Benefits of modeling

Note: This document is a working draft. It will be refined in the near future.

Why Model Explicitly?

Deciding how to allocate resources across cause areas is one of the hardest problems in philanthropy. The factors are numerous, they interact in complex ways, and the stakes are enormous. In this post, we make the case that explicit models (even imperfect ones) are the best tool we have for navigating this complexity. We illustrate with concrete examples from our own work.

Why trust us: This post is part of a series from the cross-cause prioritization team at Rethink Priorities. Our companion posts cover what can go wrong with heuristic reasoning and offer a defense of explicit modeling against common criticisms.

Putting numbers on beliefs sharpens thinking

It's common to reason about cause allocation in qualitative terms: the *scale* of an area, its *tractability*, the *probability* that work bears fruit.[1] These are the right questions. But the specific numbers matter enormously.

Consider someone reasoning about wild animal welfare. They might think "the probability that work in this area succeeds is quite low." But *how* low? There's a vast difference between 10% and 0.001%. Until you try to put a number on it, you may not know what you actually believe.

Prior evidence from Hubbard, Tetlock, and GiveWell all point in the same direction: **"putting a number on it," even if highly uncertain, importantly improves your thinking.**

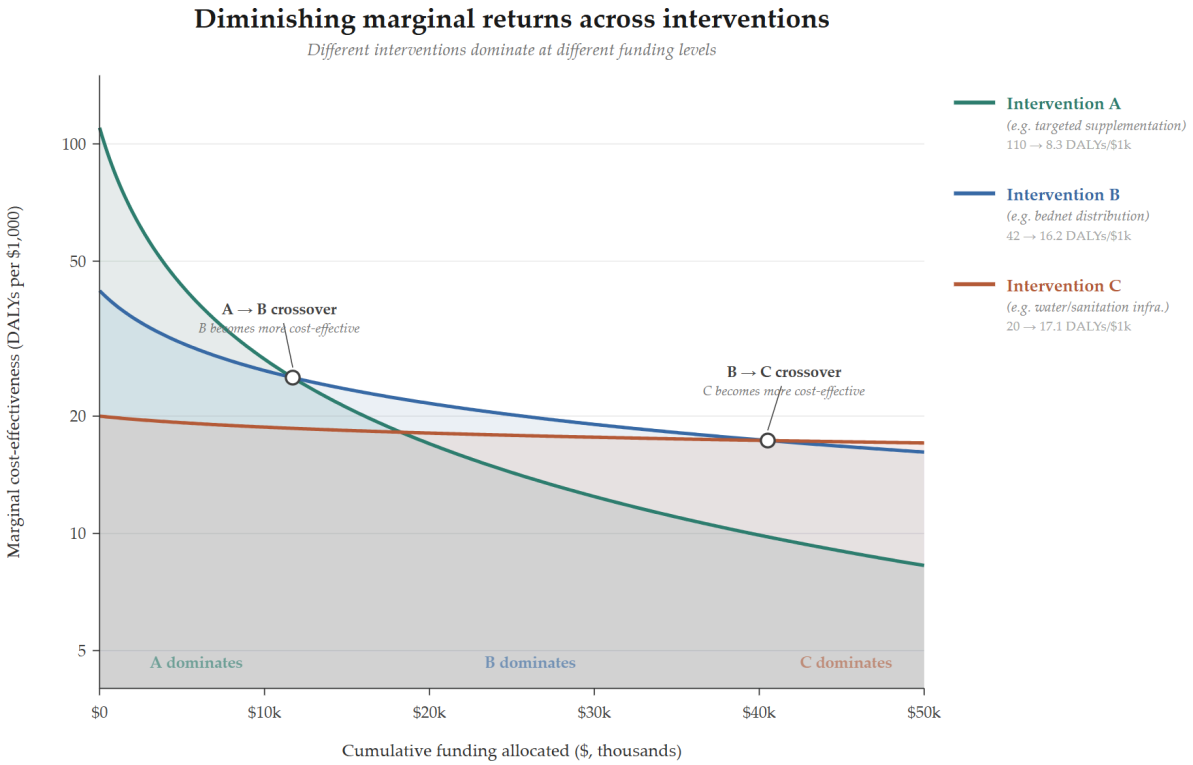
The payoff can be surprising. You might believe that the potential impact of AI safety or wild animal welfare work is huge, but that success is very unlikely, and conclude that these areas aren't worth much spending. Yet when you actually run the numbers, you may find that the expected value only drops significantly under implausibly low assumptions about the probability of success. The implications of your beliefs may not be obvious until you do the math.

