).

Dimmery and Munger (2025) *Enough?* Observational Studies

Experimentation in a changing social world

Partial Identification

- L: bound on TE change
- t_1: Target context
- t_0: Source context
- M: bound on HTEs
- $W_1(p_0, p_1)$: Wasserstein distance



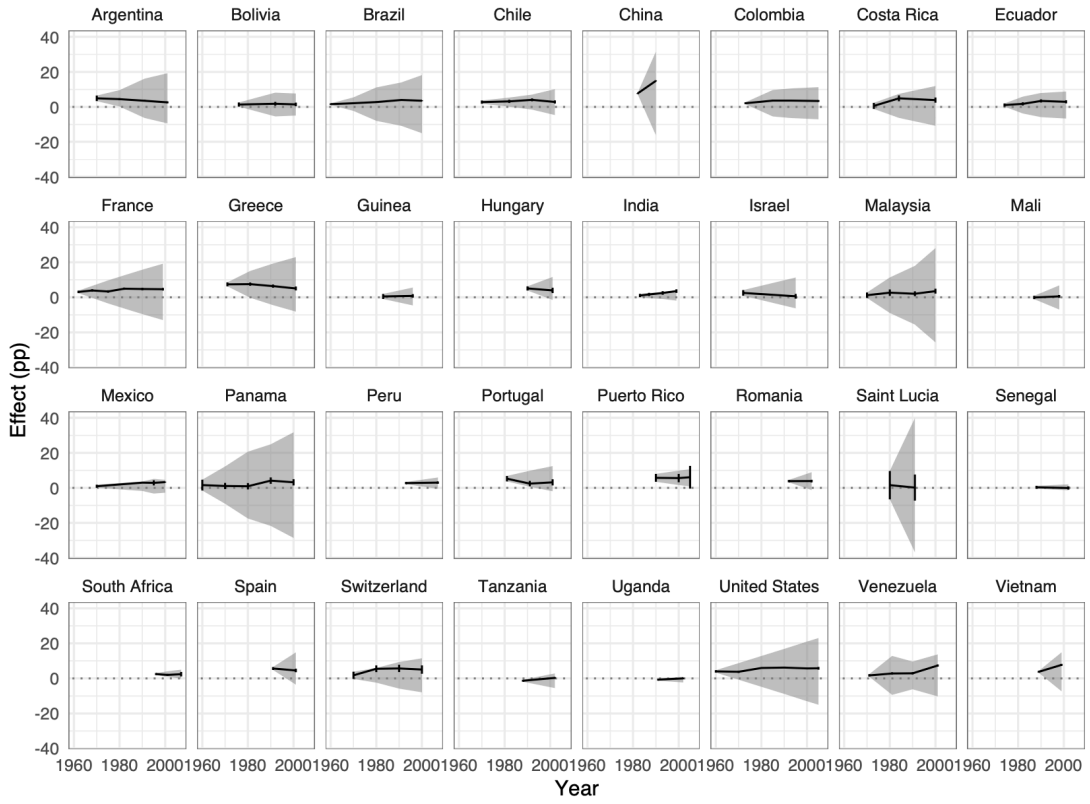
Dimmery and Munger (2023) *Measure Twice, Treat Once: The Necessity of Descriptive Knowledge for Causal Generalization* working paper

Experimentation in a changing social world

Partial Identification

- L: bound on TE change
- t_1: Target context
- t_0: Source context
- M: bound on HTEs
- $W_1(p_0, p_1)$: Wasserstein distance

