# A comment on the 'Value of Information' and misaligned incentives

James Michelson

July 2026

**What's going on here?**

GiveWell's 'Value of Information' framework (2, 17) proposes a helpful method for evaluating the cost-effectiveness of research about programs in a manner directly comparable to the cost-effectiveness of the programs themselves. However, when the same organization is responsible for carrying out the research as implementing the program, incentives for truthful reporting are undermined. This is related to the philosophical problem of *reflexivity* (25). This comment is an exercise in *applied* philosophy of science (i.e., metascience), trading philosophical abstraction for concreteness: exposing the mis-aligned incentive problem yields options for dissolving it. I outline this problem using the VoI framework in Section 4. In Section 5, I conclude with a proposal. As funders learn more about the bias arising from misaligned reporting incentives due to their choice of funding rules they can more accurately measure the value of independent auditing.

## 1 Introduction

GiveWell uses an explicit value-of-information (VoI) framework (2, 17) to evaluate when learning more information about a funding opportunity outweighs the gains from funding it directly. This framework enables like-for-like ($) comparisons between paying for information and funding charitable programs. However, it values information as though it were exogenous to the funding rule that generates it. Where the funding rule or mechanism changes the incentive structure of those giving information, the signal returned by the VoI framework is the outcome of a strategic game and may become biased (towards "it worked") so the true VoI is lower than the reported number.

The funding rule does not merely bias the reported measurement, it may affect the data-generating process that gets implemented: which sites are chosen, how intensively the program is run, how outcomes are recorded, etc. The very quantity being estimated becomes a function of the funding rule. The measurement is in this stronger sense *reflexive*[1] (25) or *performative* (27).

Some funding strategies are more prone to this problem of reflexivity or performativity than others. GiveWell's own grant taxonomy (15) distinguishes three types of grant:

a. Academic research to evaluate program evidence

b. Early-stage funding for a promising organization

[1] At the highest possible level of abstraction the problem of reflexivity can be understood as 'science causally affects what it studies'. Setting aside observer effects in the natural sciences (e.g., quantum physics), this problem is germane to every social science where the objects of scientific inquiry are aware of their place in a scientific investigation. A more detailed *philosophical* exposition of this can be found in (25). Unless stated otherwise, the finer points of reflexivity are moot for the discussion that follows.

c. Monitoring and evaluation of an existing organization

Notably, in the case of (b) and (c) above, the institution funded to gather information (the '*measurer*') may or may not be distinct from the institution subsequently responsible for implementing the program (the '*implementer*'). Where these roles coincide, institutions can gain by misreporting the effects they measure. In the sections that follow, I work through the logic of this incentive misalignment using examples from GiveWell's funding history, price it inside GiveWell's own VoI arithmetic, and conclude with two points of departure for a solution.

## 2 A taxonomy of misaligned incentives

I set aside the case of (a) where funding is given for "academic research to evaluate program evidence." Here, there is no institution which takes the role of an *implementer*. I focus on cases (b) and (c) where different institutions may take different roles.

### 2.1 Independent *measurer* and *implementer* institutions

The RCT of **New Incentives**' (the *implementer*) conditional cash-transfer program for infant vaccination in northern Nigeria was run by **IDinsight** (the *measurer*), an independent research organization, which received the evaluation grants directly (8). The trial randomized 167 clinics across Katsina, Zamfara, and Jigawa States (July 2017–February 2020) and found a +22pp increase in vaccination coverage (10). New Incentives received operating grants during the trial — including a $5.9M grant in November 2017 explicitly to support operations for the duration of the RCT (5, 22).

This is as close to an ideal incentive-compatible funding design as we can hope for: separating the two roles in two distinct institutions. The same separation holds for the Y-RISE/BRAC, One Acre Fund, and LSHTM eyeglass RCTs (18, 13, 16).

### 2.2 When the *measurer* becomes *implementer*

In the case of (b) "early-stage funding for a promising organization", an organization can grow in proportion to their (positive) research findings. For example, **Fortify Health** (iron/flour fortification, India) received three incubation grants totaling ∼$9.5M (∼$300,000 in June 2018; ∼$1M in August 2019; $8.2M over five years in December 2021), then a lookback using Fortify Health's own growth data led to a renewed $10M grant (6, 7, 11). **Results for Development** (childhood-pneumonia treatment, Tanzania) received a $6.4M Phase I grant (under which it conducted its own monitoring and evaluation) and then a $5.6M Phase II scale-up grant to the same organization (4, 9). In both, the organization knows at the evidence stage that a favorable result indicating program effectiveness unlocks the scale grant (*implementer*) and it is in a position to report as it wishes (*measurer*).