---

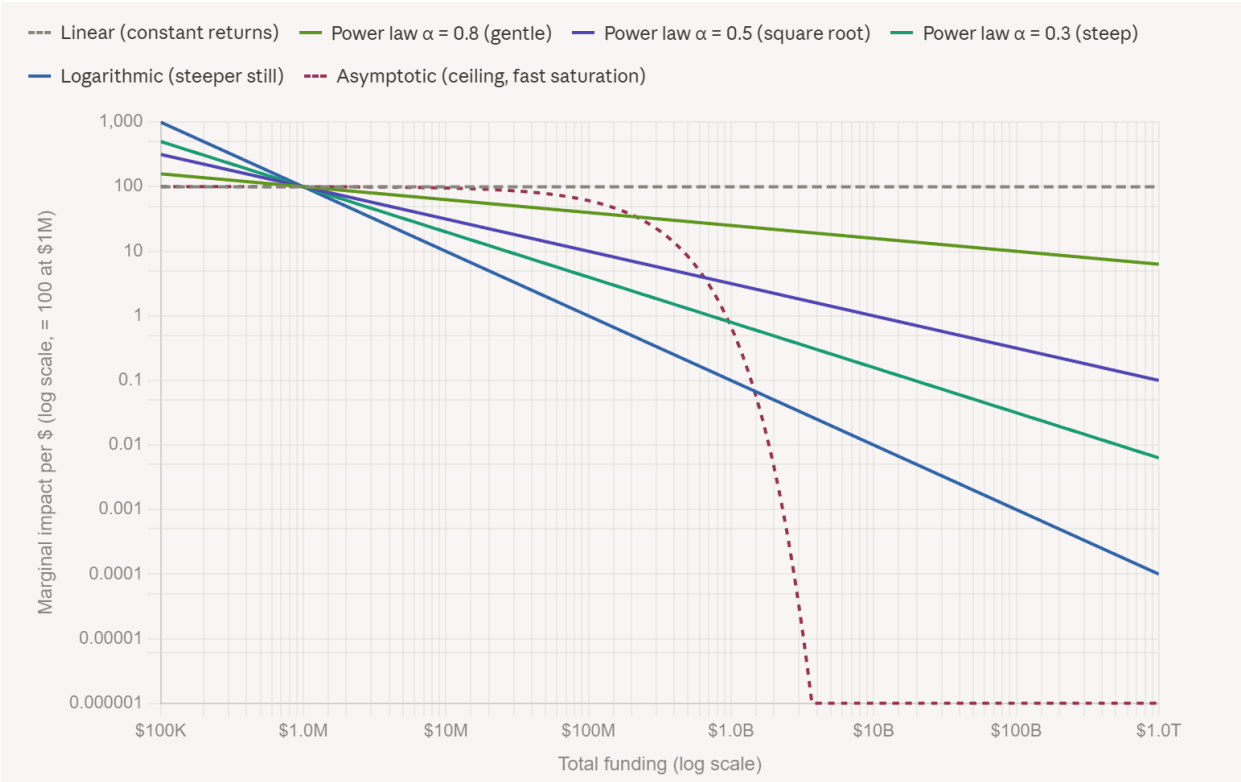
[1] This document is targeted at people who want to make the *best* decision they can about where to allocate their resources. We do not assume you necessarily want to do the *most good you can*. If you want to donate in a way that guarantees doing some good, or avoids doing harm, explicit models are valuable too. Note also that while we frame this in terms of allocating across funds (the scenario the Donor Compass covers), similar considerations apply to other resource-allocation cases, though there may be some differences (e.g., if you are donating directly to organizations within a specific area based on insider knowledge).And, of course, people's reasoning can be far more complex and sophisticated than the simple examples here. Donors may have considered many nuanced arguments about AI timelines, for example. But explicit models are no less helpful in these cases. Indeed, they may be *more* important when trying to weigh many complex considerations.