$$L(t_1 - t_0) + M\, W_1(p_0, p_1)$$



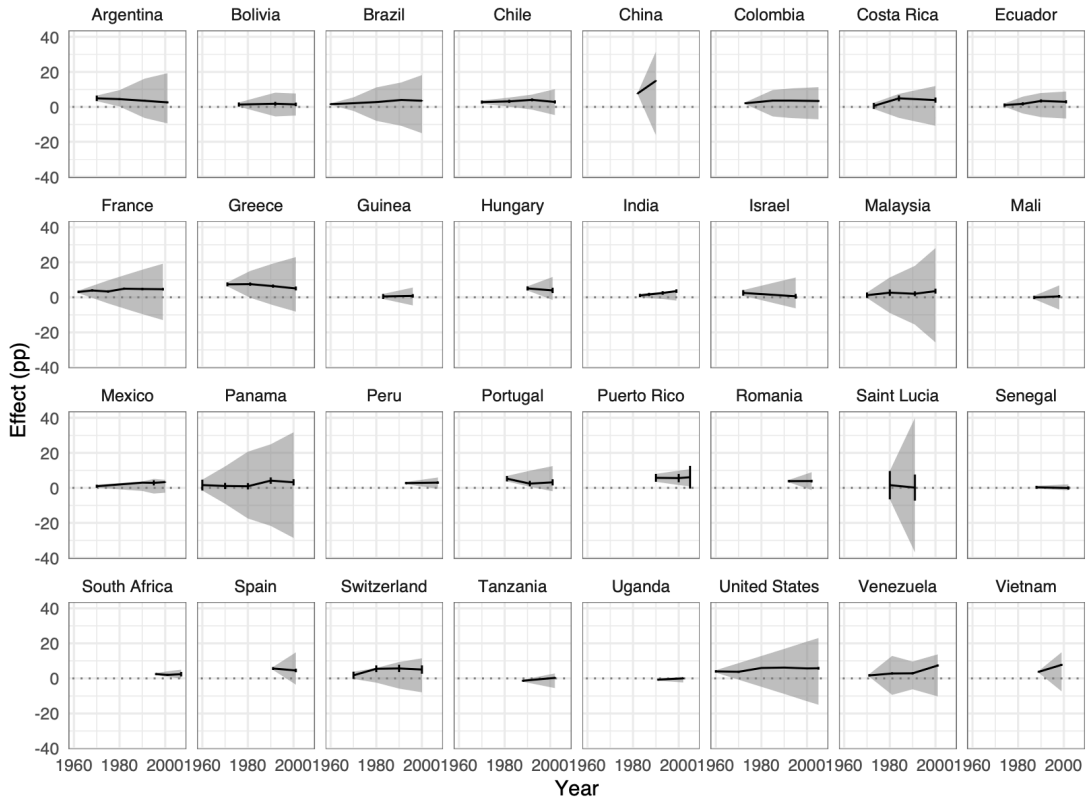
Dimmery and Munger (2023) *Measure Twice, Treat Once: The Necessity of Descriptive Knowledge for Causal Generalization* working paper

Experimentation in a changing social world

Partial Identification

- L: bound on TE change
- t_1: Target context
- t_0: Source context
- M: bound on HTEs
- $W_1(p_0, p_1)$: Wasserstein distance

$$L(t_1 - t_0) + M\, W_1(p_0, p_1)$$



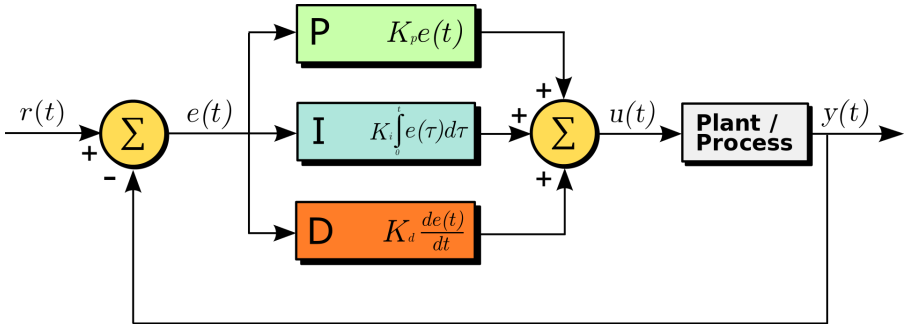
The possibility of change rapidly outpaces the local knowledge we have gained.

