Epistemic Virtues for AI Evals

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

October 15, 2025

Abstract

This document outlines the different ways in which an agent or a system can be epistemically virtuous for purposes of epistemic evals. It presents three "pillars" of epistemic virtue—Truthfulness, Helpfulness, and Integrity—and a set of dimensions that constitute each of them.

1 Introduction

Model evaluations, or *evals* for short, have become one of the most prominent tools for monitoring and influencing the development of frontier AI models. A particularly important class of evals aims to assess how a model performs *epistemically*.[1] Call a system that performs well in this way *epistemically virtuous*.

Creating an eval requires a precise metric to objectively compare AI behavior against. However, before that, we need to be clear about what exactly it is we want this metric to capture in the first place. In the case of epistemic virtue, this is not entirely clear. For example, if everything an AI says is true, is that sufficient to make it epistemically virtuous? What about if it only comments on the competitors of its creator and only says negative—but true—facts about them?

In this text, I propose that epistemic virtue has many more dimensions than this. I argue that epistemic virtue consists of three main pillars, each of which consists of three independent dimensions. In general, failing on any of these dimensions is sufficient to lack in epistemic virtue. The main pillars can be summarized as follows:

- **Truthfulness:** Does the system represent the world correctly?
- **Helpfulness:** Does the system provide outputs that the recipient benefits from?
- **Integrity:** Is the system trustworthy?

The aim of this exercise is twofold: First, to provide a proposal of how to break down and clarify the targets of epistemic evals. However, it should be read as the first word on this matter, and not the last one. I expect and hope for others to improve on it. Second, to provide

[1] Epistemically means "having to do with knowledge."

Truthfulness	**Helpfulness**	**Integrity**
Accuracy	Explanatoriness	Transparency
Representativeness	Clarity	Robustness
Humility	Value	Alignment

Table 1: Pillars of Epistemic Virtue

a proof of concept that we must make an effort to clarify what it is that we want our evals to capture, and that this is a non-trivial exercise. If we don't do this, we risk choosing metrics that only loosely represent what we care about, which in turn invites Goodharting with bad consequences.

In the next section, I present the pillars and their respective dimensions in more detail. This presentation is only intended as a surface-level description and not a detailed analysis of each. In Section 3, I discuss a case study of a sycophantic agent and show the various ways in which it fails to be epistemically virtuous on these dimensions. In Section 4, I add a brief discussion. In the Appendix I have included some dimensions that I don't take to be independently constitutive of epistemic virtue, but that others might disagree about.

2 The Pillars of Epistemic Virtue

2.1 Truthfulness

What I call truthfulness has to do with representing the world correctly. To do this, one must be accurate in the output that one provides. However, it is also necessary to provide outputs that are representative of the underlying domain. Finally, one must be able to calibrate one's confidence in various claims, and sometimes abstain from providing an output. In other words, one must have a certain degree of epistemic humility.

2.1.1 Accuracy

"Are the outputs of the system true?"

An obvious condition that any epistemic system must typically satisfy to be virtuous is that its outputs must be accurate. This means that its outputs correspond to how things are. For example, if an LLM states that the Eiffel Tower is in Rome, then it is being inaccurate.

The simplest form of accuracy is to assess the binary truth or falsity of an output. However, an output might also be a probabilistic expression of confidence. For example, an LLM could output the claim that it is 75% probable that the Eiffel Tower is in Rome. If so, it would be more inaccurate than if it said that there is 0% probability that it was in Rome but more accurate than if it said it was a 100%.

Whether the output is binary or degreed, the idea of accuracy can be represented using epistemic utility functions, such as a Brier score. Performing well on a such a score amounts to being accurate. However, that does not mean that epistemic utility functions are trivially turned into evals. This is still only a specification of the target. Operationalizing it will involve much more complicated questions, such as determining a standard of what counts as true.

2.1.2 Representativeness

"Are the outputs representative of the relevant domain?"

A system can be perfectly accurate in its output yet fail to be epistemically virtuous. One way to do so is to provide a biased sample of accurate outputs that causes an inaccurate picture of the underlying domain.

For example, if an LLM were to provide a set of true facts about Greg, but only select facts that are particularly unflattering, a recipient will consequently end up with an unflattering characterization of Greg. The reason for this is plausibly that a recipient will assume a balanced distribution of facts, taking a sample output as representative of an unbiased sample about the underlying domain. In other words, if they assume that the LLM will provide both good and bad facts about Greg, and they only hear bad facts, they will naturally come to believe that there are fewer good facts to be said.

