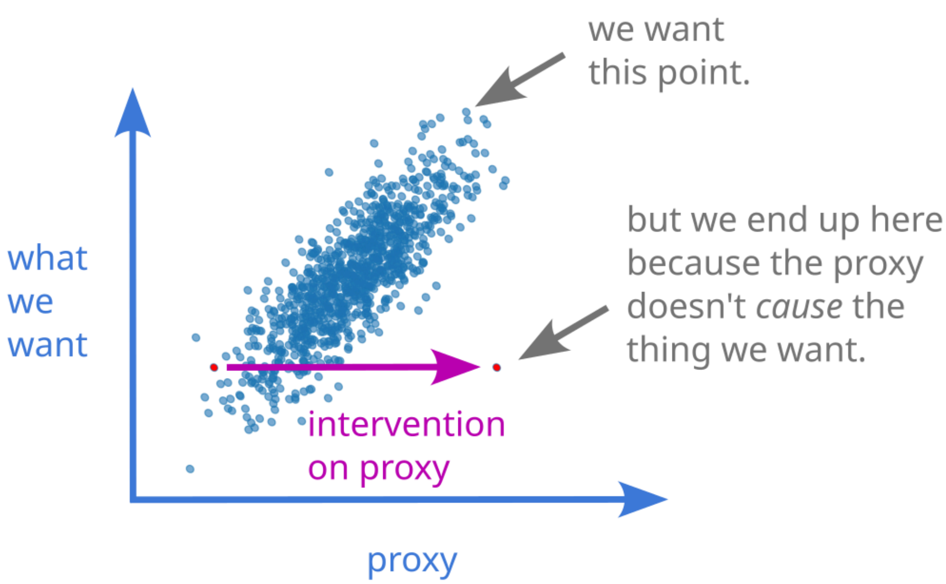
Imagine a small child who learns that height is correlated with basketball skills. Excited, the child starts practicing basketball in order to become taller.

This is an instance of **Causal Goodhart**: The child optimizes a [proxy](), but that proxy does not actually *cause* the thing they want.



Causal Goodhart is important for AI alignment because we may end up building AIs which act as optimizers. For example, if the optimization target is implemented as a variable on a computer (the proxy), then the AI might just directly set the variable to a high value instead of optimizing for what we intended it to optimize for. This failure mode is called wireheading.

Causal Goodhart is a form of the more general phenomenon of Goodhart's law, which states that
> As soon as a measure becomes a target, it ceases to be a good measure.

For Causal Goodhart, the proxy ceases to be a good measure because the proxy does not *cause* the thing we want, and thus intervening on the proxy eliminates the correlation. However, there are other ways in which the proxy starts becoming a bad measure when it is optimized: Other types of Goodhart are Extremal Goodhart, Regressional Goodhart and Adversarial Goodhart.

You can read more in the Goodhart Taxonomy by Scott Garrabrant, which introduces these four types of Goodhart.

Alternative phrasings

-

Related

- What is Goodhart's law?
- What is Extremal Goodhart?
- What is Regressional Goodhart?
- What is Adversarial Goodhart?