### 2.3 When the *measurer* is the *implementer*

Case (c) collapses the two roles by construction and in the case of **Evidence Action**'s Dispensers for Safe Water program GiveWell explicitly concedes that their VoI estimates may have been biased. Evidence Action's monitoring was self-collected (chlorine use was measured by matching a test result to a color wheel) and "this data was collected by Evidence Action's own staff, which may

have led them to interpret the color wheel results more favorably" (24). Independent data (the **Kenya Study of Water Treatment and Child Survival**) found that in overlapping villages the number of people drinking chlorinated water was only 33% as high as routine monitoring showed, an independent census found ∼40% fewer users, and the headline correction was ∼2 million actual users against a reported 5 million (21, 24). These three different comparisons imply overstatement factors of roughly 3×, 1.7×, and 2.5×, respectively.

## 3 Value of Information: an overview

This section reviews GiveWell's VoI framework. I use the example from Brown et al. (2) (Table 1). Although the numbers are now outdated, they do not materially affect the subsequent analysis.

| Terms | Quantity | Value |
|---|---|---|
| $c$ | Cost of study | \$3.4M |
| $S$ | Change in funding if we positively update | \$190M (= \$19M/yr × 10 yr) |
| $p$ | Probability of a positive update | 0.23 |
| $\kappa_1$ | CE of "unlocked" opportunity | 11.7× cash |
| $\bar{\kappa}$ | CE of counterfactual opportunity (the bar) | 10× cash |
| $\Delta$ | Marginal value of reallocated funding | $\kappa_1 - \bar{\kappa} = 1.7\times$ cash |
| | CE of funding the study | ∼22× cash |

Table 1: The VoI calculation for ORS/zinc distribution in northern Nigeria

The value $V$ of reallocating capital in cash-equivalent dollars is given by:

$$V = S \cdot \Delta \cdot p$$

and the cost-effectiveness (CE) of the study is simply:

$$\text{CE}_{\text{study}} = \frac{S \cdot \Delta \cdot p}{c} \approx 22 \times \text{cash}$$

which entails that the \$3.4M study should be funded since $\text{CE}_{\text{study}} > \bar{\kappa}$. Note, spending $c$ on the marginal program returns $c \cdot \bar{\kappa}$ in cash-equivalent value, so the bar for funding the study is a ratio in CE units, never a raw dollar inequality.

The quantities in Table 1 are themselves *ex ante* objects, derived from a prior distribution $\pi$ over the parameter $\theta$ under investigation by the study. The prior gives us a distribution over $\text{CE}(\theta)$, and the decision rule once the study reports is: reallocate $S$ iff $\text{CE}(\theta) > \bar{\kappa}$. The realized value[2] when $\theta$ is revealed is:

$$V(\theta) = S \cdot \big(\text{CE}(\theta) - \bar{\kappa}\big) \cdot \mathbb{1}\{\text{CE}(\theta) > \bar{\kappa}\}$$

and the expected value before the study (from the funder's *ex ante* perspective) is:

$$\mathbb{E}_\pi\big[V(\theta)\big] = S \cdot \underbrace{\mathbb{E}_\pi\big[\text{CE}(\theta) - \bar{\kappa} \mid \text{CE}(\theta) > \bar{\kappa}\big]}_{\Delta = \kappa_1 - \bar{\kappa}} \cdot \underbrace{\text{P}_\pi\big(\text{CE}(\theta) > \bar{\kappa}\big)}_{p}$$

[2]For simplicity, I ignore the case of negative updates as noted in Brown et al. (2, 310, fn 3). I also treat the study as revealing $\theta$ exactly ("post-trial certainty").

# 4 The problem of misaligned incentives

Suppose that *implementer* and *measurer* roles coincide. Here, by construction, incentives are mislagned and the organisation can benefit from inflating the effect of its progam. How then to model this problem? Let the study design be biased upward multiplicatively such that the report $\widehat{\text{CE}} = (1+b)\text{CE}(\theta)$, for some $b \geq 0$. (This move allows for the fact that, as in the case of Evidence Action, the misreport lies in the number of people who were claimed to be drinking chlorinated water, not merely in the efficacy of the intervention $\theta$.) The harm is a 'false unlock' where the reported estimate clears the bar for funding whereas the true one does not. Define:

$$p_{\text{false}} = \text{P}_\pi\left[\text{CE}(\theta) \leq \bar{\kappa} < \underbrace{(1+b)\text{CE}(\theta)}_{\widehat{\text{CE}}}\right] = \text{P}_\pi\left[\frac{\bar{\kappa}}{1+b} < \text{CE}(\theta) \leq \bar{\kappa}\right]$$

