Formalizing the Question of a Sharp Left Turn

David Reber

April 18, 2023

1 Motivation

There's been quite a bit of disagreement and confusion around Nate Soares' 'Sharp Left Turn' hypothesis, which he posits is the 'hard part' of alignment, and which would render using AGI even on seemingly positive objectives (e.g. as an automated alignment researcher) inherently dangerous.

I think we can do better at resolving this disagreement. The goal of this post is to articulate the "sharp left turn" hypothesis formally, and articulate experiments which can either verify or falsify it. To ensure that the formalism is not unduly favoring one perspective, I have iterated with Nate Soares, [someone], and [someone]. Each broadly endorses this document: specific comments are marked in brackets with the individual's initials (NS, and ?? respectively).

2 Formalism

All frameworks are wrong, but some are useful. This framework is the map, not the territory. My hope is that by proposing a decent framework, and demonstrating how to derive testable implications of that framework, others will have an easier time exploring implications of (realistic!) variations of the framework. If the majority of the testable implications of realistic framework-variations are empirically leaning one direction, this constitutes overall strong evidence for that conclusion. In the worst case that the empirical results end up being highly sensitive to the differences of multiple realistic frameworks, then at least we'll have succeeded in articulating how our assumptions affect our interpretations.

Here, by "realistic" assumption I'm gesturing at "an assumption which is likely to be valid in the situations we expect to end up happening". I expect, and encourage, disagreement about which assumptions are "realistic".

2.1 What is a Goal?

Let's assume that human concepts are crisp in the sense that they are amenable to being assigned mathematical variables This shouldn't be a problem, because it seems like human concepts can be written down and formalized, even if they are intrinsically soft or fuzzy.

Definition 1. ***Concept Space*** *A concept consists of a set of states. If concept X takes on value x, we write $X = x$. The concept space is the cartesian product of all concepts, with elements of the form $(x_1, x_2 \dots)$*

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. Consequently, we'll refer to goals instead of utility/loss functions.

Definition 2. ***Goal*** *An indicator function over concept space that determines which states are satisfactory, and which are not.*

Now we also want to restrict our attention to concept-tuples which are sensical. This is because we want to avoid the problem of having a goal that is impossible to achieve, or that is so unlikely to be achieved that it's not worth pursuing. So, we'll define a set Ω which is the set of all concept-tuples which make sense:

- Equality relations (synonymns, concerved quantites in physics, etc)
- heirarcial relationships like "Socrates is a man" and "a man is a human", plus the transitive closure of such relationships

With a lot of work, we could probably come up with a formal definition of sensical concept-tuples, but for now we'll just assume that we can do it.

Definition 3. ***Sensical Concept-Tuples*** *Call the entire set of sensical concept-tuples Ω.*

Note that Ω is a subset of the set of all concept-tuples, which is the set of all possible worlds.

Now really, there's no reason to believe that Ω is a σ-algebra, but it's a good assumption to make for now: otherwise, it get hard to talk about probabilities. We can always relax it later.

Assumption 1. *Ω is a σ-algebra.*

For intution, just consider $\Omega = \mathcal{R}^n$ (though the results here don't rely on this). This is a σ-algebra, and it makes sense for modeling 'nice' concepts that are amenable to real-values.

Definition 4. ***Hard Goal (Informal)*** *We call a goal g* ***hard*** *if g is very restrictive; that is, $P(g)$ is very small.*

2.2 What is a Convergent Instrumental Subgoal?

Definition 5. ***Subgoal*** *For a given goal g^*, a subgoal is any goal $g_i \neq g^*$ used as a proxy for g^*. Unless stated otherwise, it will be assumed that subgoals are nontrivial: $g_i \neq \emptyset$ and $g_i \neq \Omega$.*

Definition 6. ***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 7. ***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 itself be a concept. In our $\mathcal{R}^n$ example, this would correspond to several goals having high probability density along the dimension corresponding to $\hat{g}$. For instance, pretty much any goal g_i which involves free human mobility on Earth in the short-term is going to require the convergent instrumental goal of 'Earth's oxygen level of 21%', which is a particular value of the 'Earth's oxygen level' dimension.

If convergent instrumental subgoals (CISs) exist, and since AGI by its nature is trained to be optimal with respect to a variety of goals, it seems like AGI **must** pick up any relevant instrumental goals. This intuition is explored more formally in Section 2.4.

2.3 What is 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.

2.4 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.

2.5 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”.

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.

[TODO: verify] The number of iterations for Algorithms 1 and 2 follow geometric distributions so the expected number of iterations are

$$\mathbb{E}_1 = \frac{1}{P(g^* = 1)} \qquad \mathbb{E}_2 = \frac{1}{P(g^* = 1 | \hat{g} = 1)}$$

So the speedup given by the ratio

$$\mathbb{E}_1 / \mathbb{E}_2 = \frac{P(g^* = 1 | \hat{g} = 1)}{P(g^* = 1)}$$

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 (such that $P(g^* = 1)$ is very small), then it seems reasonable that if a CIS exists, it will be used.

2.6 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.

3 The Sharp Left Turn, as a Concern. . . or Not

What is the Sharp Left Turn? A combination of:

- Discontinuous Shifts
  - We expect large goal-perturbations
- Capabilities generalize better than safety
  - "Really bad stuff" is hard to avoid
  - Convergent instrumental subgoals mean that even if goals shift, at least the CIS will still be hit

In what ways might we not be concerned about a Sharp Left Turn?

- "Really Bad Stuff" is unlikely to be hit
  - Challenging the "Safety doesn't generalize" argument
- Smooth Shifts + Reliable Control
  - If we could ensure goals never shift by more than a small amount, and we can quickly detect and correct this shift, then we aren't exposing ourselves too much to really bad stuff
- Smooth Shifts + Short Usage
  - If the likely mechanisms to induce goal shift can only move goals smoothly (like, there's expected to be a small upper bound on how much they can move per hour of deployment), and we only need to use it for a short amount of time before we get what we need, then we have a low risk of hitting "really bad stuff" before achieving our goal.
  - This seems to be the primary idea behind the "use AGI to advance alignment research"

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?

4 Assumptions, ranked by Credulity

4.1 Fast change

- There are realistic situations where, due to e.g. distributional shift, reflection, etc, goals can rapidly change so that when the new goal is satisfied, the old goal is unlikely to be satisfied: $P(g_0|g_1) \ll P(g_0)$

4.2 Reliable Control

4.3 Available CIS will be Exploited

5 Experiments

5.1 Continuous / Discontinuous Shifts

- The fast change assumption can be verified by demonstration of cases where goals have high sensitivity to the sort of perturbations we expect in relevant situations in the real world.

5.2 Causal confusion

- Changes in environment

5.3 Likelihood of “Really bad stuff”

- Social studies on “fragility of value”

6 Why it matters

for this section, try to quote directly and evenly from both sides, and stick to cruxes that would hold only in either case (not ones that people would likely stick to no matter what)

- Timing of when to slow AGI development
- OpenAI’s “slow down clause”
- Whether it’s safe to use AGI to further alignment research

7 Acknowledgements

This research is made possible via a Long-Term Future Funds grant from Effective Ventures.