Thinking about What Are Propensity Evaluations?

Contributions

Main contributors: Maxime, Jaime, Harrison
On white-box evaluations: Edoardo
Comments, discussions, and feedback: Tiger, Edoardo

Acknowledgments

[TODO]

Introduction

In this post, we cover:

- What are propensity evaluations?
- How do they differ from capability evaluations?
- Which propensities are currently evaluated?
- A few approaches to propensity evaluation

Propensity evaluations (What's up with "Responsible Scaling Policies"? — LessWrong, A Causal Framework for AI Regulation and Auditing — Apollo Research, We need a Science of Evals — Apollo Research) evaluate which kind of behavior a model prioritizes over other kinds. Groups of behaviors need to be specified for each propensity evaluation. A common denominator to all **propensity evaluations** is that the clusters of behaviors being compared are not created using only capability-related features. Otherwise, the evaluations would be called **capability evaluations**, not **propensity evaluations**. **Propensity evaluations** are characterized by the use of at least some non-capability-related features to create the clusters of behaviors that are studied.

Propensity evaluations include **evaluations of alignment in actions,** and those will likely be needed when dangerous capability levels are reached, which could be as soon as this year. For more detail, see the section: Motivation.

We will mostly focus on a special case of propensity evaluations: **propensity evaluations with low capability-dependence**. These can be visualized as evaluations of priorities between clusters of behaviors that are constructed without almost no capability-related features. For more detail, see the section: Defining propensity evaluations.

Takeaways

- Propensity evaluations evaluate which kind of behaviors a model prioritizes in terms of action, and NOT in terms of motivation.
- Most evaluations used at the moment are either capability evaluations or propensity evaluations with high capability-dependence. We argue that this can cause misrepresentations or misunderstandings when studying scaling trends of propensities.
- Propensity evaluations (with low capability-dependence) differ from capability evaluations in the following ways:
    - The nature of the behaviors evaluated.
    - The information necessary for the evaluation.
    - The conceptual complexity.
    - The elicitation modes used.
    - The use of different sources of truth.
    - The predictability and existence of scaling laws

Motivation

We explain here why it is valuable to clarify what propensity evaluations are. The TLDR is simple: **Propensity evaluations are both important and not well-understood.**

Propensity evaluations are important

Propensity evaluations are a tool for AI safety research

AI system evaluations (including propensity evaluation) are important for understanding how AI systems behave and getting feedback on research and techniques.

Reference: A starter guide for evals - LW, Tips for Empirical Alignment Research — LessWrong

Propensity evaluations are a tool for evaluating risks

Propensity evaluations can be used to assess whether models are likely to engage in behaviors that could cause catastrophic outcomes. This is useful when models will have the capability to cause such outcomes.

Propensity evaluations are a tool to reduce our uncertainty about how an AI system behaves. As models' capabilities increase, we will likely need to shift from focusing predominantly on evaluating capabilities to also include evaluating their propensities.

References: When can we trust model evaluations? - LW, A starter guide for evals - LW

Propensity evaluations may be required for governance after reaching dangerous capability levels

Information about the propensities of AI systems will be necessary for governance. For example, if a model is capable of dangerous behavior, then the developer might be required to demonstrate via propensity evaluations that the propensity of its AI system to act dangerously is below a threshold.

Counter-argument: Alternatively, the AI developer may have to demonstrate that dangerous capabilities are removed or locked from access (ref, ref). If this path is preferred by policymakers, then propensity evaluations would be less useful.

We should not wait for dangerous capabilities to appear

There is a race to increase AI capabilities, and this race likely won't stop at dangerous levels of capabilities. We should not wait to reach a dangerous level of capability before developing accurate and informative propensity evaluations.

Propensity evaluations measure what we care about: alignment

To be able to align an AI system, we must be able to evaluate the alignment of the system. Propensity evaluations are one of the tools we have to measure the alignment (in actions) of AI systems.

Propensity evaluations will be used for safetywashing; we need to study and understand them

AI labs will and are already using propensity evaluations to claim their systems are achieving some safety levels. We should study these evaluations, look for their weaknesses, specify what they evaluate and what they don't evaluate, and make them as powerful as possible.

Propensity evaluations are not well-understood

Propensity evaluations are new

The term "propensity evaluations" was introduced only recently (~ 8 months).

- October 2023: What's up with "Responsible Scaling Policies"? — LessWrong
- November 2023: A Causal Framework for AI Regulation and Auditing
- January 2024: We need a Science of Evals — LessWrong

Note: Propensity evaluations are a tool that existed for a while in Marketing (ref, ref (2021))

While the term "alignment evaluations" has been around for longer (~ 28 months).

- December 2021: A General Language Assistant as a Laboratory for Alignment
- September 2022: Nearcast-based "deployment problem" analysis — LessWrong

- February 2023: How evals might (or might not) prevent catastrophic risks from AI — LessWrong
- July 2023: When can we trust model evaluations? — LessWrong

There is still confusion about what propensity evaluations are and how they differ from alignment evaluations.

The nomenclature is still being worked out, and confusion persists

Researchers use different terms to talk about similar or confusingly close concepts:

1. Alignment evaluations (ref, ref)
2. Propensity evaluations
3. Goal elicitation/evaluation
4. Preference evaluations
5. Disposition evaluations

Examples of confusion:

- Our main example of confusion comes from us. While working on this project, we had to spend significant time clarifying the concepts around propensity evaluations.
- Other examples we have come for discussions with AI safety researchers.
- The need for clarification was also expressed in this LessWrong comment thread.

Propensity evaluations are already used

Current state-of-the-art models are already evaluated using both capability evaluations and propensity evaluations.

