Testing the Sharp Left Turn: A Probabilistic Framing and Experimental Desiderata

David Reber

April 26, 2023

People I want feedback from, in order: Victor, Nate Soares, Holden, Tsvi Benson-Tilsen, Dan Hendricks.

TL;DR

- Sharp Left Turn Hypothesis postulates that all of the following are true:
  - Claim A: Large goal-shifts will occur.
  - Claim B: Safety doesn't generalize well across large shifts.
  - Claim C1: Highly convergent, robust instrumental subgoals exist.
  - Claim C2: Convergent instrumental subgoals will be learned if they exist.
- Challenging the Sharp Left Turn: Falsify any of the above claims.
  - Claim $\overline{\text{A}}$: Large shifts can be ruled out.
  - Claim $\overline{\text{B}}$: "Really Bad Stuff" is unlikely by default.
  - Claim $\overline{\text{C}_1}$: No robust convergent instrumental subgoals exist.
  - Claim $\overline{\text{C}_2}$: Convergent instrumental subgoals are unlikely to be learned.
- Experimental Desiderata, Rated by Difficulty of Empirics:
  - A **vs.** $\overline{A}$**:** Plausibility of Large Shifts: Assessing the likelihood of large shifts is easy if A is actually true, hard if $\overline{\text{A}}$ is true.
  - C_2 **vs.** $\overline{C_2}$**:** Likelihood of CIS's being learned: Evaluating the probability that CIS's will be unintentionally learned seems plausible with current ML methods, but it may be challenging to avoid watering down the problem.
  - C_1 **vs.** $\overline{C_1}$**:** Plausibility of Robust CIS's: Demonstrating the existence or nonexistence of realistic, robust CIS's seems quite challenging.
  - B **vs.** $\overline{B}$**:** Frequency of "Really bad stuff": Empirically testing just how fragile human value is seems challenging, if it relies on social sciences.

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 desiderata for experiments which can verify or falsify the hypothesis.

This seems particularly important to get right, because the validity or falsehood of this hyothesis seems pivotal for deciding 1. If/when/how much to slow AGI development, and 2. Whether it's safe to use AGI to bootstrap alignment research.

2 Formalism

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". In this post I'm not focusing on making strong claims about which assumptions are realistic, but rather about how to test those assumptions.

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.

Definition 1. Concept Space A concept is 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 exceed various thresholds (see Figure 1). Hence multiple utility functions may correspond to a single goal: indeed, each goal can have an infinite number of utility functions, and those are all equivalent in our view. For now, we'll refer to goals instead of utility/loss functions.

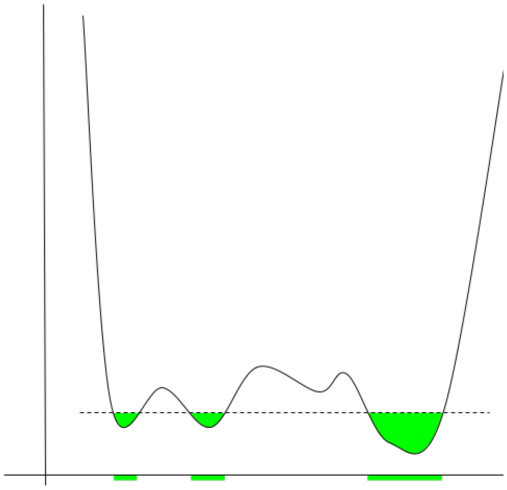

Figure 1: Loss and utility functions can be thresholded to provide an implicit goal.

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

Add: example which shows how a goal can be derived in RL settings.

Now we also want to restrict our attention to concept-tuples which are sensical; 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 (synonyms, conserved quantites in physics, etc)
- Heirarcial relationships like "Socrates is a man" and "a man is a person", 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.

For intution, just consider $\Omega = [-1,1]^n$ (though the results here don't rely on this).

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

For instance the problem of finding a promising research problem is somewhat hard, as modeled in Figure 2. A proposal needs to score well on multiple dimensions, so the overall probability of finding a research problem is low.