Dimmery and Munger (2023) *Measure Twice, Treat Once: The Necessity of Descriptive Knowledge for Causal Generalization* working paper

Cybernetic Social Science

How could social science work?

- Stop trying to ***learn*** things
  - The "Detour through knowledge" Pickering 2010
  - Or at least don't make "success" depend on it
- Seek to ***do*** things
  - Cyberneticists like Stafford Beer sought to build machines that *acted* without "*thinking*"
  - e.g.: a PID controller / cruise control
- Trade **generalization** for **regulation**



Towards a cybernetic social science

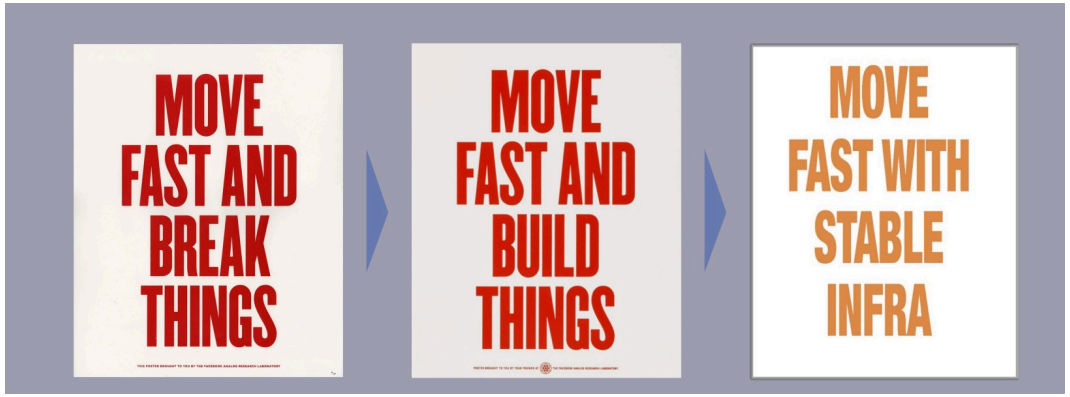
- **Think** of something
  - Radical pluralism Feyerabend 1975
  - Formal models
  - Agent-based models
  - A good novel
  - A painting
- **Act** in the world
  - Ontological case for RCTs Dimmery & Munger 2025
- **Observe** what happens
  - Deploy rigor here Campbell 1973
- **Ought** should be left to someone else
  - Like democratic institutions Dewey 1927, Dorf and Sabel 1998

How Bing Image Creator imagines Cybernetic Social Scientists

Cybernetic social science exists and is eating the world

Iterative improvement eventually wins

- The Internet Trap Hindman 2018
- Agile software development / Scrum
  - "ship early, ship often"
  - "highest priority is to satisfy the customer through early and continuous delivery of valuable software." Beck et al. 2001
- coordinate […] via feedback loops, […] reality-based […] central to […] higher level behavior. Sutherland 2004



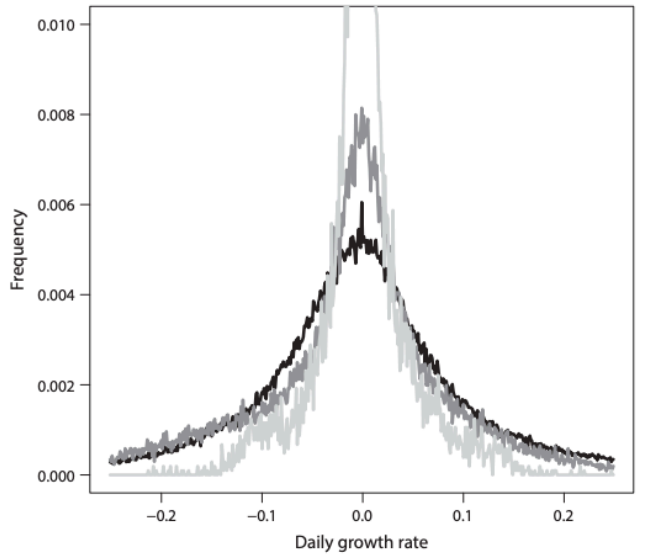
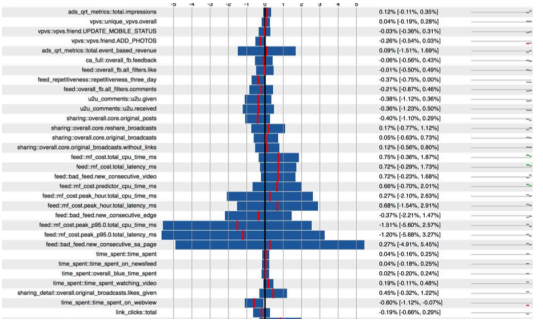