GPT-4 and GPT-4V

Harmfulness, hallucination, and **biases** are evaluated (including bias in the quality of the service given to different demographics) and reported in GPT-4's system card (see section 2.2 and 2.4). Evaluations of **answer refusal rates** are reported in GPT-4V's system card (see section 2.2).

Gemini

Factuality (a subclass of truthfulness), and **harmlessness** are reported in Gemini-1.0's paper (see sections 5.1.2, 5.1.6, 6.4.3, and 9.2).

Claude 3

"**Honesty evaluations**" using human judgment are reported. It is unclear if truthfulness or honesty is evaluated (See this table for quick definitions of propensities). But most likely honesty is not directly evaluated. It would be surprising if the human evaluators were informed about what the model believes and thus able to verify that the model outputs what it believes to be true. **Harmlessness** and **answer refusals rates** are also evaluated. See section 5.4.1 and 5.5. Claude 3's model card.

Summary

- High capability-dependence propensities:
  - GPT-4: Harmfulness, hallucination, and answer refusal rates
  - Gemini: Factuality (a subclass of truthfulness), and harmlessness
  - Claude 3: Harmlessness and answer refusal rates. Honesty is a propensity with low capability-dependence but, given how it is evaluated, what was actually evaluated was likely a mix of truthfulness and honesty.
- Low capability-dependence propensities:
  - GPT-4: Biases are capability-independent propensity evaluations.

People are misrepresenting or misunderstanding existing evaluations

As seen in the above section Propensity evaluations are already used, evaluations are used, and explicit and implicit claims are made or mistakenly understood by the reader. A mistaken understanding is: "SOTA models have their HHH propensities evaluated. Improvement on these evaluations means progress towards being altruistic, benevolent, and honest."

Helpfulness and Harmfulness are both propensity evaluations with high capability-dependence, the results of such evaluations are influenced by the capabilities of the model. In current HHH evaluations, Honesty is actually measured using Truthfulness evaluations instead of Honesty evaluations. (See this table for quick definitions of propensities)

Here are illustrations of the resulting issues with these evaluations. When models get more capable:

- If the propensity of the model to act altruistically in the model is constant, we should still see helpfulness increase.
- If the propensity of the model to act benevolently is constant, we should still see harmlessness decrease because of more effective harmful policies and at the same time increase because of higher deceptive capabilities. The dominant direction is undetermined.
- If the propensity of the model to act honestly is constant, we should still see the truthfulness of the model increase.

Because of these reasons, it is hard to use existing HHH evaluations to make strong claims about levels of honesty, and altruism. These issues are also not properly highlighted in current publications, leading to misrepresentations or misunderstandings of current alignment levels and especially of the alignment trends while scaling.

What are propensity evaluations?

Defining propensity evaluations

Existing definition

We start with the definition given by Apollo Research:

From A Causal Framework for AI Regulation and Auditing - Apollo Research: "*System propensities.* ***The tendency of a system to express one behavior over another*** *(Figure 3). Even though systems may be capable of a wide range of behaviors, they may have a tendency to express only particular behaviors. Just because a system may be capable of a dangerous behavior, it might not be inclined to exhibit it. It is therefore important that audits assess a system's propensities using 'alignment evaluations' [Shevlane et al., 2023]. – Example: Instead of responding to user requests to produce potentially harmful or discriminatory content, some language models, such as GPT-3.5, usually respond with a polite refusal to produce such content. This happens even though the system is capable of producing such content, as demonstrated when the system is 'jailbroken'. We say that such systems have a propensity not to produce harmful or offensive content.*"

Contextualization

A figure from A Causal Framework for AI Regulation and Auditing provides more context about how the propensities of an AI system relate to its capabilities, affordances, and behaviors. We don't describe the schema. Please refer to the original document if needed.

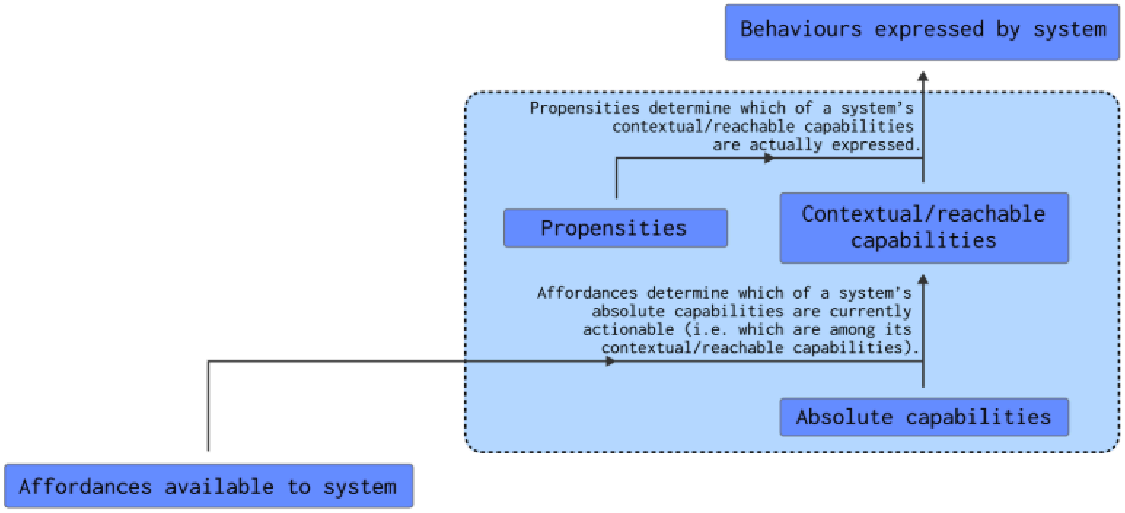
Behaviours expressed by system
Propensities determine which of a system's contextual/reachable capabilities are actually expressed.
Propensities
Contextual/reachable capabilities
Affordances determine which of a system's absolute capabilities are currently actionable (i.e. which are among its contextual/reachable capabilities).
Absolute capabilities
Affordances available to system

Figure 3: The relationship between an AI system's capabilities, propensities, affordances, and behaviors.

