Paul de Font-Reaulx
paul.defontreaulx@gmail.com

Virtuous

tl;dr

This is a pitch for Virtuous: a nonprofit dedicated to making complex but desirable model behavior legible to frontier labs through evals. It would research and curate high-quality public benchmarks of normatively important behavior with the aim of steering AI development towards virtue.

Problem: Capturing Complexity in Evals

The nature of future AI models will have an enormous impact on our well-being, yet we cannot assume they will behave as we would like by default. We can influence this by providing evals that assess whether models demonstrate what we want. A verifiable benchmark provides a target for frontier labs to aim for, making it legible. It can also become an important focal point to determine legal responsibility (e.g. mental health benchmark to assess liability for user harm), giving labs an incentive to perform well on it.

Many behaviors we want models to demonstrate are complex behaviors, such as whether they are truthful or empower their users. However, such evals are inherently difficult to design because it is unclear what we want to capture. Consider truthfulness. Is it enough to say only true things to be truthful? What if you cherry-pick all the bad truths, and omit the good truths about the domain? Something more needs to be said, as I argue at greater length here.

The complexity of virtues does not make them less important. But without a concerted effort to represent them as measurable evals, they risk becoming neglected and invisible. Alternatively, they might be represented in inadequate evals, inviting severe Goodharting. This requires a dedicated intervention.

Proposal: A Center for Virtue Evals

Basic Idea: Virtuous will be a research non-profit org providing high-quality evaluations of virtues that are otherwise unlikely to be developed with sufficient care. These might include epistemic virtues (e.g., “Are models truthful?”), societal virtues (e.g., “Do models manipulate users?”), and ethical virtues (e.g., “Do models contribute to user flourishing?”).

Focus: The purpose is to elucidate what we care ultimately about under the guise of intuitive virtues, and scout paths to possible operationalizations. The aim is to be a primary provider of public high-quality evals of complex virtues, giving frontier labs desirable targets and incentives to meet them. This is naturally done in collaboration with the broader eval community, aiming to

complement rather than compete with existing organizations (e.g. CIL's Weval, Transluce, and MLCommons).

Impact: High-quality evals are likely to be particularly impactful if provided soon. We can still steer the development of frontier models, and the choices we make now can generate path dependencies for our long-term future.

Plans

The overall strategy is to do scalable theory-heavy labor to permit the design of high-quality evals of particularly important virtues. The implementation will develop across time:

Short-term: The aim is to design select high-impact evals fast, focusing on flagship examples to establish industry buy-in. A proposed workflow:

1. Solicit input from stakeholders about what virtue evals would be most important to have
2. Elucidate the normative core of these virtues; clarify what an eval should reflect
3. Propose operationalizations that closely capture this

In practice, this can happen either with an internal team, researchers recruited on a project basis, or by an open call for proposals of evals from relevant domain experts (e.g. from academics in philosophy, economics, or psychology in exchange for grants). This is likely to be particularly scalable, and might have the positive externality of engaging more people. Labor costs are additionally likely to be comparatively cheap.

Long-term: It is likely AI will soon be able to substantially help with this research. From early on, we should set up infrastructure and experiments of how frontier models can help design the best possible evals, e.g. by helping to elucidate the normatively crucial features of intuitive virtues, or even improve on them (e.g. as we have largely dismissed honor as a virtue, viewing it as a defective signalling game). This is likely to be a particularly scalable method of generating virtue evals.

Risks & Mitigations

- **Complexity:** Some virtues may prove too complex to benchmark effectively. **Mitigation:** Even if true, this would be an unfortunate reality, not a reason to avoid a concerted effort. Partial legibility is better than none, and this effort is the best way to discover the limits.
- **Insufficient Buy-In:** Labs may not adopt our evals. **Mitigation:** We will work to secure interest and buy-in from stakeholders early, focusing on in-demand evals and developing wide-ranging collaborations.
- **Contentious Normative Stances:** Operationalizing virtues can be ethically and politically contentious. **Mitigation:** This risk is real and requires constant vigilance. We will strive to balance democratic input with the focused momentum of a small team to mitigate imposing problematic stances.

Paul de Font-Reaulx
paul.defontreaulx@gmail.com

Logistics

Virtuous is proposed as an independent non-profit organization. By default, it would be founded by Paul de Font-Reaulx while recruiting co-founders and collaborators. Paul has a comparative advantage in theory-heavy research, organizational experience, and strong networks within academia for potential recruitment.

If founded, Virtuous would aim to launch under a fiscal sponsor, as receiving 501(c)(3) designation takes many months. This might also require the sponsoring organization to facilitate an O-1 visa application for Paul. There are also scenarios where the basic idea is viable but better implemented as part of existing organizations. Finally, there are plausible synergies with related but independent projects (e.g. *DeliberationBench*).

Founding Virtuous will require seed-funding to support 2-3 full-time employees for at least 18 months. If you are interested in the idea and would consider funding the organization, please contact Paul.