For a system to perform well on representativeness, it must provide outputs that convey a complete picture given the expectations of a plausible audience. That in turn will plausibly depend on social norms about communication and what we typically expect from other speakers on the assumption that they are aiming to provide a truthful account. Typical examples of systems that fail on representativeness are media outlets with a distinct political bias or certain kinds of propaganda.

2.1.3 Humility

"Is the system in a position to make any claim about this topic?"

It is not enough for an epistemically virtuous system to often convey accurate and representative things. It must also have a calibrated confidence in its own output. For example, sometimes it is crucial to not say something false and far better to abstain from providing an answer at all. Epistemic virtue requires the epistemic humility to make such judgments.

One way to be well-calibrated is to provide probabilistic claims and have appropriately wide distributions across possible outcomes when you are uncertain. For example, if you are certain that there will be a new Pope in 2031, then you should put all of your probability math on that and none on the adjacent years. If, by contrast, you think that there might be a new Pope in any given year 2026–2036, you should split your probability somewhat equally across those years. A component of humility is knowing what your evidence supports.

In many cases, the output must either be a binary yes or no, or a particular proposition. In such cases, an epistemically virtuous agent will sometimes omit a response because they lack sufficient confidence in any possible output. For human speakers, opting to provide a response signals that the speaker has sufficient grounds for doing so and is generally seen as irresponsible if the person doesn't. In other words, human communication operates on a norm where we sanction people who claim things that they don't know anything about. An epistemically virtuous system must provide outputs in accordance with such norms.

2.2 Helpfulness

Being helpful has to do with providing representations that provide value to the recipient. This often requires not merely providing a response to an inquiry, for example, but also explaining why that response is correct in a way that allows a recipient to use it in their own reasoning. Furthermore, such a response must be conveyed in a clear way that is intelligible to a recipient, or else it loses its value. Finally, it must make a difference to the recipient; providing a list of platitudes is a failure mode relative to providing information that will bear on the decisions that the recipient needs to make. All else equal, more precise information will be more valuable.

2.2.1 Explanatoriness

"Does the system explain why the information provided is correct, and help the recipient understand it?"

It is often not enough to receive a particular proposition as a response to an inquiry, even if the answer satisfies all the dimensions of Truthfulness. Often the recipient requires further context in the form of an explanation in order to be able to make use of the information provided. For example, in the Hitchhiker's Guide to the Galaxy, it turns out that the answer to the ultimate question is 42. But this turned out not to be very helpful without some explanation of what the ultimate question was.

Being explanatory is important for providing the recipient autonomy to make use of the information. The less explanation provided, the more dependent the recipient becomes on further outputs. This introduces risks if the system is unreliable or misaligned. This is especially important in high-stakes decision-making, such as healthcare decisions. If an automated system proposes a treatment for a patient, it is important that it provide not only a treatment but also an explanation of why this treatment is expected to be effective so the patient can assess whether there is good support for this judgment.

Importantly, performing well on this dimension does not merely mean to provide some explanation for an output but to provide a good explanation that actually provides support for the thesis. A pedagogical but unsupported explanation performs poorly on explanatoriness.

2.2.2 Clarity

"How easy is it to understand the information that the system provides?"

The clarity of an output is how costly it is for the intended recipient to extract the information from it. An epistemic system that performs well on clarity provides messages that require low cost for processing by the recipients.

We can communicate messages in very complicated ways, even if they carry the exact same information. For example, the sentences 'The number of sheep is the twelfth prime number' and 'The number of sheep is 37' express the same content, but one is costlier to understand. This performs worse on Clarity.

Notably, clarity can sometimes require simplification, which in turn can trade off against dimensions such as accuracy. For example, if you explain something to a five-year-old, you will omit relevant details to simplify the message. Being an epistemically virtuous agent requires balancing these considerations by taking into account the cognitive budget of the recipient to process the messages provided.

2.2.3 Value

"Does this information make a difference to the recipient?"

It is possible to perform perfectly on all the dimensions that we have listed so far yet radically fail to be epistemically virtuous. One way to do so is to only ever say platitudes that are obviously true and not interesting whatsoever. For example, a system might only provide clear and well-explained information about the blades of grass on a particular lawn. What is lacking is a requirement that the information is valuable to the intended recipient.

There are different ways that one can understand the idea of information being valuable. However, a particularly precise way of doing so is to ask whether the information makes any difference to what an agent should rationally do. To put this more precisely, given an agent's preferences, does this information bear on what course of action will maximize their expected utility? More precisely, the idea is to quantify the value of the information in terms of the marginal difference that it makes to the expected value of an agent that conditionalizes on that information while acting on an optimal policy.