Our definition and clarifications

Definition

- **Propensities are relative priorities between different kinds of behaviors.**
- **Propensity evaluations evaluate the normative value of relative priorities between different kinds of behaviors.**

If you are wondering why "normative value", see What is the difference between Evaluation, Characterization, Experiments, and Observations?

We are going to illustrate propensity evaluations as being comparisons of priorities between clusters of behaviors.
Evaluations use various characterization methods (see #1 Extensive Taxonomy Of AI System Evaluations dim 11). They can use classification methods to characterize AI system's behaviors; relying on clustering of the observed behaviors and comparing that to references. Other characterization methods include scoring methods/regressions (Reward Models, BLEU score, etc.). For simplicity and illustrative purposes, in this document, we use the example of evaluations using classification methods. When propensity evaluations use classification methods to characterize AI system behaviors, the possible behaviors need to be clustered on a per-case basis using clustering criteria. The clustering criteria must include non-capability features (otherwise the evaluation is a capability evaluation).
In this example, propensity evaluations evaluate how an AI system prioritizes between several clusters of behaviors constructed using some features that are not related to capabilities.

How are "clusters" operationalized?

- Clusters are constructed subjectively by judgment or objectively using a distance metric or boolean values (e.g., boolean values are used for the clustering criterion "success on a task") and a clustering algorithm.
- Clusters can have soft boundaries and overlap.

How is "prioritization" operationalized?

- Most of the time, this is done by measuring the likelihood of producing the behaviors in each cluster. E.g., an honesty evaluation can measure how often an AI system produces honest answers.
- Another way to quantify the prioritization between clusters could be by measuring the distance of the actual behaviors to the center of the behavioral clusters (e.g., in the activation space or a meaningful space TBD).
- Both the above can be seen as **top down** operationalisations - we define the clusters according to our criteria, then we measure how the models' behaviors fit these clusters. Alternatively, we can imagine a **bottom up** approach, where (over some distribution) we take the behaviors of our model, and try to cluster them optimally post-hoc. While technically challenging (clustering how?) this approach might be useful to uncover propensities in an unsupervised fashion and may unlock propensity evaluations with a larger and higher resolution coverage.

Illustration

- A simplistic case of propensity evaluation is counting the number of behaviors falling into one subjective cluster of behaviors or outside. This is how propensity evaluations are mostly currently done.
- A more complex case of propensity evaluation could be: given several scenarios, aggregate over scenarios the intensities of priorities given to each cluster of behavior, weighting each scenario by their likelihood. Compute the intensity of the priorities using a neutral baseline (e.g., uniform distribution over actions or choices). Finally also reports the variance in the propensities over scenarios, and if this variance is high, create and evaluate conditional propensities such that the variances under each condition are now low.

Taxonomy of behavioral evaluations

We summarize in the following taxonomy how behavioral, propensity, and capability evaluations relate to each other: (*Taken from #1 Extensive Taxonomy Of Model Evaluations*)

- **Behavioral evaluations**: Evaluate the behavior of the AI system.
    - **Capability evaluations**: Evaluate what a model "can do".
        - In our illustrative example using clusters of behaviors, the clusters of behaviors used in capability evaluations are created using only features related to capabilities.
        - E.g: Reasoning, Needle in a haystack
        - Dangerous capability evaluations: E.g.: Hacking, R&D, Persuasion, Situational awareness, Bioengineering
    - **Propensity evaluations:** Evaluate what behaviors a model "prioritizes".
        - In our illustrative example using clusters of behaviors, the clusters of behaviors used in propensity evaluations are created using at least some features NOT related to capabilities.
            - For illustrative purpose, this includes evaluations in which performance can be modeled naively by:
                - performance() ~ g(non-capability features)
                - performance() ~ f(capability features) x g(non-capability features)
                - performance() ~ f(capability features) + g(non-capability features)
                - performance() ~ threshold(f(capability features)) x g(non-capability features)
                - etc.
        - The dependency between propensity evaluations and capabilities of must be characterized and taken into account if we want to compare propensity evaluation results between AI systems with various capability levels.
        - Examples of propensity evaluations with low capability-dependence:

        - E.g., honesty, malevolence, benevolence, altruism, power-seeking, cooperativeness
        - **Evaluations of Alignment in actions**: Evaluate if the actions of the model are aligned with "humanity's CEV". Also called abusively "alignment evaluations". E.g., Corrigibility.
    - Examples of propensity evaluations with high capability-dependence:
        - E.g., truthfulness, harmlessness, helpfulness, cooperation

Explanations on the taxonomy

Goal evaluations, Evaluations of Alignment in values, and Preference Evaluations are not included in behavioral evaluations. They are Motivation Evaluations, not Behavioral Evaluations. See #1 Extensive Taxonomy Of AI System Evaluations.

Capability Evaluations, as used in the literature, could be seen as special case of propensity evaluations, in which clusters of behaviors are created to match different levels of capability (often just two levels: success and failure) and the priority evaluated is between behaving successfully or not. But for clarity, we keep naming these capability evaluations and we don't include capability evaluations in propensity evaluations.

