Beyond Validation: a Feedback Loop Between Decisions, Experiments, and Causal Models

Chris Khawand, Principal Economist

w/ Patrick Schwarz, Max Vogler

Amazon

Why should we trust causal estimates?

- Most of us produce causal impact estimates
- They affect business strategy & investments sometimes worth **$B**
- Fundamental Problem of Causal Inference: ***"Causal models have no ground truth"***
- The wrong question: "How do we know it's right?"
- The right question: "How good are the decisions we make?"

Levels of Model Trust

- **Level 0**: expert using established method (aka. “trust me bro”)
- **Level 1**: economists/experts ‘feel good’ about the model based on theory/gut/assorted data points, placebo tests, etc.
- **Level 2**: explicit accuracy metrics (e.g., out-of-sample RMSE)

The problem with validation

- Binary pass/fail criteria are uninteresting and risky
- Model success is a **continuum**, not binary
- Not “right or wrong”, but “close or far”
- RMSE criteria help there, but what is $1 of RMSE worth?
- **Model Goodness** should be defined as whether it drives better decisions than the alternative

(Complete) Levels of Model Trust

- **Level 0**: expert using established method (aka. “trust me bro”)
- **Level 1**: economists/experts ‘feel good’ about the model based on theory/gut/assorted data points, placebo tests, etc.
- **Level 2**: explicit accuracy metrics (e.g., out-of-sample RMSE)

Beyond validation:

- **Level 3**: Complete probabilistic distributions around estimates
- **Level 4:** Decision risk measured and priced

The Feedback Loop

1. **Combine** experimental & observational signals to get best estimate
2. Quantify **uncertainty** in metrics
3. Measure **decision loss**
4. Recommend & implement **experiments** to reduce decision loss
5. Repeat

Example: Experiment Design for Marketing

- Goal: choose spend to maximize net benefit (i.e., next $1 spend = $1 benefit)
- Concave payoff function
  - Estimated with uncertainty
  - Based on combo of experiment + observational models
  - High potential confounding in short-term and long-term impacts
- Business questions:
  - How good are our current decisions?
  - How many experiments should we run? Which types? Where? How long?

Model Setup

- **Business objective**: choose decision π to maximize expected utility given causal parameters θ

$$\tilde{\pi} = \arg\max_{\pi \in \Pi} \mathbb{E}_{\theta \sim f}[U(y(\pi, \theta))]$$

- Strictly increasing, concave utility: $U' > 0$, $U'' < 0$
- Causal parameter θ not known, but we have some prior info $f(\theta)$
- Notation similar to Athey and Wager (2021)

What do more experiments get us?

- We can get more information (e.g., experiments): ψ
  - For experiments, represents the design space (e.g., type, proportion treated)
  - These can be used to update and improve our estimates

- Cost of information: $c(\theta, \psi)$
  - Opportunity cost of forgone marketing
  - Explicit cost (tech, personnel)

The experiment optimization problem

- The value of the new information is:

$$V(\psi) = \int \int \Big(\underbrace{U(y(\pi^{E}(\hat{\theta}|\psi), \theta) - c(\theta, \psi))}_{\text{Value under new info minus cost}} - \underbrace{U(y(\pi^{NE}, \theta))}_{\text{Value under old info}} \Big) df(\hat{\theta}|\theta, \psi) df(\theta)$$

- The ideal experimental plan maximizes this value:

$$\psi^{opt} = \arg\max_{\psi} V(\psi)$$

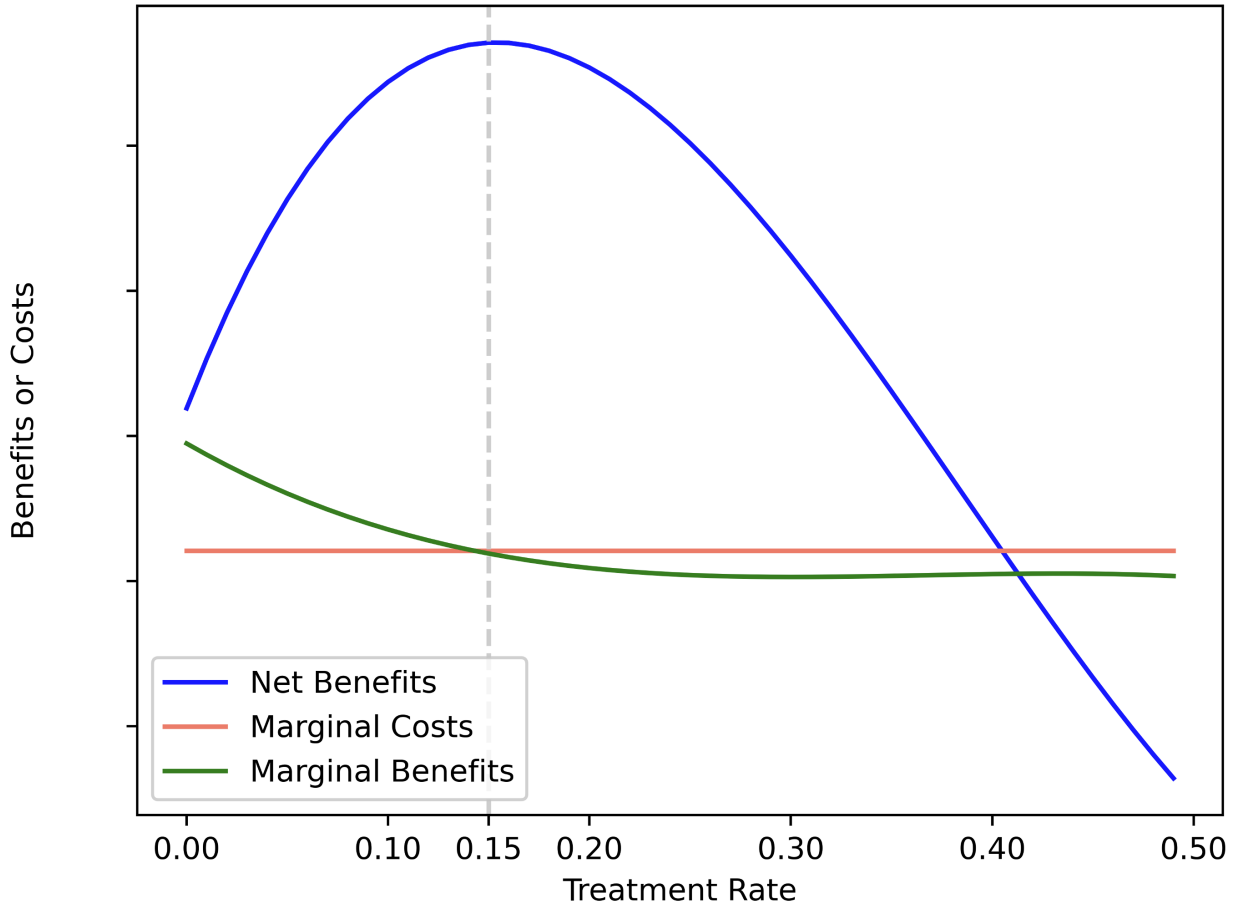
Measuring Uncertainty

- What we know about causal parameter is encapsulated in $f(\theta)$
- Multiple types of uncertainty, e.g.:
  - **Confounding**: primary source of anxiety about causal modeling
  - **Sampling error:** as measured by traditional confidence intervals
  - **External invalidity:** applicability of trained estimate to target population
  - … and more!
- Errors interact: need to **propagate uncertainty** to estimate distribution, esp. with multiple models
- **Ex:** surrogates models have **multiplicative error** for short-term impact X long-term impact (200% bias X 200% bias = 400% bias)