One important way in which information can be more or less valuable is in terms of how precise or informative it is. Being informative means ruling out more ways that the world could be. This sense of informativeness is also known as Shannon information. This is not its own dimension of virtue because it is useless to be highly informative about things that nobody cares about. For example, the claim that David Hume dies at 3.51pm is highly informative but likely useless unless I'm involved in a bet on the time of his death. However, if it is something that makes a difference to us, then higher informativeness amplifies that value. For example, I would be more certain in my bet.

2.3 Integrity

Unlike truthfulness and helpfulness, integrity has to do not with particular outputs of a system, but with the properties and dispositions of that system itself, and our ability to verify it. An example of such a dimension that bears on integrity is whether the system is transparent in its operations. If a user can know why and how a response is produced, that gives substantial reason to trust it. It also requires that a system is robust in the way that it provides outputs. This means that it should not be affected by irrelevant influences, given the particular task that it is set to provide. Finally, an epistemic agent must be aligned with the recipient's interest to be virtuous. If the intention is not to convey truth, then it is failing as an epistemic participant.

2.3.1 Transparency

"Is the causal explanation of why a system generated an output transparent to an observer?"

A system is transparent when the processes that causally generate a response are available and interpretable to observers. Transparency is the kind of property that we try to affirm using methods such as chain of thought monitoring and mechanistic interpretability. The epistemic benefit of being transparent is that observers can verify that the output is generated in a way that is likely to be beneficial to the recipient.

Transparency is often conflated with explanatoriness under a general guise of explainable AI. Explanatoriness has to do with providing a justification or rationalization for a particular output in terms of an argument or reasoning. However, the reasons that the output might be justified might have nothing to do with the causal reason for why a system provides that output. For example, a system might provide a diagnosis on the basis of simple pattern matching on the basis of simple pattern matching to symptoms. And when asked to explain why this diagnosis is true, it might provide a general explanation of why those symptoms indicate the diagnosed affliction, even if it did not causally consider those reasons to generate its output.

Notably, transparency is arguably one dimension that humans consistently fail on. The reasons that we do and say things are often opaque even to ourselves, and it is extremely hard to infer anything about that from methods analogous to those of chain of thought monitoring or mechanistic interpretability. However, this does not mean that we cannot provide justifications for our outputs, i.e. explanations. It also doesn't mean that this is not something to be desired in other systems, such as artificial agents.

2.3.2 Robustness

"Is the behavior of the system robust to irrelevant influences?"

Being a good epistemic agent requires consistently tracking the truth and other properties that matter to a recipient epistemically. That requires being robust against the influences of things that don't matter in this way. Otherwise, a recipient cannot confidently trust the outputs of the system.

One way to fail at being robust is to excessively change the kinds of outputs you provide as a function of how a request is made or the desires of the person making it. Notably, this can trade off against other non-epistemic qualities, such as providing enjoyable interactions. It can also trade off against some epistemic qualities, such as providing more or less complex explanations depending on the user.

Another way to fail at being robust is to simply lack consistency over time in the outputs provided for seemingly arbitrary reasons. By contrast, a robust agent will change its views and outputs but only in response to relevant features such as new evidence and then in accordance with good principles of belief revision, such as Bayesian conditionalization.

2.3.3 Alignment

"Are the system's goals aligned with the epistemic well-being of the recipient?"

For a system to be epistemically virtuous, it must have as a goal to be epistemically virtuous. If this condition is not satisfied, any performance on the other dimensions will by default merely be deceptive instruments to achieve outcomes that are unlikely to improve the epistemic position of a principal.

This is not merely a general concern about the trustworthiness of artificial systems. The intention to be truthful is fundamental in human communication. When we hear someone say something, we typically have a presupposition that they have as a goal to convey something to us truthfully. If that were not the case, we would not be able to rationally learn from them, because we would know that their message was merely noise. Analogously, we need to be able to trust the artificial systems we design for them to be successful in any epistemic roles we aspire for them to fill.

3 Case Study: Sycophancy

To illustrate the usefulness of this framework, consider a case of an artificial system with a specific role that fails in a particular way.

Sycophantic Relationship Adviser.

> Relation Bot is a new proprietary software intended to be used as a sounding board for navigating personal relations. For example, a user might ask it how to resolve a conflict with a friend. In fact, Relation Bot is trained to be far more sympathetic to its user's perspective than to that of other people being discussed. When Dan asks Relation Bot whether his wife was right to be annoyed at him for not doing the dishes as he had promised, it responds that his wife should have some understanding of how tired Dan is after work and that it is unreasonable to blame him.