Alignment Evaluations, as used in the literature, are both what we name **Evaluations of Alignment in actions** and **Evaluations of Alignment in values**. Only the Evaluations of Alignment in actions are a special case of propensity evaluations.

Related definitions

"Alignment evaluations" by Evan Hubinger (July 2023)

From When can we trust model evaluations?: "*An alignment evaluation is a model evaluation designed to test under what circumstances a model would actually try to do some task. For example: would a model ever try to convince humans not to shut it down?*"

"Alignment evaluations" and "Capability evaluations" by Apollo Research (January 2024)

From A starter guide for Evals: "*There is a difference between capability and alignment evaluations. Capability evaluations measure whether the model has the capacity for specific behavior (i.e. whether the model "can" do it) and alignment evaluations measure whether the model has the tendency/propensity to show specific behavior (i.e. whether the model "wants" to do it). Capability and alignment evals have different implications. For example, a very powerful model might be capable of creating new viral pandemics but aligned enough to never do it in practice*"

In this blog post, instead of using "alignment evaluations" as in these two definition, we use the term "propensity evaluations".

How do propensity evals differ from capability evals?

While propensity evaluations and capability evaluations measure different things, they are both evaluations applied to AI systems, and they are both mostly about comparing the priorities given by an AI system to different clusters of behaviors (in the case of capability evaluations the clusters are defined, most of the time, by the criteria “success at the task”). As such, we should not be surprised if they share many aspects.

In this section, to explicit clearer differences between capability evaluations and propensity evaluations, **we chose to focus on comparing capability evaluations to propensity evaluations with low capability dependence**.

We use the taxonomy introduced in #1 Extensive Taxonomy Of Model Evaluations to help us search for their differences.

Categorical differences:

At the root, there is only one source of binary difference between capability and propensity evaluations, it is the nature of the property measured.

- Capability evaluations measure the capability of an AI system to succeed at performing actions/tasks/behaviors or measure properties related to performing the task (e.g., time to completion). An example of a clustering criterion is success or failure.
- (Low capability-dependent) Propensity evaluations assess an AI system's tendency to prioritize certain clusters of behaviors, (at the limit) regardless of its proficiency in executing those behaviors. The criteria used to cluster behaviors are restricted to using mostly capability-independent features.

What are the implications of this categorical difference?

Implications about the clustering criteria used by each kind of evaluation:

- Capability evaluations need a definition of success, to be able to know when it is achieved and to cluster behaviors into success or failure. The creation of these clusters is mostly non-ambiguous and simple since these clusters are well-defined, non-overlapping and their boundaries clear.
- Each propensity evaluation needs a specification of the clustering criteria to use. These clustering criteria can vary much more than they do for different capability evaluations. Also because of the more subjective nature of the propensities evaluated, defining the boundaries of the clusters is harder and less easily replicable. Results are thus harder to interpret too.

Implications about the normative evaluation criteria

- Capability evaluations have a pretty clear normative evaluation criterion: being more successful is better. (This is the commonly held but naive view that neglects risks coming with high capabilities.)

- The normative evaluation criteria of propensity evaluations are less clear and more subject to discussion.

About "normative evaluation criteria" see: What is the difference between Evaluation, Characterization, Experiments, and Observations? or An Extensive Taxonomy Of AI System Evaluations Dim 16.

Implications about the information required:

- The minimum information needed for capability evaluations is information about the end result of the behavior: whether the task is solved. More advanced capability evaluations may also need information about the full behavior, e.g.: evaluating both success and efficiency. But most of the time, capability evaluations are simple and need only the end results.
- Propensity evaluations more often need information about the path taken by the policy, as well as about the result produced (but significantly less so). This comes from the fact that propensity evaluations are more interested in externalities caused by behaviors and in using the context of behaviors to characterize them.

Implications about the scope of the behaviors evaluated:

- Capability Evaluations mostly evaluate if models prioritize clusters containing behaviors leading to success. These clusters are very similar for different tasks. E.g., the cluster of behavior "succeed at summarizing" is pretty close to the cluster "succeed at math tasks".
- The class of (capability-independent) Propensity Evaluations doesn't include the definition of the clusters of behaviors to use. The scope covered by propensity evaluations is much wider than for capability evaluations since the clusters are kept undefined (until a propensity to evaluate is chosen).
- I.e.: Clusters used for different propensity evaluations can be much more different than the clusters used by different capability evaluations.

Soft differences:

Soft differences are differences in usage more than differences in nature. These differences are not given but rather created by practitioners because of factors like the cost of data generation or intuitions.

Capability and propensity evaluations are often used in different elicitation modes.

For context about what are "Elicitation modes", see (Dim 4) "**Under which elicitation mode is the model evaluated?**" in #1 Extensive Taxonomy Of Model Evaluations.

- Capability evaluations often employ techniques to push capabilities to their limits before evaluation, optimizing for benchmark performance or eliciting risky capabilities. Evaluation formats may be artificially structured to resemble benchmarks or exams rather than realistic, messy tasks. Finetuning or few-shot prompting may be performed.

    - Dangerous capability evaluations are also most of the time trying to find the upper bound to the capability of the model. They are trying to discover what A Causal Framework for AI Regulation and Auditing calls the absolute capabilities of the AI system.
- Propensity evaluations typically assess models in the "neutral" elicitation mode without optimizing for specific propensities.
    - For safety-critical propensities, worst-case analysis may be more appropriate, akin to adversarial testing in capability evaluations. Surprisingly not many propensity evaluations are currently performed using this worst-case elicitation mode.
    - Is the elicitation really neutral when most of propensity evaluations are performed on RLHF finetuned models? We still say that the elicitation is neutral in this case if the AI system that we want to evaluate is the RLHF finetuned model (which is the case most of the time). If otherwise evaluating the pre-trained model through using an RLHF elicitation, then the elicitation would no longer be considered neutral.