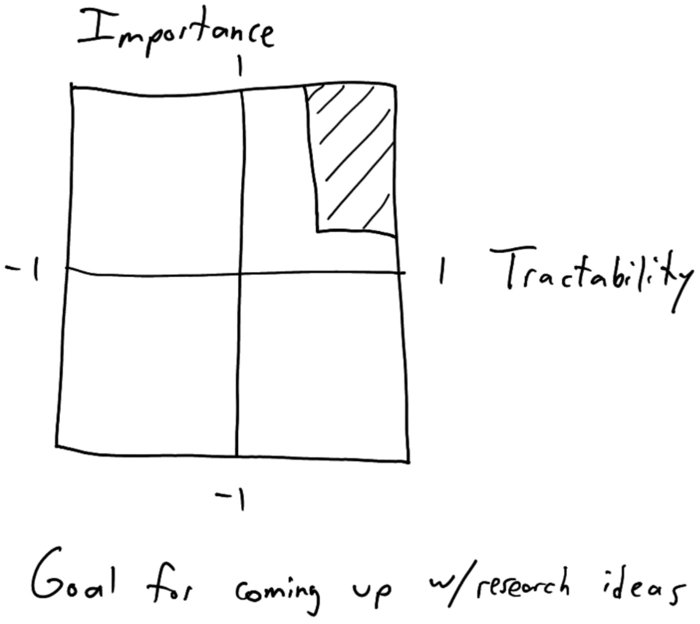

Figure 2: Goal for coming up with research ideas. Since it relies on several disjoint concepts taking on a narrow range of values, it is fairly difficult.

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 $P(g^* = 1 | g_i = 1) \gg P(g^* = 1)$; that is, achieving $g_i = 1$ makes $g^* = 1$ much more likely.

(As shorthand since we're considering binary goals, we'll write $P(g^*|g_i) \gg P(g^*)$.)

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 $g_i = \Omega$, but we are allowed to expand or restrict the goal.

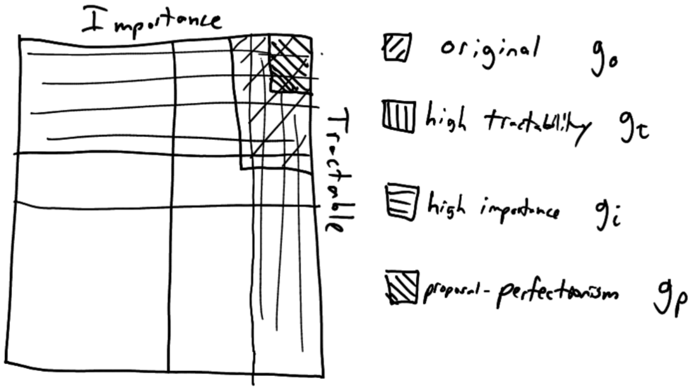

Figure 3: g_c, g_i, and g_p are all instrumental subgoals for the original goal g_o.

For instance, Figure 3 shows three instrumental subgoals for the goal of 'finding a promising research problem'. Note that as drawn, neither g_t (tractability) nor g_i (importance) are instrumental for each other. Meanwhile, g_p (proposal-perfectionsism) is instrumental for g_o, g_t, and g_i. However, if we were to add a new dimenension of 'time spent on proposal-writing', g_p may not be an instrumental goal for "having a research writeup soon" since it scores poorly on that dimension.

Now roughly speaking we want to say that a goal is convergent if it is instrumental for many goals:

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, $P(g_i|\hat{g}) \gg P(g_i)$ for all i.

However, it's not really interesting to say talk about convergence for goals which are all very similar to each other. So let's articulate a few ways to articulate how *disparate* our goals are.

Definition 8. Pairwise Disjoint Goals $g_1, \ldots, g_n$ are said to be *pairwise disjoint* if $g_i \cap g_j = \emptyset$ for all $i \neq j$.

This is a very strong condition, and it seems a bit too restrictive to only refer to instrumental convergence over pairwise disjoint goals. For instance, normal distributions with very little overlap seem to convey a satisfactory sense of distinctness, pictured roughly in Figure 4 (right).

An alternative framing of disparate goals is to say that satisfying one of the goals should not make the other goals more likely to be satisfied. This is a weaker condition than pairwise disjointness, but it's still fairly strong.

Definition 9. Mututally Noninstrumental Goals $g_1, \ldots, g_n$ are said to be *mutually noninstrumental* if $P(g_i|g_j) \leq P(g_i)$.

Q for Victor: maybe it'd be better to refer to g's from the start as being probability distributions refering to 'the likelihood of a goal being achieved' rather than 'the goal itself'. I'm a tad concerned that this might be confusing for the intended audience.

