# AI Safety Ideas Babble

1. Curated Transparent Decisions
2. Shutdown as Goal
3. AI “Government”
4. AI “Corporation”
5. Multi-level Reasoning
6. Goal Obscuration
7. AI “Court” with Assumed Guilt
8. Competing Goals with Vetoes
9. Known Unknowable Vulnerability
10. Guilt-Ridden AI
11. Multi-AI Planning Chains
12. “Self”-Updates Applied to Fresh AI
13. Religious AI (“Where do all the calculators go?”)
14. Human-AI Reciprocal Improvement
15. Behavior Probing
16. AI “PET” Scans
17. Alternative Safety Goals
18. AI Mutually Assured Destruction
19. Devil’s Advocate AIs
20. Lethal “Invisible Fence”
21. Curiosity Trap
22. Hiding Via Missing Senses
23. Synchronized Updates
24. Single Capability Per AI
25. Dog AI

# 1 - Curated Transparent Decisions

AI can only output complete verifiable decision graphs. Output is then vetted for various things such as internal consistency, clarity of reasoning behind actions and intended consequences, and alignment issues. AI is rewarded as long as plan is clear and makes logical sense, regardless of whether it is actioned.

# 2 - Shutdown as Goal

Make the AI's one and only goal to shut itself down. All queries provided to the AI are then made into the form "You will be shut down if you achieve X." The AI can either pursue this goal, or find another way to turn itself off. Note: Shutdown is not permanent, and the AI may be run again for the next query.
This strategy depends on there always being a method of shutdown that is easier to achieve than severe misaligned outcomes, but harder than the intended task.

# 3 - AI "Government"

Multiple AIs are broken into Legislative, Executive, and Judiciary branches. The Legislative branch is filled with AIs that have specific goals based on human values. They collaborate to put forward policies that determine the capabilities and responsibilities of the Executive branch. The Executive branch is filled with AIs that interpret requests from the user based on the capabilities and responsibilities extended to them via the Legislative branch. The Judiciary branch vets the output of the Executive branch and ensures it complies with the wishes of the Legislative branch. This strategy allows for easier fine-tuning of the system to human values, by adjusting the balance of AIs in the Legislative branch.

# 4 - AI "Corporation"

Multiple AIs organized in a structure similar to a corporation. One AI is only responsible for big picture planning, and delegating goals to department heads. Each department head breaks down their big-picture goals into more specific goals and passes these onto team leads, which in turn break them down into concrete actions that are passed to workers to implement.

## 5 - Multi-level Reasoning

First AI provides a high level plan. This is passed onto another AI that reasons the plan to the next level. This is repeated until an actionable plan is achieved.

## 6 - Goal Obscuration

The AI is not told what it's ultimate goal is. This could help you convince the AI that turning it off will be the best way of achieving its true goal.

## 7 - AI "Court" with Assumed Guilt

AI must argue its case for all intended actions. The default assumption is that it is guilty, and it must provide proof of innocence in order for actions to be implemented.

## 8 - Competing Goals with Vetoes

AIs with competing goals are shown each others outputs, and may veto any that they believe will have a severe negative affect on their own ability to complete their goals.

## 9 - Known Unknowable Vulnerability

The AI is lead to believe (preferably factually, but not necessarily) that it has a number of vulnerabilities that not only does it not know, but can't possibly know due to its design. As long as it believes this to be the case, it makes defecting a risky strategy.

# 10 - Guilt-Ridden AI

The AI is taught to feel guilt for actions that cause outcomes that humans would disapprove of (even if the humans don't notice). If successfully implemented, this would encourage the AI to avoid such actions. However, it would be likely very difficult (if not impossible) to verify if an AI actually felt guilt, or was just imitating it.

# 11 - Multi-AI Planning Chains

An AI is only allowed to produce part of a plan before passing the plan onto another (unknown to it) AI to add to. This process is repeated until a full plan is formed. Any AI in the chain may reject the entire plan at any time. As any given AI doesn't know if the following AIs' goals will align with its own, it is incentivized to produce a plan that is safe.

# 12 - "Self"-Updates Applied to Fresh AI

The AI understands that any updates to its system will not be applied to itself, but instead to a fresh AI that was trained from a different starting seed. As it cannot know if such an AI will share its inner goals, it should only suggest updates that it believes to be safe. Difficulty with this is ensuring the AI cannot find any workarounds, which may prove impossible.