Models handle complexity that intuition cannot

Some of the factors that matter most for cause allocation are genuinely hard to reason about without a model. Diminishing returns is a good example. It's common to talk about the cost-effectiveness of donations on the *current* margin, but cost-effectiveness typically changes (potentially dramatically) as funding levels rise. The interaction between different rates of diminishing returns and different funding levels is very difficult to estimate informally.



An explicit model lets you see these implications directly. For instance, you might model diminishing returns as logarithmic, find that this implies cost-effectiveness drops 100x after $100m of funding, and then ask whether that seems too steep, too gentle, or about right.



Uncertainty is another area where models earn their keep. Cause prioritization is rife with uncertainty at every level (about how much good interventions would do, how likely they are to work, and how valuable their outcomes would be). A simple qualitative comparison or a back-of-the-envelope calculation with point estimates can only go so far. It cannot capture, for example, the difference between being unsure whether an intervention has a 5% or 10% chance of success and being unsure whether it has a 0.01% or 50% chance. Not just the *range* of uncertainty matters, but the shape of the distribution.

Rate each factor on a rough scale

Quick and intuitive. You can compare causes at a glance — but "High" importance and "Low" tractability can't be combined into a single number.

	IMPORTANCE	TRACTABILITY	NEGLECTEDNESS
• **AI safety**	High	Low	Medium
• **Global health**	Medium	High	Medium

> **Verdict:** *Unclear. Each cause has different strengths. There's no systematic way to weigh "High importance, Low tractability" against "Medium importance, High tractability."*

Assign numbers. Compute a ratio.

Now you can combine factors and directly compare. But single numbers hide how confident — or uncertain — you actually are.

```
CE = (impact if successful) × P(success) ÷ (cost per unit effort)
```

• AI safety

Impact	**10,000,000 QALYs**
×	
P(success)	**0.1%**
÷	
Cost/unit	**$100,000**
CE RATIO	**100 QALYs/$1k**

• Global health

Impact	**500 QALYs**
×	
P(success)	**85%**
÷	
Cost/unit	**$5,000**
CE RATIO	**85 QALYs/$1k**

Replace each number with a distribution

Every input is uncertain. Express that uncertainty honestly, propagate it through the calculation, and see what the output *actually* looks like.

```
Each input is a distribution → multiply/divide → output is a distribution
```

• AI safety

IMPACT IF SUCCESSFUL

median ~10M · spans orders of magnitude

P(SUCCESS)

median ~0.1% · could easily be 0.01% or 1%

COST PER UNIT EFFORT

median ~$100k · high variance

• Global health

IMPACT IF SUCCESSFUL

median ~500 · relatively well-studied

P(SUCCESS)