Figure 5.3 This figure shows the distribution of daily log growth rates for large, midsize, and small sites in our data. All three sets of sites show a clear bell curve, with the distribution much more dispersed for midsized and (especially) smaller sites. Fully observed data would show even fatter and more symmetrical left-hand tails for the smallest sites, since they are most affected by leakage.

Experimentation at Facebook

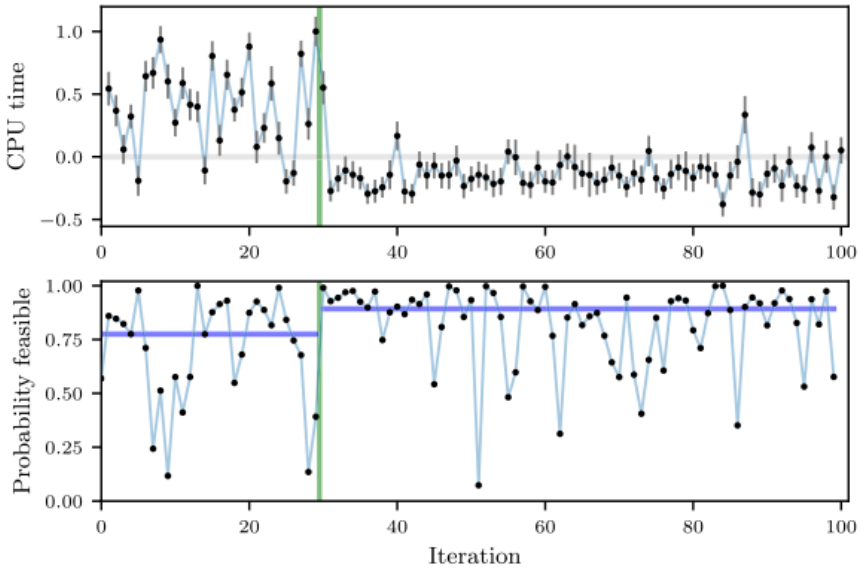
- Performance of employees is largely measured by AB tests
  - » ∴ everything is tested
- "Positive" changes retained
- Massively multi-objective optimization problem
- There's no real "learning" about **how** things work
  - » Just a test of **whether** they do



Dimmery, Bakshy and Sekhon (2019) *Shrinkage Estimators in Online Experiments* SIGKDD
Wu, Tan, Li, Garrard, Obeng, Dimmery, Singh, Wang, Jiang & Bakshy (2022) *Interpretable personalized experimentation* SIGKDD

Methods help scale and automate

- Bayesian Optimization for optimizing complex systems
  - » Literally also called “black-box optimization”
- They are not even attempting to “feel the lock turn”
- “Facebook” **embodies the values** of {Meta, Zuck, the product owner}



Facebook

Bringing People Closer Together

January 11, 2018
Adam Mosseri, Head of News Feed

Letham, Karrer, Ottoni and Bakshy (2019) *Constrained Bayesian Optimization with Noisy Experiments* Bayesian Analysis

Formalizing this institutional perspective

Experimentation as gradient descent

Setup: An institution controls policy $\theta_t \in \Theta \subseteq \mathbb{R}^d$ with diameter D