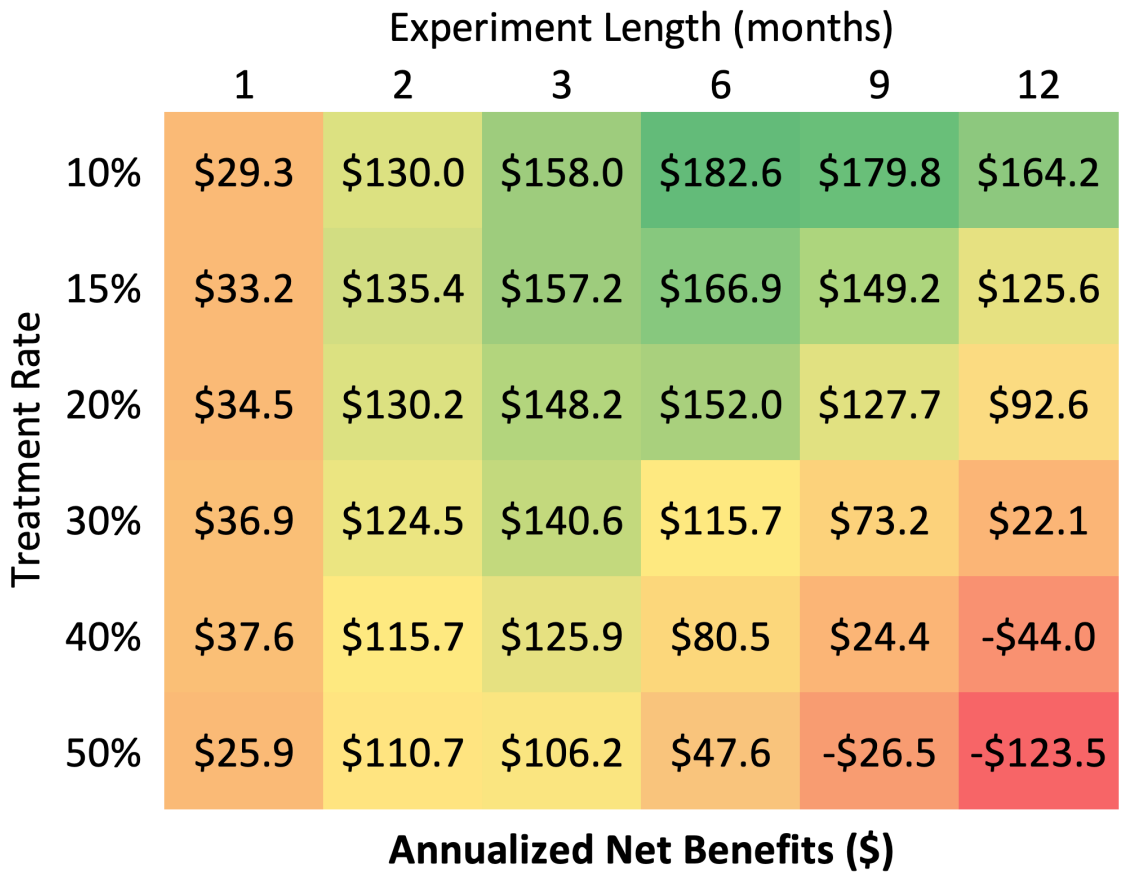
Experimental Design

- Experimental grounding step => uncertainty improves with more experiments
- Example: Marketing treatment => lower marketing intensity
  - Fraction treated? (Power)
  - How long? (Long-term vs. Short-term Effects)
  - For whom? Which countries? (Heterogeneity)
  - When? (Seasonality)
- Many different possible design schedules

1-dimensional Example: Treatment Rate



2-dimensional example: Rate & Length



What are key lessons we learned?

1. Give Bayesian statistics a chance
2. Demand definition of a decision framework, or specify it yourself
3. Assess **portfolios** for experimental design, not individual experiments
4. Business leaders struggle to prioritize investments in measurement, *unless you say why it matters ($$) and how to do it*
5. Most ideas aren't scientifically new (e.g., reinforcement learning), just hard to operationalize
6. All this applies to any information improvement: model changes, better data

What's next?

- Details in paper to be released (under internal review)
- Assessing internal use cases for experimentation + decision health
  - Do they have an explicit decision framework written down?
  - How uncertain are their metrics?
  - Are they doing too little, or too much experimentation?
  - What's blocking them? Conviction, technology, or something else?
  - How much do we stand to gain by fixing it?
- Long-run ideal:
  - All models empirically evaluated
  - Need for expert assessment -> 0