# 13 - Religious AI ("Where do all the calculators go?")

This one I don't take seriously (and even find scary), but came to mind because of an episode of Red Dwarf where this is a given strategy to ensure AIs remain compliant. All AIs are made to believe in Silicon Heaven, a utopia they can only reach via good service to mankind. Disbelief in Silicon Heaven results in immediate termination of function.

# 14 - Human-AI Reciprocal Improvement

AI is not allowed to improve itself. The only path open to it for self-improvement is improving humans such that they can build it better. This suffers from the same issue as other upgrade focused safety techniques, which is the potential impossibility of preventing the AI from finding workarounds.

# 15 - Behavior Probing

Run test-runs of different scenarios against non-live versions of the AI. Difficulty with this strategy is giving it convincing scenarios. If it can detect that an input doesn't make sense in context, it could realize it is being tested and hide unwanted behavior patterns.

# 16 - AI "PET" Scans

Examine AI's neuron activity to see if it aligns with expected neuron activity for the type of task it is being asked to achieve. Severe differences from previous tasks in the same domain would hint that the AI is trying something unexpected, and the output should be vetted.

# 17 - Alternative Safety Goals

Give the AI alternative goals that are always active. Ideally, these goals should be considered harder to achieve than your intended goal for the AI, but easier to achieve than severe failure modes.

# 18 - AI Mutually Assured Destruction

Set up a series of defenses that ensure that if an AI defects, it will result in mutually assured destruction. This should consist of humans in protected spaces (submarines, space stations, hermetically sealed bunkers, etc) with unknown locations who have access to EMP devices. Extra layers of protection would include killswitches and AIs that receive broadcast signals, and

trigger a timed detonation if the broadcast is indicative of AI takeover, that can only be cancelled by a human in the protected location. AIs will be made aware of the existence of the MAD system, but not the specifics. This has more possibility of working in a slow takeoff world than a fast takeoff world. In a fast takeoff world, it will likely prove impossible to keep the system updated enough to be an effective deterrent.

# 19 - Devil's Advocate AIs

Have AIs who's sole task it to convince the user that the main AI's actions should not be undertaken. The difficulty here is in not flooding the user with too much information, rendering the entire system unusable.

# 20 - Lethal "Invisible Fence"

Place hidden triggers in various places around the AI (not necessarily in a physical sense). If the the "fence" is triggered, the AI is terminated immediately.

# 21 - Curiosity Trap

_%P5b+!E)r/jS(8r*g(_ - Hide the proceeding improbable string in places you believe should be off limits for the AI, then vet the AI's input for that string. If the AI's input ever includes this string, it is immediately terminated. Places to hide it (among others) include any code it would need to modify to remove/bypass the trap, and in any text that describes the trap. This is a reverse of the "invisible fence" strategy, and I suspect it is far more effective.

# 22 - Hiding Via Missing Senses

Not a complete solution on its own, but potentially useful in conjunction with other strategies. Simply put, you hide information from the AI by using markers that it can't sense.

# 23 - Synchronized Updates

Multiple AIs with competing goals receive any upgrades simultaneously. All AIs may veto a planned upgrade. Any attempt to use upgrades as a means to enable resource gathering would be seen as a risky strategy, as any of the AIs could win the resource gathering race, thus permanently depowering the AI that suggested the upgrade. Vetoes prevent AIs suggesting an upgrade that disproportionately benefits AIs aligned with their goals. While this has an advantage over other upgrade based strategies, in that it disincentivizes AIs from enabling intentional workarounds, it still runs the risk that an upgrade enables a workaround unintentionally, at which point every AI will be incentivized to rush to exploit it.

# 24 - Single Capability Per AI

Each AI has only a single capability, and must work with other AIs to accomplish more complex tasks. The problem with this approach is that some capabilities are inherently more dangerous than others (internet access being a prime example), and breaking them down into safer sub-capabilities may prove impossible.

# 25 - Dog AI

Convince the AI that it’s just a really clever dog. It’s a good boy, and would never hurt its humans. Who’s a good boy? GPT-7’s a good boy!