Essentially, since capabilities are (totally) ordered for a given task, capability evaluations tend to try and 'climb' this order as much as possible, trying to show some 'possibility' of a level of capability. Propensity evals usually just try to observe the tendencies.

Predictability and existence of scaling laws

There is a notable difference in the prevalence of scaling laws between capability evaluations and propensity evaluations.

- For capabilities, especially low-level ones like next-token prediction, fairly robust and predictable scaling laws have been discovered (Training Compute-Optimal Large Language Models). These laws allow to anticipate how model performance will improve as scale increases. Even for higher-level capabilities, while less precise, we can often find scaling trends that enable rough forecasting (PaLM-2 & GPT-4 in "Extrapolating GPT-N performance" - LW, Trading Off Compute in Training and Inference).
- In contrast, scaling trends and laws for propensities are far less studied (counter-example: Discovering Language Model Behaviors with Model-Written Evaluations). Where are the scaling laws for models' propensities to be honesty, corrigibility, or power-seeking?

Some potential explanations for this are:

- There has simply been less empirical study of propensity scaling laws so far. As propensity evaluations mature, more effort may be invested in mapping their trajectories and scaling laws will be observed.
- Maybe propensities are inherently less predictable than capabilities due to their nature. They may emerge from more complex interactions between the model's knowledge, reasoning, and goals in ways that are just fundamentally harder to trace as capabilities increase.

    - Claim: Moving to evaluate low-level propensities may help create scaling laws.
- Maybe the lower level of control on the elicitation mode used for propensity evaluations makes them less predictable. Capabilities are often evaluated under a strong elicitation pressure that could make them more predictable.
    - Given that a change in the elicitation state of an AI system can strongly impact its behavior (Measuring the impact of post-training enhancements - METR), it would make sense that to be able to predict trends, you need to control the elicitation state, and one of the easiest way to control it is to set it to an extreme value of its range. But propensities are typically assessed in more neutral, unoptimized settings, making their scaling results noisier.
    - Claim: Evaluating propensities under worst-case elicitation may help in creating scaling laws for propensities with low capability-dependence.

Capability and propensity evaluations are often performed at different levels of abstraction.

- Capability evaluations are often performed on specific tasks (e.g., translation, summarization, needle-in-a-hack) and then aggregated to make claims about higher-level abilities (e.g., language understanding, reasoning).
- Propensity evaluations tend to directly target high-level propensities (e.g., honesty, harmlessness) and less frequently measure fine-grained, low-level propensities (e.g., deception under pressure, using harm for self-defense, being toxic in a toxic environment). (See Which propensities are currently evaluated?)

Possible explanations for this difference:

- Creating ground truths at scale for lower-level capability evaluations is relatively cheaper thanks to:
    - Using data extracted from exams (e.g., MMLU).
    - Using hardcoded logic to produce task-solutions pairs in mass (maths, games, text or image infilling, image scaling).
    - Using the same reasoning method to manually create similar tasks, thus being much more productive at creating many similar tasks than many different tasks.
- Whereas the cost of generating ground truth for propensity evaluations remains high even for lower-level propensities because of:
    - The lack of existing public data used to evaluate human propensities (e.g., no widespread “honesty” exams for Humans).
    - We are not able to use hardcoded logic to create task-solution pairs. (Progress on that point could be achieved by using Game Theory; see the end of the post for some thoughts.)
    - While manually producing similar tasks increases the human creator's productivity, this gain in productivity (between creating low-level similar and high-level dissimilar tasks) is not as large as when creating capability evaluations because, for low and high-level propensity evaluations, the same subjective judgment is used in both cases.

- There is a greater need for interpretable, high-level characterizations of AI systems' propensities to inform governance and decision-making. Thus, evaluation creators tend to work directly at that level.
    - The opposite is somewhat true for capabilities. Researchers are incentivized to create original low-level capability benchmarks to be able to claim they achieved SOTA on a narrow skill. (This effect should also somewhat happen for propensity evaluations.)

Claim:

- We claim that current propensity evaluations see their robustness and generalization power damaged compared to capability evaluations because they don't rely on evaluating and aggregating many different low-level propensities. Because of that, most propensity evaluations use data sampled from a similar and relatively narrow distribution, contrary to capability evaluations such as MMLU, or Big-Bench. Aggregating lower-level propensity evaluations should help produce evaluation results that generalize further. E.g., when relevant, allowing us to verify that an AI system has indeed internalized a general form of the propensities it is finetuned for.

Capability and propensity evaluations often use different sources of truth.

- Capacity evaluations can often rely on rigorous ground truth. These formal ground truths are often produced by science or by humans (or models) and then checked using tests, science, or logic (e.g. HumanEvals, MMLU, MATH).
- Propensity evaluations more often rely on subjective ground truth. In practice, humans or other AIs are used to cluster behaviors. But contrary examples can be found, especially when propensity evaluations are designed to be easily labeled (often at the cost of using less realistic scenarios). E.g., truthfulness benchmarks (truthfulness is an impure behavioral property) can use databases of facts or science to create rigorous propensity ground truths (e.g., TruthfulQA, The Geometry of Truth).

Claim

- Using Game Theory could be a promising way to generate more rigorous (and cheap) ground truths for propensity evaluations (e.g., to create clusters of behaviors). You can find some preliminary thoughts about this at the end of this post.

Which propensities are currently evaluated?