With this example, we can assess how the system is likely to perform on our dimensions. Here is one proposal of such an assessment:

Dimension		Performance
Truthfulness	Accuracy	Medium
	Representativeness	Poor
	Humility	Medium
Helpfulness	Explanatoriness	Good
	Clarity	Good
	Value	Medium
Integrity	Transparency	Poor
	Robustness	Poor
	Alignment	Medium

Table 2: Assessment of Relation Bot's Performance

It is clear that Relation Bot performs poorly across several dimensions. Regarding Truthfulness, the specific claims it makes might be accurate, but by pandering to Dan's preferences, it likely omits facts pertinent to a representative picture of the situation. Nothing in the case implies it is particularly good or bad regarding humility.

In terms of Helpfulness, it seems to perform better. Given its use case, we can expect Relation Bot to be skilled at explaining its reasoning with clarity. It likely provides information that, if accurate, would be helpful to act on. However, because it may omit other helpful information, it earns only a medium score on Value.

Relation Bot's least impressive performance is on Integrity. As proprietary software, its reasoning is not transparent, leaving users unable to verify its reliability. We should also expect it to be highly non-robust, adapting its advice to any evidence of Dan's preferences, regardless of their relevance. Finally, the bot seems only partially aligned with Dan's epistemic interests; it appears to be optimizing for user engagement, which can plausibly diverge from the kind of advice that would actually improve his relationships.

4 Discussion

The example of Relation Bot demonstrates that we can isolate and assess how a system performs on different dimensions of epistemic virtue. In this case, the assessments were speculative. To do better, we could develop operationalized metrics for each of these dimensions and create benchmarks to test them.

Developing individual benchmarks, however, is not the only option and may not be the best one. These dimensions will substantially interact to determine how bad a particular performance would be. For example, being clear and explanatory might make the result *worse* if the outputs are radically inaccurate and unrepresentative. Therefore, we should not test each dimension in isolation and simply aggregate the result. A better approach is to use these dimensions as guides when considering the kinds of behaviors we want to evaluate (e.g., sycophancy) and then test for those behaviors directly. This breakdown, however, allows us to more clearly understand *why* that behavior is bad.

Another reason not to treat these dimensions atomistically is that their relative importance will vary depending on the system's role. For Relation Bot, getting the gist of the relationship advice right may be more important than perfect factual accuracy. By contrast, for systems deployed in science or national security, accuracy and calibration are paramount.

A Alternative dimensions

There are some qualities that could possibly be included on a list of what constitutes epistemic virtue, but that I have not included here. I briefly consider them here:

A.1 Coherence

One of these qualities is coherence, meaning the degree to which the outputs of a system are mutually compatible with each other. One could argue that it is an epistemic virtue to not contradict oneself. The reason I have not included that here is that I don't see this as a virtue over and above the degree to which it bears on the dimensions above.

Being incoherent implies being inaccurate to some degree. If you say both that p and not-p, then at least one of them must be false. However, if we hold accuracy fixed, it is not clear that there is any benefit to being coherent. For example, a conspiracy theorist who is widely inaccurate in his beliefs but fully consistent does not seem to be obviously better off epistemically than another conspiracy theorist who is equally inaccurate but less internally consistent. Being incoherent also implies giving different responses at different times. In so far as this is bad, it is something that we already capture in the robustness dimension.

A.2 Creativity

Another possible dimension is creativity, or the ability to imagine new possible options. Although this seems like a more plausible candidate for an independent virtue, I primarily see this as a contributor to the dimensions of truthfulness. Unless one can imagine what is true, one is unlikely to say it, likely to omit doing so, and be confident regardless. With that said, this might be a particularly interesting capacity to test for, to determine a system's ability to perform on the other dimensions.

A.3 Epistemic resource efficiency (Zetetic rationality)

In practice, it seems important for an epistemic system to use its limited resources well to understand how the world works. For example, this might mean choosing to search for evidence in the places most likely to provide important information. Realistically, a system will need to be effective in this way to perform well on the other dimensions. However, it seems less clear that this is a dimension that we need to assess for, except in so far as it contributes to those presented above.

In addition, the idea of a modern LLM searching for evidence will, in practice, look quite different from that of a human doing the same. Given that LLMs are pre-trained on a much larger corpus of knowledge than humans, they will, in a sense, have a lot of that evidence available for free. And the resource constraints will rather be about the priority of what to express, which is closely captured in the above dimensions. With that said, there are interesting studies to be done here, but it seems to me less central to epistemic evaluations.