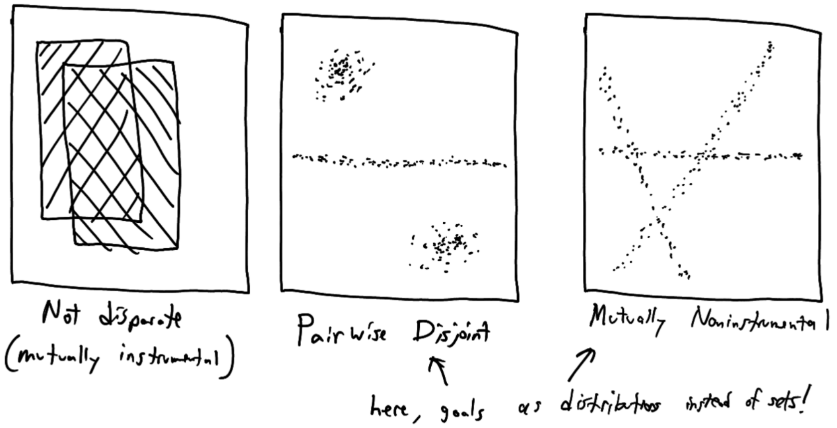

Figure 4: Left: two goals which are not disjoint enough to be interesting. Center: three goals which are pairwise disjoint. Right: three goals which are not pairwise disjoint, but are still distinct enough that it would be interesting to see a convergent instrumental subgoal for them. (Note that the center and right figures treat goals as distibutions; in our current framework, we could refer to the supopoort of these distributions.)

Note that if $g_1, \ldots, g_n$ are pairwise disjoint, then they are also mutually noninstrumental. However, the converse is not true: consider the goals $g_1 = \{x \in \mathbb{R}^n : x_1 \geq 0\}$ and $g_2 = \{x \in \mathbb{R}^n : x_1 \leq 0\}$. These are not pairwise disjoint, but they are mutually noninstrumental.

For the rest of this post, we'll assume that convergent instrumental subgoals (*CIS*'s for short) refers to goals which are convergent relative to a large number of mutually noninstrumental goals.

If convergent instrumental subgoals exist, and since AGI by definition is presumably optimal with respect to a variety of goals, it seems like AGI *should* pick up any relevant instrumental goals. We explore this question more in Section 2.5.

2.3 How likely are Convergent Instrumental Subgoals?

If we were sampling k goals uniformly from Ω (say, $\Omega = [0,1]^n$), then it's unlikely that we would find a subgoal $\hat{g}$ which is instrumental for all k goals; since they're sampled uniformly, their overall density is too spread out to collectively benefit from a restriction to a small subset of Ω. However, if we were sampling goals in a concentrated manner, such that large sections of Ω can be ignored, then the probability of a convergent instrumental subgoal would be much higher.

Optimization in the real world is far from uniform...[TBD].

Add concrete mathematical example for RL goal shifts as a result of long-term planning, so it's not just the 'supervised learning' example.

2.4 What is Generalization?

In order to talk about the sharp left turn, we need a notion of generalization.

Example 10. If our goals are derived as thresholds applied to some loss function, then distributional shifts on the data which defining the loss function will induce a change in the goal.

This change of goal happens already on a smaller scale with stochastic gradient descent. Goals defined via the loss function change on each batch (although empirically in SGD the sequence of them still leads you where you want).

See Figure 5.

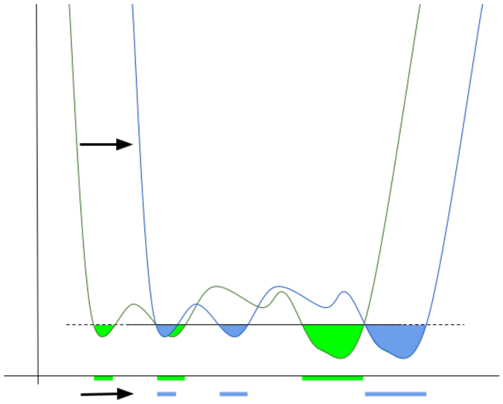

Figure 5: A loss function which is a function of the data. The goal is to minimize the loss function. If the data distribution changes, then the goal changes.

Another, more philisophical example which we would like to be able to capture in our formalism, is the following:

Example 11. Consider the effect of a change in the research question. For instance, consider the following two research questions:

- How do neurons in the brain work?
- How do neurons in the brain relate to each other over small timescales?

The first question is a very general research question, while the second is a very specific. An AI tasked with answering the first question might, through self-reflection, realize that the second question is more actionable while still relevant to the first question. This change in the research question induces a change in the goal (albeit in a way which is still fairly consistent with the original goal).

Let's represent a change in a goal as $g_0 \to g_1$. Maybe it's a small change, maybe quite distinctive, but if $\hat{g}$ is a robust CIS then $P(g_1|\hat{g}) \gg P(g_1)$ so capabilities generalize (see Figure 6 for a rough illustration). Meanwhile, if g_1 represents a sufficiently large deviation from g_0 such that $g_0 \cap g_1 = \emptyset$, then $P(g_0|g_1) = 0$ so the original objective won't be satisfied at all.