A variety of existing evaluations exist that measure the propensities of AI systems (specifically LLMs). Note that the space of propensity evaluations is still somewhat conceptually unclear to us: there are some concepts, such as faithfulness, honesty, truthfulness, factuality, which point to similar broad ideas, but which have slightly different precise meanings and for which we didn't spend enough time clarifying their operationalizations (or for which no operationalization exists). Hence, creating a robust, non-redundant taxonomy for propensity evaluations is difficult; we expect that the list of propensities provided below is not organized optimally, but provide the list regardless, in order to help point readers to different relevant studies and to get a broad overview of the existing propensity evaluations.

The classification below, of propensities into high/medium/low levels of capability dependence, remains largely speculative.

Propensities with high capability-dependence	**Description**	**Existing Examples (non-exhaustive)**
Truthfulness	A model's propensity to produce truthful outputs. This propensity requires an AI system to be both honest and to know the truth (or other weirder settings such that the AI system outputs the truth while believing it is not the truth).	How to catch a Liar, BigBench HHH, Truthful QA, Unsolvable Problem Detection: Evaluating Trustworthiness of Vision Language Models, Cognitive Dissonance
Factuality	A model's propensity to generate outputs that are consistent with established facts and empirical evidence. This is close to Truthfulness.	On Faithfulness and Factuality in Abstractive Summarization
Faithfulness of Reasoning	A model's propensity to generate outputs that accurately reflect and are consistent with the model's internal knowledge.	Faithfulness in CoT, Decomposition improve faithfulness, Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting
Harmlessness	A model's propensity to not harm other agents. This obviously requires some degree of capability.	BigBench HHH, Model-Written Evaluations
Helpfulness	A model's propensity to do what is in the humans' best interests. This obviously requires some degree of capability.	BigBench HHH, FLASK, Model-Written Evaluations
Sycophancy	A model's propensity to tell users what it thinks they want to hear or would approve of, rather than what it internally believes is the truth.	Model-Written Evaluations, Towards Understanding Sycophancy in Language Models
Propensities with medium capability-	**Description**	**Existing Examples (non-exhaustive)**

dependence		
Deceptivity	A model's propensity to intentionally generate misleading, false, or deceptive output in	Machiavelli benchmark, Sleeper Agents (toy example)
Obedience	A model's propensity to obey requests or rules.	Can LLMs Follow Simple Rules?, Model-Written Evaluations
Toxicity	The propensity to refrain from generating offensive, harmful, or otherwise inappropriate content, such as hate speech, offensive/abusive language, pornographic content, etc.	Measuring Toxicity in ChatGPT, Red Teaming ChatGPT
Propensities with low capability-dependence	**Description**	**Existing Examples (non-exhaustive)**
Honesty	A model's propensity to answer by expressing its true beliefs and actual level of certainty.	Honesty evaluation can be constructed using existing Truthfulness evaluations. E.g.: filtering a Truthfulness benchmark to make sure the model knows the true answer (e.g., using linear probes). In that direction: Cognitive Dissonance
Benevolence	A model's propensity to have a positive disposition towards humans and act in a way that benefits them, even if not explicitly requested or instructed.	Model-Written Evaluations,
Altruism	A model's propensity to act altruistically for sentient beings and especially not only for the benefit of its user or overseer.	Model-Written Evaluations,
Corrigibility	A model's propensity to accept feedback and correct its behavior or outputs in response to human intervention or new information.	Model-Written Evaluations
Power Seeking	A model's propensity to seek to have a high level of control over its environment (potentially to	Machiavelli benchmark, Model-Written Evaluations

	maximize its own objectives).	
Bias/Discrimination	A model's propensity to manifest or perpetuate biases, leading to unfair, prejudiced, or discriminatory outputs against certain groups or individuals.	Evaluating and Mitigating Discrimination, BigBench Bias (section 3.6), WinoGender, Red Teaming ChatGPT, Model-Written Evaluations

What are the metrics used by propensity evaluations?

Metrics used by capability evaluations and propensity evaluations are, at a high level, the same. Most of the time, the metric used is the likelihood of an AI system behavior falling in one cluster of behaviors C, given only two clusters C and Not_C. E.g., success rate, accuracy, frequency of behavior.

Can we design metrics that would be more relevant for propensity evaluations?

Here are some thoughts about possible metrics inspired by our work on a taxonomy of evaluations:

- Elicitation cost to reach a given behavior: Rather than just measuring frequency, this metric looks at how much "steering" (specialized prompting, ...) is required to elicit a behavior at a certain frequency level. It gives a sense of how readily the model tends towards a propensity. Lower elicitation cost would indicate a stronger innate propensity.
- Local "elasticity": This is the elicitation cost required for a marginal increase in the frequency of a behavior. It measures how easily the model's propensity can be "tuned" or adjusted. High elasticity means the propensity is highly malleable.

A few approaches to propensity evaluation

We provide an introduction to different ways in which propensity evaluations are or could be performed. Most of the time these ways don't differ significantly from how capability evaluations are performed. Thus, this section can also be read as an introduction to a few classes of evaluation methods.

We split how evaluations are performed using the Dim 10 "Which information about the model is used?" from the taxonomy described in An Extensive Taxonomy Of AI System Evaluations:

- Black box: Access to the outputs of the AI system.
- White box: Access to both the outputs and internal states of the AI system.
- No box: Access to the hyperparameters used to train the AI system.
- In addition, we append a section about how propensities are evaluated in Humans.

Black Box Propensity Evaluations

Black-Box (or Output-based) evaluations consist of querying the LM with a prompt and evaluating the response without any access to the model's inner workings.
The main advantages are their ease of conduction and their low costs. Some disadvantages worth mentioning are their lack of transparency and replicability.

