Formalizing the Question of a Sharp Left Turn

David Reber

April 18, 2023

Rough Sketch

Let's assume that human concepts are crisp in the sense that they are amenable to being assigned mathematical variables

Assumption 1. ***Finite Concept Space*** *The set of relevant human concepts is finite.*

Perhaps there are situations where an infinite-concept space would be more appropriate, but I suspect that sticking with finite won't affect this investigation too much, and finite seems already plently large enough to capture much of what us huans care about.

Assumption 2. ***Hard Concepts*** *Human concepts can be written down and formalized, even if they are intrinsically soft or fuzzy.*

Definition 1. ***Goal*** *An indicator function over concept space that determines which states are okay and which aren't.*

Note: this notion of goal is extremely flexible; not all are well-defined or feasible, and some are separate from concept-states.

Definition 2. ***Sensical*** *A concept-tuple is considered sensical if it respects reality's relationships/constraints between concepts.*

Definition 3. ***Constraint*** *Constraints are relationships between concepts, such as "a woman is a person" as well as temporal and logical relationships.*

Definition 4. ***Convergent Instrumental Subgoal*** *A subgoal that is instrumental for multiple goals.*

Definition 5. ***Sensical Concept-Tuples***
Call the entire set of sensical concept-tuples Ω.

Assumption 3. Ω ***is a*** σ***-Algebra*** *Ω is a σ-algebra.*

Definition 6. ***Hard Goal*** *We call a goal g* ***hard*** *if $P(g)$ is very small.*

Note that we can get a goal naturally from utility functions by asking where those functions are optimized (or more generally, asking for "above a threshold"). Hence multiple utility functions may correspond to a single goal: indeed, each goal can have an infinite number of utility functions, but those are all equivalent in some sense.

Now, suppose we have a goal g^*.

Definition 7. ***Subgoal*** *A subgoal is any goal $g_i \neq g^*$ such that $g_i \neq \Omega$, and is used as a proxy for g^*.*

Definition 8. ***Instrumental*** *A subgoal g_i is "instrumental" for g^* if $P(g^* = 1|g_i = 1) > P(g^* = 1)$; that is, achieving $g_i = 1$ makes $g^* = 1$ more likely.*

For intuition, it can be useful to consider the special case where $g^* = \bigcap_i g_i$, although this excludes subgoals which throw away some decent $g^* = 1$ options in order to make $g^* = 1$ more likely to be hit overall. Note by definition, we can't have trivial subgoals like $g_i = \Omega$.

Definition 9. ***Convergent*** *Let $G_1, \ldots, G_n$ be goals. We say a goal $\hat{g}$ is a convergent instrumental subgoal for $g_1, \ldots, g_n$ if it is instrumental for each; that is, $P(g_i = 1|\hat{g} = 1) > P(g_i = 1)$ for all i.*

Aside: if $\hat{g}$ is human-understandable, it may be a concept. Hence, this phenomenon (if it indeed exists) would correspond to a lot of goals having high concentration on that half.

If convergent instrumental subgoals (CISs) exist, AGI by nature is likely to pick up on them (because it's meant to be general).

1 Generalization

What does generalization mean in this context?

Example 1.

- *AI Alice is supposed to do alignment research the same way the alignment research Alice would, starting within some overall goal g_0 that Alice gives.*
- *AI-Alice analyzes thousands of neurons using mechanistic methods. At this point, AI-Alice needs to synthesize the results to get new conjectures and hypotheses to form new research questions to explore, but this is something Alice has only ever done on at most 100 neurons at a time.*

Hence this amounts to changing the research question and goals, in an unfamiliar domain: mathematically we represent this as $g_0 \to g_1$. Maybe it's a small change, maybe quite distinctive, but at any rate if $\hat{g}$ is a CIS, then likely $P(g_1 = 1|\hat{g} = 1) > P(g_1 = 1)$ so capabilities generalize. Meanwhile, $P(g_0 = 1|g_1 = 1) = 0$ if $g_0 \cap g_1 = \emptyset$, so if g_1 represents a sufficiently large deviation from g_0, the original objective will not be satisfied.