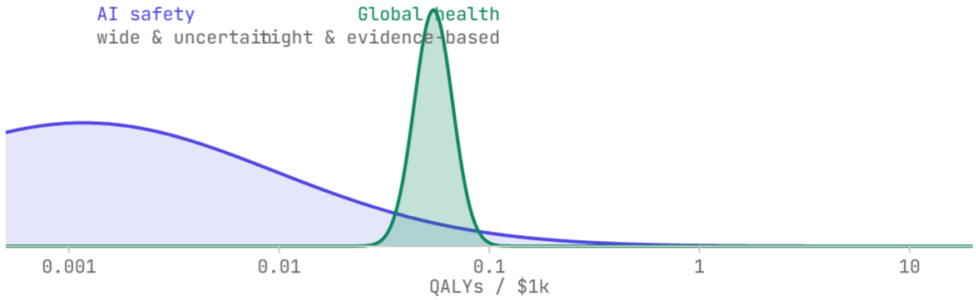
median ~85% · based on RCT evidence

COST PER UNIT EFFORT

median ~$5k · GiveWell-style estimates

The resulting CE distributions

Same formula, same medians — but now you can see what you're actually betting on



AI SAFETY · 90% CI

0.003 – 3.1 QALYs/$1k

~1,000× spread

GLOBAL HEALTH · 90% CI

0.04 – 0.08 QALYs/$1k

~2× spread

These issues compound once you consider the full set of factors that interact in cause-allocation decisions. Our own model, for instance, considers:

- **What does a dollar donated to a fund achieve?**
    - Cost per unit of impact
    - Diminishing returns
    - When the impact occurs (next 5 years vs. 20–100 years)
    - The risk profile of interventions (probability of success, backfire, or no effect).
- **How should we compare different outcomes?**
    - Human welfare tradeoffs (e.g., doubling income vs. saving a life)
    - Human-animal welfare tradeoffs
    - Risk attitudes (e.g., whether a 90% chance of saving 10 lives equals a 0.0001% chance of saving 90,000)
    - How AI risk should affect the valuation of other interventions.
- **How should we combine these valuations?**
    - For example, should you average across moral views, or give each view a slice of the budget?

Each of these is complex on its own. Together, the implications of their interactions are almost impossible to reason through without a model.

Models can reveal surprising conclusions

Because explicit models force you to trace the full implications of your beliefs through complex interactions, they can produce genuinely counterintuitive results. For example:

- In our prior work, we found that even a modest amount of risk aversion significantly reduced the value of global catastrophic risk spending (even when its cost-effectiveness was assumed to be high) and warranted diversification.
- Even if you assume that animal welfare matters much less than we think is reasonable, this does not reduce optimal animal-welfare spending as much as one might expect.
- When starting the Moral Weight Project, Bob Fischer expected it to imply large differences between non-human and human welfare. But after building the model,

he concluded that it would be hard to generate really big differences, given the assumptions.

These are insights that informal reasoning would be unlikely to surface.

Models promote transparency, discipline, and updatability

Beyond surfacing surprising implications, explicit models offer several further advantages.

They make it harder to fudge. When reasoning informally, it is easy to subtly (and unconsciously) adjust your reasoning to produce the conclusion you would antecedently find plausible. Explicit models, by requiring that you specify numbers and assumptions and then see what they imply, reduce this risk.[2]

They make it easier to find and fix errors. Because an explicit model shows you what is driving results, you can run sensitivity analyses to see which assumptions matter most, and refine the ones that seem wrong. This should increase confidence in the model's conclusions over time.

They make productive disagreement possible. If you hold different views from the model's defaults, you can swap in your own assumptions and see what follows. For example, our model allows you to see what would be recommended if you give 100% weight to total utilitarianism. This makes it clear to people with different views what is actually driving their conclusions.

They are easy to update. As circumstances change (for instance, as funding levels shift and diminishing returns kick in) an explicit model can be readily updated. In principle, you can update informal reasoning too. But an explicit model, by making clear what you have or haven't updated, encourages you to actually do so.

---

[2] This is not to say you should uncritically follow whatever a model produces, but that if you revise its assumptions, you should do so explicitly and reflectively.

What This Means for Your Giving

As we argue in our [post on the value of prioritization], the stakes of resource-allocation decisions are high, with potentially vast differences in what we can achieve. The questions are enormously complex. **Explicit models may not be perfect, but they offer the best chance of making the best decisions we can.** We outline the case in more detail in our post In Defense of Modeling and our post on what can go wrong with heuristic reasoning.