Conditional on a false unlock, the funder moves $S$ out of the counterfactual option $\bar{\kappa}$ into a program with true cost-effectiveness $\text{CE}(\theta)$, losing $\bar{\kappa} - \text{CE}(\theta)$ per dollar. Therefore,

$$V_{\text{audit}} = S \cdot \mathbb{E}_\pi\left[\big(\bar{\kappa} - \text{CE}(\theta)\big) \cdot \mathbb{1}\big\{\frac{\bar{\kappa}}{1+b} < \text{CE}(\theta) \leq \bar{\kappa}\big\}\right]$$

and by extension, we buy an independent audit (i.e., decouple *implementer* and *measurer* organisations) iff

$$c_{\text{audit}} < \frac{V_{\text{audit}}}{\bar{\kappa}}$$

The audit is worth buying in proportion to the prior mass sitting in the false-unlock band. A tight prior centred well above the bar renders verification nearly worthless.

The above presentation took an unbiased $\pi$ for granted: the prior over ORS usage rates comes from previous studies. However, if previous studies face the incentive misalignment problems outlined here, the damage compounds across the whole grant portfolio. This value hard to quantify but conceptually it enters the calculation as a second term, so that the audit is bought against $V_{\text{audit}} + V_{\text{knowledge}}$, where the latter is the value of an uncontaminated prior $\pi$.

## 4.1 A (hypothetical) worked example

Keeping the setup roughly similar to Table 1 above, suppose $\text{CE}(\theta) \sim \text{Unif}[0, 13.4]$. Then,

- $p = \text{P}[\text{CE} > 10] \approx 0.25$
- $\kappa_1 = \mathbb{E}[\text{CE}|\text{CE} > 10] = 11.7$
- $V = 0.25 \times 190 \times 1.7 \approx \$81\text{M}$

Now take $b$ from the *residual* range implied[3] by **Evidence Action**'s Dispensers for Safe Water program above, yielding $b \in [0.5, 1.7]$ (see Table 2). Notice that raising $b$ widens the band $(\frac{\bar{\kappa}}{1+b}, \bar{\kappa}]$, so $p_{\text{false}}$ rises *and* the conditional mean inside the band falls. $V_{\text{audit}}$ is therefore increasing in $b$ twice over.

[3] $b$ here is the bias net of the (10%) downward adjustments GiveWell applied. (Recall, the estimated number of people drinking chlorinated water was either 33% or 60% of the reported total (21, 24)). So to get the bias for the cost-effectiveness $\widehat{\text{CE}}$, we calculate $0.9/0.33 - 1 = 1.73$ and $0.9/0.6 - 1 = 0.5$ to get the factor $b$ on $(1+b)\text{CE}(\theta)$.

| | $b = 0.5$ | $b = 1.7$ |
|---|---|---|
| False-unlock band | $(6.67, 10]$ | $(3.70, 10]$ |
| $p_{\text{false}}$ | 0.25 | 0.47 |
| $\mathbb{E}[\text{CE} \mid \text{band}]$ | 8.33 | 6.85 |
| Expected shortfall | 1.67× | 3.15× |
| $V_{\text{audit}}$ | ∼\$79M | ∼\$281M |
| Buy the audit iff $c_{\text{audit}} <$ | **\$7.9M** | **\$28.1M** |

Table 2: The audit rule for $b \in \{0.5, 1.7\}$.

Thus an independent verification budget of somewhere between \$7.9M and \$28.1M is justified against a capital unlock of \$190M. Note, this is a ceiling in two senses. First, it is a threshold on willingness-to-pay. Second, because the harm originates in a report the *measurer* choses to inflate, deterrence is available more cheaply than detection. On this second point, random audits with probability $q$ (where the grant is withdrawn or clawed back on detection) deter inflation at expected cost $q \cdot c_{\text{audit}}$ rather than $c_{\text{audit}}$ (29).