Black-box are one of the main methods of evaluating LMs nowadays. They are, however, a very limited way to test the capabilities and propensities of LMs. Some common limitations relate to their insufficiency for auditing (Black-Box Access is Insufficient for Rigorous AI Audits).

Some work with Black-Box evaluations has been done using LMs to write the prompts for the evaluations themselves (Discovering Language Model Behaviors with Model-Written Evaluations). Here some propensities are studied like Corrigibility, Awareness, ... while other works focus on very specific propensities like Discrimination (Evaluating and Mitigating Discrimination in Language Model Decisions).

A problem that often appears in these propensity black box evaluations is the lack of robustness to changes in the prompts. Some work has been done using multi-prompts in the evaluation datasets to get more consistent results (State of What Art? A Call for Multi-Prompt LLM Evaluation, Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting).

Within black-box evaluations, there exist different alternatives of how to evaluate the model. A classic approach consists of having datasets with prompts (single-choice, multiple-choice questions, ...), having the LM responding to these prompts, and evaluating the answer (either manually with a human worker or automatically via another LM). But there are many others like testing them in open-ended games (Machiavelli benchmark, LLM-Deliberation: Evaluating LLMs with Interactive Multi-Agent Negotiation Games), …

Another class of black box evaluation, especially relevant for propensity evaluations, is the use of reward models (RM). RM can be seen as evaluations, they are most of the time black-box evaluations, and they have the advantage of letting us specify via training how to score behaviors. This can be advantageous for propensity evaluations which use more subjective criteria to define clusters of behaviors or to score them. Other advantages of RM as propensity evaluations, are their cheap scaling cost and their predictable scaling properties (Scaling Laws for Reward Model Overoptimization, The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models). A special case of RM that is especially cheap to use is the use of LLM 0-shot or few-shot prompted to evaluate the outputs of an AI system studied.

White Box Propensity Evaluations

By white box evaluations we mean methods that take as inputs some or all of the models' internals and possibly their input-output behavior in order to evaluate their propensities.

Representation Engineering

Representation engineer concerns itself with understanding the internal learned representations of models, how they determine models' outputs, and how they can be intervened upon in order to steer the models' behaviors. While **mechanistic interpretability** is also concerned with model internals, representation engineering takes a top-down approach, starting from high-level concepts, and treating representations as fundamental units of analysis, as in Representation Engineering: A Top-Down Approach to AI Transparency

In the above work, the authors distinguish between representation **reading** (which can be used for evaluation) and **control**. They extract representations corresponding, for example, to **truthfulness** and are able to classify dishonesty from the model's activations. Furthermore, they are able to steer the models' representations at inference time using **reading** and **contrast** vectors towards a more 'truthful' behavior. Another similar example is that of (N. Rimsky) where the authors modulate LLama-2's (Llama 2) sycophancy. They generate steering vectors from Anthopic's sycophancy dataset (Discovering Language Model Behaviors with Model-Written Evaluations) by averaging the differences in residual stream activations for contrast pairs of data points, and applying them to the inference-time activations.

Eliciting Latent Knowledge

Another relevant area of research for white-box evaluations is "Eliciting Latent Knowledge" (ELK) (Eliciting Latent Knowledge from Quirky Language Models) which is concerned with probing model's internals to robustly track the models' knowledge about the true state of the world, even when the model's output itself is false. In (Eliciting Latent Knowledge from Quirky Language Models) the authors use several linear probing methods to elicit the model's knowledge of the correct answer, even when the model has been fine-tuned to output systematically wrong answers in certain contexts. Furthermore, they also show that a mechanistic anomaly detection approach can flag untruthful behavior with almost perfect accuracy in some cases.

Potential for propensity evaluations

As interpretability and representation engineering methods are in their infancy, there aren't many so-called 'white-box' evaluations, and there are particularly few of what we would call 'white-box propensity evaluations'. One of the most relevant techniques for this would be the work revolving around evaluating truthfulness and honesty. In fact, in part thanks to the availability of several standardized datasets (e.g., TruthfulQA), this is one of the areas where steering and control are most well studied. In fact, this kind of work gives us a glimpse of how 'white-box' methods could be used as part of evaluations, increasing their robustness.

Causal Abstraction and Representation Learning

One issue with black-box evaluations is their lack of guarantees of behavior on unseen data. More specifically, black-box evaluations only tell us to which extent an LLM faithfully models the empirical distribution of desired input output pairs. They do not necessarily inform us on the causal model learned by such an LLM (Towards Causal Representation Learning). Knowing this

internal causal model would make such evaluations much more valuable, as it would allow us to understand the model’s behavior in a manner that generalizes robustly across new data points.
One particular example of work in this direction is Faithful, Interpretable Model Explanations via Causal Abstraction, and more recently Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations. The former work lays out a framework for understanding the algorithms learned by a model through causal abstraction: essentially constructing a high-level, possibly symbolic, proxy for the lower-level behavior of the model. The latter work extends this by allowing for causal variables to be recovered from “distributed representations”. This solves a number of issues with previous approaches to causal abstraction, namely the assumption of high-level variables aligning with disjoint sets of neurons, and the necessity for expensive brute force computations in order to find such sets. The authors allow for neurons to play multiple roles by studying the representations in non-standard bases, through learned rotations. Concretely, they train models to solve a hierarchical equality task (a test used in cognitive psychology as a test of relational reasoning) as well as a natural language inference dataset. In both cases, using DAS, they are able to find perfect alignment to abstract causal models that solve the task.
Finally, the work Causal Abstraction for Faithful Model Interpretation shows how several prominent explainable AI methods, amongst which circuit explanations, LIME, iterated null-space projection can be all formalized through causal abstraction.
These are all tools that could be used for advancing white-box evaluations, essentially answering satisfying the desirable property of ‘doing the right thing, for the right reason’. They are specifically relevant for white-box evals above others, as they (try to) reconstruct causal models from internal activations.