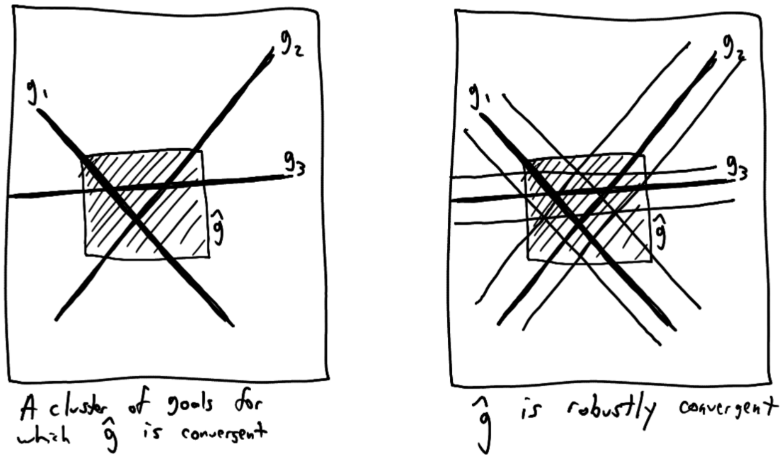
g_1
g_2
g_3
$\hat{g}$
A cluster of goals for which $\hat{g}$ is convergent
g_1
g_2
g_3
$\hat{g}$
$\hat{g}$ is robustly convergent

Figure 6: Convergent instrumental subgoals are likely to be robust to distributional shifts.

On the other hand, let's consider the goal of "avoids really obvious bad stuff"; call this goal α. If α is hard it's not robust to changes in the goal (as claimed in Value is Fragile). If α is large in Ω, it's very generalizable. In a limited 2-d drawing, this would look like Figure 7.

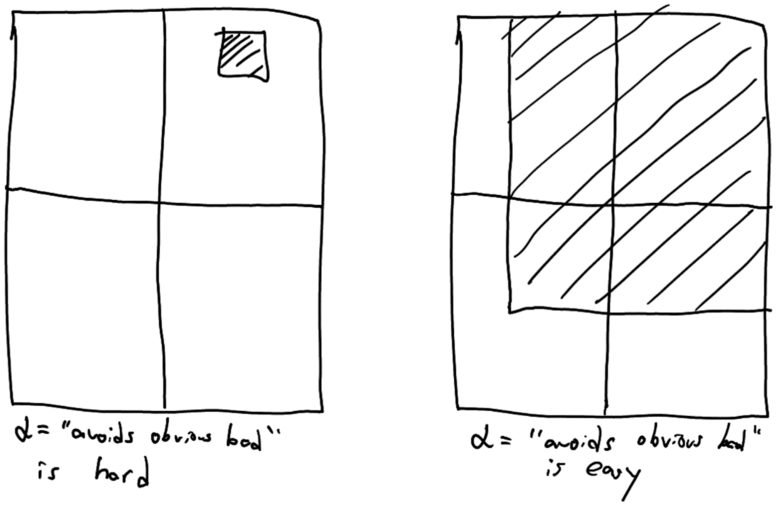

Figure 7: Illustrative way in which the goal α of "avoids really obvious bad stuff" could end up being hard (left) or not (right).

Hence, the claim that "capabilities generalize more than safety" relies on the assumption that the goal α is hard.

2.5 Are CISs likely to be learned in general?

Assume $\hat{g}$ is a convergent instrumental subgoal for some goals $\{g_i\}$, but not other 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 $P(g_k)$ for some $k \in \{i\} \cup \{j\}$, and want to find the probability that $\hat{g}$ is satisfied conditional on g_k being satisfied.

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

For simplicity, assume $P(\hat{g}|g_i) \approx 1$, and $P(\hat{g}|g_j) \approx 0$. (Intuitively, this states that if $\hat{g}$ is a CIS for G^*, it will be learned with high probability; if $\hat{g}$ is not a CIS for G^*, it is unlikely to be learned.) Then $P(\hat{g}|g_k) \approx \sum_i P(g_i)$. 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 $P(\hat{g}|g_i) \approx 1$ assumption.

2.6 How likely is a CIS to be learned for a particular goal?

To check the $P(\hat{g}|g_i) \approx 1$ assumption, suppose $\hat{g}$ is a CIS for a goal g^*. That is, $P(g^*|\hat{g}) \gg P(g^*)$. 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

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

So the speedup given by the ratio

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

Since the numerator is bounded, this ratio seems largely driven by how small $P(g^*)$ might be.