- observes outcomes $Y_t = f_t(\theta_t) + \varepsilon_t$, has utility $U(Y_t)$.
- $\|\nabla \ell_t(\theta)\| \leq G \qquad \|\nabla^2 \ell_t(\theta)\| \leq H$

Dimmery (2026) *Social Science without Generalization* working paper

Experimentation as gradient descent

Setup: An institution controls policy $\theta_t \in \Theta \subseteq \mathbb{R}^d$ with diameter D

- observes outcomes $Y_t = f_t(\theta_t) + \varepsilon_t$, has utility $U(Y_t)$.
- $\|\nabla \ell_t(\theta)\| \leq G \qquad \|\nabla^2 \ell_t(\theta)\| \leq H$

Problem: Maximize $\sum_{t=1}^{T} U(Y_t)$ in an **adversarial** setting

- f_t is chosen by nature (or society) **after** the institution commits to θ_t.
- No stationarity assumed.

Dimmery (2026) *Social Science without Generalization* working paper

Experimentation as gradient descent

Setup: An institution controls policy $\theta_t \in \Theta \subseteq \mathbb{R}^d$ with diameter D

- observes outcomes $Y_t = f_t(\theta_t) + \varepsilon_t$, has utility $U(Y_t)$.
- $\|\nabla \ell_t(\theta)\| \leq G \qquad \|\nabla^2 \ell_t(\theta)\| \leq H$

Problem: Maximize $\sum_{t=1}^{T} U(Y_t)$ in an **adversarial** setting

- f_t is chosen by nature (or society) **after** the institution commits to θ_t.
- No stationarity assumed.

Key insight (chain rule):

$$\nabla_\theta U(f_t(\theta_t)) = \underbrace{\nabla_Y U(Y_t)}_{\text{known preferences}} \cdot \underbrace{\nabla_\theta f_t(\theta_t)}_{\text{causal effect}}$$

A test comparing θ_t vs $\theta_t + \delta e_j$ gives a **finite-difference approximation** to $\nabla_\theta f_t(\theta_t)$.

Dimmery (2026) *Social Science without Generalization* working paper

Experimentation as gradient descent

Setup: An institution controls policy $\theta_t \in \Theta \subseteq \mathbb{R}^d$ with diameter D

- observes outcomes $Y_t = f_t(\theta_t) + \varepsilon_t$, has utility $U(Y_t)$.
- $\|\nabla \ell_t(\theta)\| \leq G \qquad \|\nabla^2 \ell_t(\theta)\| \leq H$

Problem: Maximize $\sum_{t=1}^{T} U(Y_t)$ in an **adversarial** setting

- f_t is chosen by nature (or society) **after** the institution commits to θ_t.
- No stationarity assumed.

Key insight (chain rule):

$$\nabla_\theta U(f_t(\theta_t)) = \underbrace{\nabla_Y U(Y_t)}_{\text{known preferences}} \cdot \underbrace{\nabla_\theta f_t(\theta_t)}_{\text{causal effect}}$$

A test comparing θ_t vs $\theta_t + \delta e_j$ gives a **finite-difference approximation** to $\nabla_\theta f_t(\theta_t)$.

You don't need to understand **why** — just measure the gradient.

Dimmery (2026) *Social Science without Generalization* working paper

Regret bounds

Regret: cumulative loss relative to the best fixed policy in hindsight:

$$R_T = \sum_{t=1}^{T} \ell_t(\theta_t) - \min_{\theta^*} \sum_{t=1}^{T} \ell_t(\theta^*)$$

Sublinear R_T ($R_T/T \to 0$) = convergence.

Regret bounds

Regret: cumulative loss relative to the best fixed policy in hindsight:

$$R_T = \sum_{t=1}^{T} \ell_t(\theta_t) - \min_{\theta^*} \sum_{t=1}^{T} \ell_t(\theta^*)$$

Sublinear R_T ($R_T/T \to 0$) = convergence.

