Summary: Goal misgeneralization primarily happens because the system under training latches not upon the goal it's being trained for, but upon an *upstream correlate* of that goal — like human love for their children is an upstream correlate of inclusive genetic fitness.†

What complicates the problem are suspicions of path-dependence°. It's not overdetermined what values a given system subjected to a given selection pressure will learn. Rather, every next value learned is a function of the values the system has *already* learned, such that the entire process can only be predicted step-by-step, no "skipping to the end" allowed.

I propose that it may nonetheless be *predictable*, and provide a first-approximation model of how that may look like.

> † Current discourse often discourages this characterization, pointing out that evolution's effect upon human minds is screened off by human brains' reward circuitry. I think it's fine, though: that hard-coded reward circuitry is itself an upstream correlate of inclusive genetic fitness, and upstream-correlatedness is a transitive property.

The Setup

We'll model the environment E as some complex probabilistic causal model consisting of nodes/variables $\{e_1,\ e_2,\ \dots,\ e_n\}$. We'll denote the joint probability distribution over the environment as $P(E) = P(e_1 \cdot \dots \cdot e_n)$.

$S : \{e_i\} \to \mathbb{R}$ is the selection pressure, a function which takes in the values of some pre-specified set of variables, and returns a score. Inclusive genetic fitness for animal populations, reward circuitry for human brains, loss functions for ML models. We'll denote the variables S is defined over as g_i^0, and the set of all of them G^0.

The environment contains an agent. The set of variables $\Pi \subset E$ is the agent's policy function, and $A \subset E$ are the agent's action-variables.

To start off, let's define what an "action" is. In essence, those are variables the agent can freely control — variables which are downstream of Π, and whose values are arbitrarily set by Π.

To formalize this, we first introduce the following function, which quantifies the extent of causal influence of X upon Y:

$$I^c_{\rightarrow}(X;Y) := 1(X \rightarrow Y) \cdot \left(1 - \frac{H(Y|X)}{H(Y)}\right) \in [0;1]$$

Where $H(\cdot)$ measures entropy, and $1(X \rightarrow Y)$ represents a function that returns 1 if X is upstream of Y, and 0 otherwise. $I^c_{\rightarrow}(X;Y)$ itself, then, returns the fraction of uncertainty in Y that is banished by X — i. e., "how much" of Y is "controlled" by X.

From there, defining actions is straightforward: $A = \{e_i \mid I^c_{\rightarrow}(\Pi;\ e_i) \approx 1\}$. Actions are variables whose values are entirely or almost-entirely controlled by the agent's planning process.

In a similar fashion, we can concretize *upstream correlates of goals*. They're variables or sets of variables such that they are, to wit, causally upstream of the goal-variable, and whose values largely determine the goal-variable's value.

Formally: let $\text{corr}(g^m_i)$ be a function which returns a set of variables $\{e_k\}$ such that, for all $m > 0$, it satisfies the following condition:

$$(I^c_{\rightarrow}(\text{corr}(g^m_i); g^m_i) \approx 1) \wedge (\forall g \in \text{corr}(g^m_i) : I^c_{\rightarrow}((\text{corr}(g^m_i) \setminus g); g^m_i) < 1)\}$$

The second clause here ensures that the set is non-redundant: you can't take out any variable out of it without it losing control over g^m_i.

The individual correlates shall be denoted $g^{m+1}_k \in \text{corr}(g^m_i)$.

In a similar fashion, we define a saturated set of mth-order correlates G^m, for $m > 0$:

$$G^m := \{g^m_i\} \mid \exists G^{m-1} : (I^c_{\rightarrow}(G^m; G^{m-1}) \approx 1) \wedge (\forall g \in G^m : I^c_{\rightarrow}((G^m \setminus g); G^{m-1}) < 1)$$

Starting from G^0, G^1 is some set of variables that can jointly control all variables in G^0 — and there may be several such sets. G^2, then is some set of variables that can robustly control one of the possible G^1s, and so on.

Further, we'll posit that the actions in A are some αth-order correlates of G^0, i. e. $A \subset G^\alpha$. That assumption is basically free: all the e_is are part of the same causal model, and if we're assuming our agent can conform to the selection pressure/achieve its goals, its actions *have* to be upstream of the goal-variables.

As the final building block, we define a *heuristic*:

$$h_i^m(t) := P(\text{corr}(g_i^m) \mid (g_i^m \sim t))$$

A heuristic for g_i^m takes in some target distribution t_i^m, and returns a probability distribution over the values of $\text{corr}(g_i^m)$ such that, if they held these values, g_i^m would be distributed as t_i^m.

All of this allows us to define the following planning algorithm (with some algorithmic glue omitted):

$$\text{cascade}(T^m) := \begin{cases} T^m & \mid m = \alpha \\ \bigcup\limits_{t_i^m \in T^m} \text{cascade}(h_i^m(t_i^m)) & \mid m < \alpha \end{cases}$$

Where $T^m = \{t_i^m\}$, i. e. some set of targets for mth-order correlates.

In essence, what $\text{cascade}(\cdot)$ does is propagating the condition "the goal is achieved" down through goal-correlates, until it computes the values to which the action-variables themselves must be set — i. e., what actions the agent must take to achieve its goal.

