True names are precise mathematical formulations of intuitive concepts that capture all the properties that we care about for those concepts. “True names” is a term introduced by alignment researcher John Wentworth, possibly inspired by the idea from folklore that knowing a thing's “true name” grants you power over it.

Wentworth gives many examples of true names. Concepts like "force", "pressure", "charge" and "current" were all once poorly understood, based on vague intuitions about the physical world, but have now been robustly formalized mathematically.

To put it another way, a “true name” can be thought of as a mathematical formulation that robustly generalizes as intended. An important property of true names is that they are not susceptible to failing via Goodhart's law when faced with the immense optimization pressure of a future superintelligence. Since alignment researchers are interested in finding mathematical measures that are “non-Goodhartable”, they also care about finding true names. However, non-Goodhartability is just one property of true names. Robustness to optimization might be a necessary condition to conclude that we are dealing with a true name, but it isn't the definition of true names.

Many alignment researchers care about human values. It would be a huge boon for AI alignment efforts if we could discover a robust formulation or a “true name” of human values. Currently, we use proxies of what humans truly care about in AI models in order to measure how well a given model performs. The use of these proxies often results in side effects through things like reward misspecification or specification gaming. However, if we had a “true name” for human values which we could optimize for, then we would not need to worry about undesired side effects or unforeseen consequences.

In addition to human values, alignment researchers also seek true names for components of agency such as optimization, goals, world models, abstraction, counterfactuals, and embeddedness.

Related

- What is "agent foundations"?
- What are “type signatures”?
- What is the deconfusion branch of ai alignment research?
- What is John Wentworth's research agenda?
- What are "selection theorems" and can they tell us anything useful about the likely shape ...

Scratchpad

Yaakovs comments

I would suggest this alternative setup for this answer:

1. True names are a precise mathematical formulation of a certain intuitive concept, which captures all of the properties that we intuitively care about from that concept
2. An important property (and perhaps a sufficient condition?) of true names is that they are not subject to the problem of goodharting, they will remain effective even under strong optimization pressure. This property is a major reason that they are important for alignment. Another way of framing this is a "True Name" is a mathematical formulation which *robustly generalizes as intended*.
3. One set of true names that are important for alignment are the true names of human values
4. Alignment researchers are also interested in other true names, such as such as optimization, goals, world models, abstraction, counterfactuals, embeddedness (note that this is the one which is emphasised in John Wentworth's post)

Cinera's comment

I think this is an overly restrictive definition of "true name".

In particular, myself and other alignment researchers have used "true name" in a broader context than just talking about human values.

E.g. Janus referred to the true name of large language models in their "Simulators" post:
> I want to hypothesize about LLMs in the limit, [...] I could directly extrapolate the architecture responsible for these feats and talk about "GPT-N", a bigger autoregressive transformer. But often some implementation details aren't as important as the more abstract archetype that GPT represents – I want to speak the true name of the solution which unraveled a Cambrian explosion of AI phenomena with inessential details unconstrained, as we'd speak of natural selection finding the solution of the "lens" without specifying the prototype's diameter or focal length.
https://www.lesswrong.com/posts/vJFdjigzmcXMhNTsx/simulators#:~:text=I%20want%20to,or%20focal%20length.

And I have contended that "agent" is not the true name of intelligent systems produced by optimisation processes:
> To be precise, I am quite unconvinced that "agent" is the "true name" of the relevant intelligent systems. There are powerful artifacts (e.g. the base versions of large language models) that do not

match the agent archetype as traditionally conceived. I do not know that the artifacts that ultimately matter would necessarily conform to the agent archetype[7]. Theorems that are exclusively about the properties of agents may end up not being very applicable to important systems of interest (if e.g. the first AGIs are created by a [mostly] self-supervised training process).

https://www.lesswrong.com/posts/G2Lne2Fi7Qra5Lbuf/selection-theorems-a-program-for-understanding-agents?commentId=ZFQ6s6oeapRx95pcC

I think that by restricting yourself to "values", you're failing to capture "true name" as it is actually used by alignment researchers in practice.

Removed things

- unnecessary

It is the goal of some alignment researchers to discover either these ‘true names’ for human values or a “pointer” to human values – something from which the “True Name” of human values could be automatically generated. They hope that if we find either the “true names” by studying the fields listed above, or if we can find a generator function for these “true names” it will be a significant step towards solving the alignment problem.

- ...

As part of the comment thread underneath the post in which ‘true names’ were introduced, the author says that “*robustness to optimization is not the True Name of True Names, but it might be a sufficient condition*” for achieving/finding “true names”. So it might be generally acceptable to understand them in this way.

- Overly restrictive definition

Currently, we use proxies for human values to measure how well an AI model performs. However, these proxies often fail when faced with optimization pressure. A “true name” in the alignment context is one or more human values that do not fail under optimization pressure. This means that if we try to maximize the attainment of those values it does not result in undesired side effects or unforeseen consequences.

- Confusing Example

As an example, we use the proxy of aesthetics to measure how good food tastes. So food that looks good quite often also tastes good. However, when the proxy becomes the target, we might over-optimize for making food 'look good' vs. actually 'be good'. After a certain point making the food look good becomes more important than actually tasting good which results in overall worse-tasting food. The "true name" of food would be a dish that does not become nutritionally worse or taste worse when we optimize for the appearance of that same dish. One common example observed in everyday life is fruits and vegetables optimized for supermarkets. These foods are optimized for appearance but are often less nutritionally valuable[1] than foods that are not in the supermarket, but are directly from farms or gardens. These fruits and vegetables might be duller in color and oddly shaped but tend to have higher nutritional values and better taste.

[1] Optimization pressure for appearance over nutrition is a factor, but this is just meant to serve as an illustrative example of optimization pressures and is not meant as an authoritative claim from a botanist/biologist/...The main culprit in certain fruits and vegetables having lesser nutritional value is depletion of the top soil layers from overfarming, resulting in less nutritional dense foods overall.