With A/B tests (perturbation size δ, Hessian bound H):

$$R_T = O\left(DG\sqrt{T} + \underbrace{\delta H \cdot D \cdot T}_{\text{bias from finite } \delta} \right)$$

Regret bounds

Regret: cumulative loss relative to the best fixed policy in hindsight:

$$R_T = \sum_{t=1}^{T} \ell_t(\theta_t) - \min_{\theta^*} \sum_{t=1}^{T} \ell_t(\theta^*)$$

Sublinear R_T ($R_T/T \to 0$) = convergence.
With A/B tests (perturbation size δ, Hessian bound H):

$$R_T = O\left(DG\sqrt{T} + \underbrace{\delta H \cdot D \cdot T}_{\text{bias from finite } \delta} \right)$$

- Ideal ($\delta \to 0$): $R_T = O(\sqrt{T})$ — vanishing average regret
- Finite δ: bias term is **linear** — average regret hits a floor
- Optimizing $\delta = T^{-1/4}$: recovers $R_T = O(T^{3/4})$

How to attain linear regret: NHST

The dominant paradigm — null hypothesis significance testing — is a **thresholded sign-update**:

$$\theta_{t+1} = \theta_t + \Delta \cdot \tilde{\tau}_t$$

$$\tilde{\tau}_t \equiv \mathbb{1}\left\{\left|\frac{\hat{\tau}_t}{\widehat{se}_t}\right| > z_{\alpha/2}\right\} \cdot \mathrm{sign}(\hat{\tau}_t)$$

How to attain linear regret: NHST

The dominant paradigm — null hypothesis significance testing — is a **thresholded sign-update**:

$$\theta_{t+1} = \theta_t + \Delta \cdot \tilde{\tau}_t$$

$$\tilde{\tau}_t \equiv \mathbb{1}\left\{\left|\frac{\hat{\tau}_t}{\widehat{se}_t}\right| > z_{\alpha/2}\right\} \cdot \text{sign}(\hat{\tau}_t)$$

Three information losses vs. gradient descent:

1. Uses only the **sign**, discarding magnitude — i.e. **"sign generalizability"**
2. **Fixed** step size Δ regardless of precision
3. **No update** when result is non-significant

How to attain linear regret: NHST

The dominant paradigm — null hypothesis significance testing — is a **thresholded sign-update**:

$$\theta_{t+1} = \theta_t + \Delta \cdot \tilde{\tau}_t$$

$$\tilde{\tau}_t \equiv \mathbb{1}\left\{\left|\frac{\hat{\tau}_t}{\widehat{se}_t}\right| > z_{\alpha/2}\right\} \cdot \text{sign}(\hat{\tau}_t)$$

Three information losses vs. gradient descent:

1. Uses only the **sign**, discarding magnitude — i.e. **"sign generalizability"**
2. **Fixed** step size Δ regardless of precision
3. **No update** when result is non-significant

Compare: $\theta_{t+1} = \Pi_\Theta[\theta_t + \eta_t \cdot \hat{\tau}_t]$ — uses full magnitude, updates every period, scales step to precision.

NHST = Perceptron

	NHST	Perceptron (Rosenblatt, 1958)
Update	$\theta_{t+1} = \theta_t + \Delta \cdot \tilde{\tau}_t$	$w_{t+1} = w_t + \eta \cdot y_t$
Signal	**sign** of effect only	**sign** of error only
Step	**fixed** Δ	**fixed** η
Gate	significant $\Rightarrow$ update	misclassified $\Rightarrow$ update

NHST = Perceptron

	NHST	Perceptron (Rosenblatt, 1958)
Update	$\theta_{t+1} = \theta_t + \Delta \cdot \tilde{\tau}_t$	$w_{t+1} = w_t + \eta \cdot y_t$
Signal	**sign** of effect only	**sign** of error only
Step	**fixed** Δ	**fixed** η
Gate	significant $\Rightarrow$ update	misclassified $\Rightarrow$ update