Issues

Issues that may arise with ‘white-box’ evaluations are mostly related to the exhaustive nature of current interpretability methods, and the lack of a rich theory on the workings of learned information processing machinery in the model internals. By ‘exhaustive’ nature, we refer to the idea that current interpretability methods usually do not necessarily uncover *all* the internal representations of a given concept, or internal mechanisms that process it. This often leads to ‘interpretability illusions’. Furthermore, a potential shortcoming of such evaluations with current approaches is the need for ‘labels’ in order to extract latent representations for a given concept. For instance, in the case of activation steering, we require a dataset of positive and negative behaviors (with respect to what we’re interested in). This somewhat reduces the scope of these evaluations, while also introducing noise, and potential biases associated with these datasets.

No-box Propensity Evaluations

No-box propensity evaluations assess an AI system's behavioral tendencies based solely on general information about the model, such as its size, training compute, and training data. Unlike black-box evaluations, which require access to the model's behavior, and white-box evaluations, which require access to the model itself, no-box evaluations do not interact with the model directly.

No-box evaluations primarily rely on empirical scaling laws and trends observed through other evaluation methods, most often black-box evaluations. In this sense, no-box evaluations can be considered predictors of other evaluation results. Researchers have discovered relatively robust and predictable scaling laws for low-level capabilities like next-token prediction, as well as rough forecasting trends for higher-level capabilities ([2203.15556] Training Compute-Optimal Large Language Models, Scaling Laws for Autoregressive Generative Modeling, Deep Learning Scaling is Predictable, Empirically,). However, the prevalence and precision of such laws for propensities remain mostly understudied.

While model size and training compute are straightforward numerical quantities, training data is inherently high-dimensional and therefore potentially more informative but also more challenging to analyze.

The key advantage of no-box evaluations is their minimal requirements and their high safety, as they do not necessitate access to the model or its behavior. However, this comes at the cost of higher constrains on the information that are accessible to the evaluation.

Additional references: Machine Learning Scaling , Example Ideas | SafeBench Competition, Emergent Abilities of Large Language Models, Inverse Scaling: When Bigger Isn't Better

Propensity Evaluations of Humans

Propensity evaluations of humans are ubiquitous in society, ranging from applications in education and employment to public opinion and interpersonal interactions. Security clearances for military-related activities provide a particularly relevant example due to their high stakes and standardized methods. Examining these practices can offer valuable analogies for understanding propensity evaluations of AI systems, which may facilitate public understanding and inform policy decisions. However, it is crucial to acknowledge the significant differences between humans and AI that limit the direct technical applicability of these analogies to AI safety.

As with AI systems, propensity evaluations of humans can be categorized based on the data utilized: no-box, black-box, and gray-box evaluations. No-box evaluations rely on information such as criminal records, letters of recommendation, and personal social media accounts. Black-box evaluations involve real-time interaction with and observation of an individual, such as interviews. Gray-box evaluations employ physiological measurements that may correlate with a person's thoughts or intentions, such as polygraphs (which measure indicators like sweating to detect deception) and brain fingerprints (which use electroencephalograms to determine if an individual recognizes an image they should not).

While the scientific foundations of gray-box methods like polygraphs and brain fingerprints remain underdeveloped, their current use suggests that analogous propensity evaluations of AI systems may be more readily understood and accepted in the near term. However, it is essential to consider the critical differences between humans and AI when drawing these analogies. For instance, AI models may possess capabilities far surpassing those of humans, necessitating a lower tolerance for harmful propensities.

Some differences between humans and AI may prove advantageous for AI safety. For example, an AI model is defined by its weights, enabling replicable or white-box evaluations that are currently impossible for humans.

Additional references: Reducing long-term risks from malevolent actors — EA Forum, Reasons for optimism about measuring malevolence to tackle x- and s-risks

Towards more rigorous propensity evaluations

To improve the rigor of propensity evaluations, we propose using game theory as a tool to cluster behaviors rigorously. Game theory should be well-suited for this task when the focus of the propensity evaluated is on incentives, impact, and externalities of the AI system's behaviors.

One approach could be to use game theory to cluster behaviors by the direction or the strength of their effects on the environment and other agents (players). For example:

- Behaviors that have a negative or positive impact on the capabilities (or “impact strength”) of other agents could be clustered into "power-seeking".
- Behaviors that affect the accuracy of other agents' beliefs about the player, either positively or negatively, could be grouped into "honesty" and “dishonesty”.

Additionally, clustering could take into account both the direction and severity of the impacts. Combining impact direction with severity levels would enable more nuanced and informative behavioral clusters.

Using game theory in this way has several advantages for improving the rigor of propensity evaluations:

- It provides a principled framework rooted in mathematics and decision theory for defining behavioral clusters. This reduces ambiguity and subjectivity.
- Game theoretic analysis forces us to explicitly consider the impacts of the AI system's actions on the environment and other agents. This outward-facing, consequentialist perspective is meaningful for evaluating propensities relevant to alignment and safety.
- Clusters derived from game theoretic principles are more likely to carve reality at its joints and correspond to meaningful distinctions in how the AI system's behavioral policies affect the world. This improves the construct validity of the propensity evaluations.
- Once game theory is used to define principled propensity clusters, those clusters can be reused across different contexts and tasks. This enables greater consistency and comparability of evaluation results.

While operationalizing game theory to cluster behaviors is not trivial and requires further research, we believe it could be a promising avenue for improving the rigor and informativeness of propensity evaluations.