Now, let h be the goal of "avoids really obvious bad stuff". If h is hard (aka value is fragile) it's not robust to changes in the goal. If h is large relative to

the coherent states ω, it's very generalizable. This change of goal is probably best illustrated intuitively with stochastic gradient descent. Goals (as defined here) change on each batch, but empirically the sequence of them still leads you where you want.

2 Are CISs likely to be learned in general?

Assume $\hat{g}$ is a convergent instrumental subgoal for some $\{g_i\}$, but not the goals $\{g_j\}$. Given some random process for selecting one of $\{g_i\} \cup \{g_j\}$, how likely is $\hat{g}$ to be satisfied? That is, we have fixed $P(G_k = 1)$ for $k \in \{i\} \cup \{j\}$, and want to find $P(\hat{g} = 1)$.

$$P(\hat{g} = 1|g_k = 1) = \sum_i P(\hat{g} = 1|g_i = 1)P(g_i = 1) + \sum_j P(\hat{g} = 1|g_j = 1)P(g_j = 1)$$

For simplicity, assume $P(\hat{g} = 1|g_i = 1) \approx 1$, and $P(\hat{g} = 1|g_j = 1) \approx 0$. Then $P(\hat{g} = 1|g_k = 1) \approx \sum_i P(g_i = 1)$. That is, the likelihood of the convergent instrumental subgoal is going to be roughly the probability that any g_i is picked. But this conclusion rests heavily on the $P(\hat{g} = 1|g_i = 1) \approx 1$ assumption.

3 How likely is a CIS to be learned for a particular g^*?

Suppose $\hat{g}$ is a CIS for a goal g^*. That is, $P(g^* = 1|\hat{g} = 1) > P(g^* = 1)$. If we assume an extremely naive learner which samples x from Ω *uniformly* to see if $x \in g^*$, then switching to sampling $x \in \hat{g}$ uniformly will represent a speedup of "operations til goal-satisfaction".

Conjecture 1. *Why wouldn't this be the case numerically?*

Algorithm 1: sample x from Ω uniformly until $g^*(x) = 1$, then terminate.

Algorithm 2: sample x from $\hat{g}$ uniformly until $g^*(x) = 1$, then terminate.

According to GPT-4, the number of iterations for Algorithms 1 and 2 follow geometric distributions [TODO: verify] so the expected number of iterations are $\mathbb{E}_1 = 1/(P(g^* = 1))$ and $\mathbb{E}_2 = 1/P(g^* = 1|\hat{g} = 1)$.

So the speedup given by the ratio of $\mathbb{E}_1/\mathbb{E}_2 = P(g^* = 1|\hat{g} = 1)/P(g^* = 1)$. Since the numerator is bounded, this ratio seems largely driven by how small $P(g^* = 1)$ might be.

Key takeaway: if your goal g^* is hard, then the assumption that if a CIS exists it will be used, seems more reasonable than not.

4 Takeaways

Now, back to our generalization question. If g_0 is the original goal, and it is hard, then any instrumental subgoal $\hat{g}$ is likely to be learned, **and** it's more

likely that changes to the goal will cause the original goal to be unfulfilled. If $\hat{g}$ is widely convergent, this increases the likelihood $\hat{g}$ will be learned further.

Cruxes:

- that any convergent instrumental subgoal even exists
- that g_0 (specifically h) is a hard goal.

Can this be verified experimentally?

Questions

- Do convergent intstrumental subgoals (CISs) actually exist?
- Are CISs expected to be frequent, theoretically?
- What experimental setup could discover whether CISs exist, and how frequent they are?

Given some random process, how likely is the subgoal to be satisfied?

The likelihood of the convergent instrumental subgoal being satisfied depends on the probability that any given goal is picked. However, this conclusion rests heavily on the assumption that $P(\hat{g} = 1|g = 1)$ is close to 1.

- How likely is a CIS to be learned for a particular goal?
- Suppose $\hat{g}$ is a CIS for a goal g^*. If we assume an extremely naive learner that samples uniformly to see if a concept satisfies g^*, then switches to sampling from $\hat{g}$ uniformly, this represents a speedup in "operations until goal satisfaction."

The speedup depends on the ratio of the expected number of iterations for the two algorithms. If the goal g^* is hard, then assuming a CIS exists, it will be used more frequently, making the original goal more likely to go unfulfilled.

- Can this be verified experimentally?