Separability = **the optimum doesn't move**

- for any θ, θ', the sign of $U_t(\theta) - U_t(\theta')$ is **fixed over time**.
- **If not**: $R_T = \Omega(T)$ — **linear regret**
- Gradient descent needs only **Lipschitz continuity** $\Rightarrow$ $O(T^{3/4})$ **even when it does**

Comparison of approaches

Mode	Procedure	Adversarial R_T
One experiment only	None	$\Omega(T)$
Significance testing	Perceptron	$\Omega(T)$ or $O(1)$ *
Controlled experiment	OGD + finite-diff.	$O(\sqrt{T} + \delta HT)$ $\overset{\delta = T^{-1/4}}{\Longrightarrow} O(T^{3/4})$
One at a time	Coordinate descent	$O(d\sqrt{T} + \delta HT)$
Full gradient†	OGD	$O(\sqrt{T})$

* requires **linear separability** (generalization succeeds).

† Theoretical benchmark; requires $\delta \to 0$.

› Every row assumes a known U. What are we actually converging to?

The values question

- The framework is **indifferent** to U — every disagreement reduces to a disagreement about **the objective**
- Who decides?
  - In tech: product owners.
  - In society: **democratic deliberation**
- **Transparency virtue**: value choices are smuggled in through choice of outcome, treatment, framing — the cybernetic framework **forces them into the open**

The values question

- The framework is **indifferent** to U — every disagreement reduces to a disagreement about **the objective**
- Who decides?
  - In tech: product owners.
  - In society: **democratic deliberation**
- **Transparency virtue**: value choices are smuggled in through choice of outcome, treatment, framing — the cybernetic framework **forces them into the open**

> "The ends of men are many, and not all of them are in principle compatible with each other" — Isaiah Berlin (1969)

- Encoding values in scalar U presupposes **commensurability**.
- Domains that resist quantification are **outside the framework**
- Within those boundaries: **formal guarantees**.

Conclusion

- An experiment identifies effects for a **specific** sample, site, realization, and time
- Generalization requires assumptions no less fragile than those we reject for identification

Conclusion

- An experiment identifies effects for a **specific** sample, site, realization, and time
- Generalization requires assumptions no less fragile than those we reject for identification

- Replace generalization with **iteration**
  - Repeated experimentation = online gradient descent
  - Provable convergence even in non-stationary worlds
  - Only requires **internal validity**

Conclusion

- An experiment identifies effects for a **specific** sample, site, realization, and time
- Generalization requires assumptions no less fragile than those we reject for identification

- Replace generalization with **iteration**
  - Repeated experimentation = online gradient descent
  - Provable convergence even in non-stationary worlds
  - Only requires **internal validity**

- NHST (adopt/reject) = perceptron ⇒ linear regret when the world changes
- Using full gradient information ⇒ sublinear regret **regardless**

Conclusion

- An experiment identifies effects for a **specific** sample, site, realization, and time
- Generalization requires assumptions no less fragile than those we reject for identification
- Replace generalization with **iteration**
  - Repeated experimentation = online gradient descent
  - Provable convergence even in non-stationary worlds
  - Only requires **internal validity**
- NHST (adopt/reject) = perceptron ⇒ linear regret when the world changes
- Using full gradient information ⇒ sublinear regret **regardless**
- The remaining question is not statistical — it's **political**: who chooses U?
- Meta prioritizes **time spent** — and the product embodies that choice

Conclusion

- An experiment identifies effects for a **specific** sample, site, realization, and time
- Generalization requires assumptions no less fragile than those we reject for identification

- Replace generalization with **iteration**
  - Repeated experimentation = online gradient descent
  - Provable convergence even in non-stationary worlds
  - Only requires **internal validity**

- NHST (adopt/reject) = perceptron ⇒ linear regret when the world changes
- Using full gradient information ⇒ sublinear regret **regardless**

- The remaining question is not statistical — it's **political**: who chooses U?
- Meta prioritizes **time spent** — and the product embodies that choice

Is that the world we want to optimize for?

Thank you!

me@ddimmery.com