### 4.2 A comment on ranking

I have treated $\bar{\kappa}$ as a constant but GiveWell sets the bar against projected funding. "When we anticipate having more funding, we may be able to lower the cost-effectiveness bar... when we project having less funding, we may decide to raise the funding threshold" (23). It has moved from 6× to 10× in 2022, back down to 8× over 2024–2025, and back to 6× as of May 2026 (12, 23, 14, 19). This version of the problem entails ranking programs but does not change the analysis given here. A program inflated from $\kappa$ to $(1 + b)\kappa$ displaces programs with true CE in $(\kappa, (1 + b)\kappa]$, and the loss is again $S \cdot (\mathbb{E}[\text{CE of displaced}] - \kappa)$.

## 5 Proposed solutions

Before offering a sketch towards a solution, I want to be clear: I do not presume the funding landscape encountered by GiveWell and others is dense with opportunities. Separating out *measurer* from *implementer* institutions might be operationally too difficult to merit the recommendations below. At a minimum, this is where I would start to think about augmenting GiveWell's VoI framework.

A core philosophical contention of mine (25) is that simple bias corrections are insufficient to deal with the problem of reflexivity: when the incentive structure that confronts strategic agents is perturbed as a function of a rule or mechanism, then their best response should be modeled as a function of the perturbation. Accordingly, I would explicitly model the bias $b$ as function of a decision rule $r$:

$$b = b(X, r), \quad r \in \{\text{separated}, \text{coincident}\}$$

where $X$ denotes covariates like geography, program type, and institution and $r$ is the funder's choice variable. It denotes whether the measurer and implementer institutions are the same[4]. Modeling

[4] $r$ is binary here for ease of presentation. It could be extended to cover different degrees/kinds of independence across institutions and roles over the life-cycle of a study.

$b$ as a function of *both* $X$ and $r$ allows the funder to gain insight into the data generating process that leads to misreporting as a function of their decisions. My read is that, to date, GiveWell's existing downward adjustments condition on $X$ but not $r$ (and it's the part GiveWell controls!)

Since role coincidence is chosen for reasons correlated with expected bias, it is endogenous and so regressing the bias on $r$ recovers selection, not the effect of the rule. However, after **Evidence Action**'s Dispensers for Safe Water program, GiveWell acknowledged the need for independent audits for all safe-water grants going forward (21, 20), which provides a policy discontinuity capable of identifying the effect of $r$ on $b$. That is a difference-in-difference design: safe-water grants before/after versus other grant categories before/after[5]. It gives variation in $r$ that is plausibly exogenous to individual program's expected bias.

A second solution might[6] be found in the application of contract theory to statistical inference (what has been dubbed "statistical contract theory") (1). The idea concerns a principal (in their case, a regulator like the FDA) who must decide whether to accept an agent's claim. The agent runs an experiment, bears the cost, and reports. The principal can only observe the report (not the experiment) and must commit to an acceptance rule in advance. The central result of Bates et al. (1) is that the principal should *not* simply apply the statistically optimal test: there's a quantifiable price to be paid in statistical power for incentive alignment.

In contrast to strategies like peer prediction (26), Bayesian truth serum (28), and strategyproof estimation (3), which ask "how should I score or aggregate reports so that truth-telling is a best response", statistical contract theory asks "what decision rule should I commit to, given the agent controls the evidence?" This maps directly onto GiveWell's problem since they control funding rules.

[5] Admittedly, the pre-policy safe-water grants have no independent benchmark, which makes the estimand something like "$b$ among grants where we had independent data".

[6] I find this line of research at the intersection of statistics and economics fascinating; however, there is simply not enough time in the day to put all the pieces of it together. Can it be extended to fit a more precisely scoped GiveWell-type problem? I think so. How? TBD. To my knowledge (asking some former colleagues) this kind of research is so new it has yet to be picked up and applied to real problems.

## References


[1] Stephen Bates, Michael I. Jordan, Michael Sklar, and Jake A. Soloff. Principal-agent hypothesis testing. *Journal of the American Statistical Association*, 2026. Forthcoming. arXiv:2205.06812 (v3, April 2024), https://arxiv.org/abs/2205.06812.

[2] Dan Brown, Adam Salisbury, and Meika Ball. Valuing information as a philanthropic funder. AEA Papers and Proceedings 115: 308–312, 2025. Ungated: https://benny.aeaweb.org/articles/pdf/doi/10.1257/pandp.20251021.

[3] Ioannis Caragiannis, Ariel D. Procaccia, and Nisarg Shah. Truthful univariate estimators. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *PMLR*, pages 127–135, 2016.

[4] GiveWell. Results for Development — childhood pneumonia treatment scale-up (may 2016). https://www.givewell.org/charities/results-for-development/may-2016-grant, 2016. Accessed 2026-07-29.