So in the idealized case, with a magical training loop that perfectly aligns the agent to the outer objective, we'd get an agent that orients itself to the current state of the environment (updates its world-model to the current training episode), then runs $\text{cascade}(T^0)$ over it, where T^0 is hard-wired and contains some values for G^0 that are highly reinforced by S (values such that S returns high score for setting g_i^0 to them).

Goal Misgeneralization

If we translate the Gooder Regulator Theorem° to this context, it would say something like the following: any sufficiently capable planning process Π with an action-set G^α exhaustively subjected to the selection pressure S defined over G^0 would end up developing a world-model representing $P(G^\alpha \cdot \ldots \cdot G^0)$. (Well, I think the exact claim is a bit weaker: the planner's WM would end up representing only those g_i^m whose values *actually changed during the training*. But that subtle distinction will be rendered moot by the next point.)

In practice, however, agents aren't usually trained to be "sufficiently capable" in this sense; neither ML models nor genomes nor humans. Instead, the training stops early, and the agent ends up with an *incomplete* suite of heuristics: one corresponding to $P(G_s^\alpha \cdot \ldots \cdot G_s^x)$, where $x > 0$ and $G_s^m \subset G^m$.

That's what causes inner-alignment problems. Agents end up with an unfinished "bridge" from actions to the selection pressure, with godshatter°/proto-values/shards that are weird upstream correlates of the selection pressure. Instead of passing T^0 to cascade$(\cdot)$, they pass some $T^? = \{t_i^m,\ t_k^q,\ \ldots,\ t_j^n\}$, and what goes into $T^?$ may even change from one episode to another.

But that process may not be entirely beyond reckoning.

Agents subjected to a selection pressure learn the world-model and heuristics gradually, variable-by-variable. Heuristics for controlling the value of g_i^m only become available once the agent has heuristics for $\mathrm{corr}(g_i^m)$, and once its world-model represents $P(g_i^m \cdot \mathrm{corr}(g_i^m))$. In turn, in what "direction" an agent starts building the "heuristical bridge" isn't random either: it's some function of how hard given variables are to learn and to master relative to how much payoff controlling them can give (i. e., how strongly learning to control them is reinforced).

Simplifying, suppose we have a planner Π in an environment E subjected to the selection pressure S, and it already developed a world-model representing $P(G_s^\alpha \cdot \ldots \cdot G_s^x)$. What *next* heuristic it will learn? It should be possible to predict! For example, it may be the heuristic controlling whatever g, among those available to be learned, scores highest on the following metric:

$$\frac{I(g;\Pi) \cdot I^c_{\rightarrow}(g;G^0)}{\text{complexity}(P(\text{corr}(g)) \rightarrow P(g))}$$

Where $I(g;\Pi)$ measures how "visible" g is to the agent (how hard it is to affect *or* be affected by), $I^c_{\rightarrow}(g;G^0)$ measures g's importance to S, and the denominator measures the complexity of the function mapping $\text{corr}(g)$ to g (which is effectively how hard it is to learn given the things the planner has already learned).

In theory, *the whole* training process may be predictable in this manner. Figuring out the rule for predicting the next heuristic, then applying it iteratively to some NN we just initialized, should allow us to predict the NN we'll get at the very end°.

That, in turn, may be very useful for alignment. It'd be the "holy grail" of solutions to goal misgeneralization in heuristics-based planners (which include shard ecosystems°), allowing us to figure out precisely how to set up our training loops to cultivate the values we want.

Further work

My personal opinion, however, is that it isn't a very hopeful research direction. I've generated this framework as part of a more broad model of embedded agency, and it does seem exciting... But there's two major problems:

- The exact paths through parameter-space a model takes may be highly unstable, and sensitive to perturbations. Consider the Lottery Ticket Hypothesis°. If correct, it suggests that the initial set of heuristics learned would be heavily influenced by random noise. Similar may happen at other points in training: the specifics of inductive biases, neural architecture, or the order of training episodes the AI is put through may play key parts in the direction in which the "heuristical bridge" is being built. The process may be inherently unpredictable — so no predictive model would do very well.
- Even if we can develop a robust predictive theory of heuristics formation, I don't expect it to be a definitive solution to inner alignment/goal misgeneralization *in*

full. It'll suffice for heuristics-based planners, but our real concern is AGI. And while an AGI *would* take its shards/heuristics/proto-values as an input when deciding upon its values, much like base human urges don't equal outputs of our moral philosophy, the values the AGI will decide upon won't map 1-to-1 to its shards. And the process of reflecting upon godshatter° may itself be highly unstable°.
    - (On my view, retargeting the search° is the only solution currently proposed that's worth trying for.)

Still, it *may* be fruitful, perhaps as part of a more complete solution to inner alignment, and I'll definitely be developing this framework more.

Lastly, I'll note that a more detailed version of it may be relatively easy to empirically verify:

- Set up a giant causal model, such that we know all transition functions and all conditional probabilities in it.
- Train a ML model on it.
- Use an improved version of this framework to try and predict which heuristics the model must have learned.
- Probe the ML model to see if our predictions seem to check out. (We already have° interpretability techniques that should suffice for getting a rough confirmation.)