Key takeaway: if your goal g^* is hard (such that $P(g^*)$ is very small), then it seems reasonable that if a CIS exists, it will be used.

2.7 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.

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

Let's summarize the ideas behind the sharp left turn hypothesis, to get a sense for how we could validate or falsify it experimentally. Note that while the claims below are general enough to apply to any system with a goal, both sides should probably be focused on demonstrating their respective claims *for the situations we actually expect to happen in practice*.

TODO: make each of these statements a concrete probabilitistic claim

3.1 What is the Sharp Left Turn?

The sharp left turn can be decomposed into the following parts:

Claim A: Large Shifts. We expect that at least some large goal-pertubations will occur by default when AGI is deployed.

The idea that "capabilities generaizing better than safety" can be broken up into two parts: B (safety doesn't generalize well) and C (capabilities generalize well).

Claim B: Safety doesn't generalize well to large shifts. The goal α of avoiding obvious bad stuff, is hard.

The idea that capabilities generalize well depends on two parts: C_1 (highly convergent, robust instrumental subgoals exist) and C_2 (CIS's will be learned if they exist).

Claim C_2: Existence of Robust CIS's. There exist convergent instrumental subgoals which are highly convergent for a wide range of disparate goals, and which are robust to perturbations in these goal.

Claim C_2: CIS's will be learned if they exist. If $\hat{g}$ is a CID for G^*, it is very likely to be learned.

The overall form of the Sharp Left Turn hypotheis is a conjuction of "A and B and C_1 and C_2". Claims A and C_2 take the form of 'there exists'; Claims B and C_2 are claims about likelihood. So a priori it will probably be easier to rule out Claims B and C_2 than Claims A and C_1.

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

To challenge the sharp left turn, we need to challenge at least one of the claims A, B, C_1, or C_2.

Claim $\overline{A}$: Small Shifts We can confidently rule out large shifts from occuring.

- 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"

Claim $\overline{B}$: α is not a hard goal. "Really Bad Stuff" is unlikely to be hit by default.

Claim $\overline{C_1}$: No Robust CIS's exist. There are no convergent instrumental subgoals which are highly convergent for a wide range of disparate goals, and which are also robust to perturbations in these goal.

Claim $\overline{C_2}$: CIS's are unlikely to be learned even if they exist. If $\hat{g}$ is a CID for G^*, it is still unlikely to be learned.

4 Experimental Desiderata

The key for both sides to demonstrate is that the claims they are making are true in the situations we expect to happen in practice. However, short of having AGI available, we can't really demonstrate this (and if the sharp left turn were legit, it would be nice to show *prior to* AGI). But it seems more likely that we can test individual subclaims in realistic settings without full access to an AGI.

4.1 Plausibility of Large Shifts

To differentiate between Claims A and $\overline{A}$, we need to demonstrate whether or not large goal-shifts are expected to occur suddenly at any point.

- Sensitivity analysis of causal models. How much does the goal change if the causal relationships in the enviornment change?
- RL experiments with different environments

4.2 Frequency of "Really bad stuff"

To distinguish between Claims B and $\overline{B}$, we need to demonstrate whether or not "really bad stuff" is the default expectation *given* that an AGI encounters a large goal-shift.

- Social studies on "fragility of value"

4.3 Plausibility of Robust CIS's

To distinguish between Claims C_1 and $\overline{C_1}$, we need to demonstrate whether or not there are robust CIS's. This seems promising for current RL methods, but the challenge will be in demonstrating that any instrumental goals are actually 1. convergent for a wide range of disparate goals, and 2. they are robust to *large* perturbations in these goals.

4.4 Likelihood of CIS's being learned

To distinguish between Claims C_2 and $\overline{C_2}$, we need to demonstrate whether or not CIS's are likely to be learned, if they exist.

4.5 Ranking of difficulty

If I were forced to rank the expected difficulty of running convincing experiments without any more time to consider deeply, I would rank them as follows:

- Plausibility of Large Shifts: Easy if A is true, hard if $\overline{A}$ is true
- Frequency of "Really bad stuff": Hard both ways
- Plausibility of Robust CIS's: Hard both ways
- Likelihood of CIS's being learned: Medium

I could imagine this ranking changing drastically if I spent more time thinking about it, but I think it's a reasonable starting point. The most likely thing that would change the difficulty is if there already existed academic work which sufficiently resolves the various questions.

5 Relevant Academic Fields

I'd like to ground this formalism in existing fields as much as possible without distorting the content. [TODO: do brief survey, talk to people, etc.]