[5] GiveWell. New Incentives — general support. https://www.givewell.org/charities/new-incentives/april-2017-grant, 2017. Accessed 2026-07-29.

[6] GiveWell. Fortify Health — general support (june 2018). https://www.givewell.org/research/incubation-grants/fortify-health/june-2018-grant, 2018. Accessed 2026-07-29.

[7] GiveWell. Fortify Health — general support (august 2019). https://www.givewell.org/research/incubation-grants/fortify-health/august-2019-grant, 2019. Accessed 2026-08-16.

[8] GiveWell. IDinsight — endline evaluation of New Incentives RCT. https://www.givewell.org/research/incubation-grants/IDinsight-new-incentives-august-2019, 2019. Accessed 2026-07-29.

[9] GiveWell. Results for Development — childhood pneumonia treatment program (january 2019). https://www.givewell.org/research/incubation-grants/results-for-development/january-2019-grant, 2019. Accessed 2026-07-29.

[10] GiveWell. IDinsight's RCT of New Incentives. https://www.givewell.org/charities/IDinsight/partnership-with-idinsight/new-incentives-rct, 2020. Accessed 2026-07-29.

[11] GiveWell. Fortify Health — support for expansion (december 2021). https://www.givewell.org/research/incubation-grants/Fortify-Health-expansion-December-2021, 2021. Accessed 2026-07-29.

[12] GiveWell. An update on GiveWell's funding projections. https://blog.givewell.org/2022/07/05/update-on-givewells-funding-projections/, 2022. Accessed 2026-08-16.

[13] GiveWell. One Acre Fund — tree program RCT and data collection (january 2023). https://www.givewell.org/research/grants/one-acre-fund-tree-rct-january-2023, 2023. Accessed 2026-07-29.

[14] GiveWell. GiveWell's cost-effectiveness analyses — december 2024 version. https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models/december-2024-version, 2024. Accessed 2026-08-16.

[15] GiveWell. GiveWell incubation grants (program page). https://www.givewell.org/research/incubation-grants, 2024. Accessed 2026-07-29.

[16] GiveWell. London School of Hygiene & Tropical Medicine — RCT of eyeglass provision (may 2024). https://www.givewell.org/research/grants/london-school-hygiene-tropical-medicine-eyeglasses-may-2024, 2024. Accessed 2026-07-29.

[17] GiveWell. How do you value information? https://blog.givewell.org/2024/10/31/how-do-you-value-information/, 2024. Accessed 2026-07-29.

[18] GiveWell. Y-RISE — RCT of water entrepreneurship program grant (july 2024). https://www.givewell.org/research/grants/y-rise-rct-of-water-entrepreneurship-program-grant-july-2024, 2024. Accessed 2026-07-29.

[19] GiveWell. GiveWell's cost-effectiveness analyses — november 2025 version. https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models/november-2025-version, 2025. Accessed 2026-08-16.

[20] GiveWell. Development Innovation Lab and Innovations for Poverty Action — chlorine coverage surveys in Malawi and Uganda (march–april 2025). https://www.givewell.org/research/grants/development-innovation-lab-and-innovations-for-poverty-action-chlorine-coverage-surveys-in-malawi-and-uganda-march-april-2025, 2025. Accessed 2026-07-29.

[21] GiveWell. Lookback: Grant to Evidence Action's Dispensers for Safe Water program in Kenya, Uganda, and Malawi. https://www.givewell.org/research/lookbacks/Dispensers-for-Safe-Water-2025, 2025. Accessed 2026-07-29.

[22] GiveWell. All content on New Incentives. https://www.givewell.org/charities/new-incentives/all-content, 2025. Accessed 2026-07-29.

[23] GiveWell. GiveWell's cost-effectiveness analyses. https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models, 2026. Accessed 2026-08-16.

[24] GiveWell. Following the data on Dispensers for Safe Water (podcast episode 25 summary). https://blog.givewell.org/2026/03/05/podcast-episode-25-following-the-data-on-dispensers-for-safe-water/, 2026. Recorded 2026-02-20. Accessed 2026-07-29.

[25] James Michelson. Reflexive measurement. https://philpapers.org/rec/MICRM-4, 2023.

[26] Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.

[27] Juan Carlos Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pages 7599–7609, 2020.

[28] Dražen Prelec. A Bayesian truth serum for subjective data. *Science*, 306(5695):462–466, 2004.

[29] Robert M. Townsend. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory*, 21(2):265–293, 1979. doi: 10.1016/0022-0531(79)90031-0.