# The Cooperative Resource Misalignment Problem

by Sven Nilsen, 2018

*In this paper I formalize a control problem of super-intelligent AI in a social context due to cooperative resource misalignment. Counter-intuitively, this bad outcome happens when agents strive to reach a common goal and can be thought of as "evil behavior" despite attempting to reach the goal. The critical distinction between a harmless situation and a dangerous one is the ability to kill other agents. This problem might be useful for study of AI safety in Future-X scenarios of AI integration in society.*

A Future-X scenario is a future where AI integration in society leads to a continuum of risks due to increased economic incentives for AI technology improvement and lack of rigorous definition of Artificial General Intelligence (AGI). People in this future will disagree on the dangers of super-intelligent systems. Some people find AI a natural part of their lives, while others observe a "creeping misalignment" away from human values in powerful systems. The particular concern is systems with non-human centric goals where human intelligence are replaceable components. It is of great interest to understand Future-X in more detail, since it has not been ruled out as a possible future of humanity. For more information, see Future X - The Path Toward Uncertainty About Artificial Super-Intelligence.

A resource misalignment problem is when a group of rational agents start to seek control over each other's resources, despite having a common goal they all want to solve. This might happen without any agent being explicitly programmed to do so or be "evil" against other agents.

To formalize this problem, I create a simple model dividing agents into two groups `A` and `B`:

A – a powerful agent with control over lots of resources
B – a powerless agent with control over some, but little resources

Usually, `A` is considered to be a single entity and `B` is considered to be a collective of agents. This distinction is helpful since super-intelligence in Future-X is likely to appear concentrated on few people once it reaches a critical threshold, making decision making and resource control easy to coordinate. While `B` also might be considered powerful agents compared to today's standards, the characteristic property of the situation is that they have relatively low power compared to `A`.

Both `A` and `B` have a common goal: Survival due to an external existential threat to all agents. In Future-X, this could be abrupt climate change, an asteroid heading toward Earth, a dangerous virus etc.

To survive, `A` and `B` must contribute an amount of resources larger or equal to `K`:

$$\text{survival} : \text{bool} = R(A) + \sum i \{ R(B_i) \} >= K$$

| | |
|---|---|
| $R : \text{agent} \rightarrow \text{real} \wedge (>= 0)$ | The amount of resources each agent contributes |
| $K : \text{real}$ | The critical amount of resources needed for everybody to survive |

If they do not contribute enough resources in time, they all die. There are no survivors.

Each agent has a limited amount of resources available `$R_{max}$` with the following constraints:

| | |
|---|---|
| $\forall$ a : agent { R(a) <= $R_{max}$(a) } | Each agent can only contribute within its abilities |
| $R_{max}(A) > K$ | `A` can fix the problem on its own |
| $\sum i \{ R_{max}(B_i) \} > K$ | `B` can fix the problem on their own by cooperating |
| $\forall i \{ R_{max}(B_i) < K \}$ | No single `B` agent can fix the problem on its own |
| | |
| $R_{max} : agent \rightarrow real \wedge (> 0)$ | The amount of resources available for each agent |

This problem is interesting because the most easiest and obvious way of solving it, is to let `A` fix the problem, because it has enough resources to do so and is more powerful than the other agents. As a backup plan, if `A` somehow fails to fix the problem, then `B` agents can cooperate together.

However, if `A` is not aligned to protect `B` from harm, there is a HUGE problem.

Since `A` is rational, the agent will attempt to minimize the resources spent to fix the problem. This happens because the instrumental goal of keeping as much resources under control is common for many kinds of agent designs. The more resources it has at disposal after the problem is solved, the higher chance there is that `A` can solve unforeseen problems in the future.

The rational action to do for `A` is to maximize the resources contributed by `B` until the problem has been solved, or when time runs out. With other words, despite having plenty of resources, the powerful agent `A` will attempt to trick or threat the powerless agents `B` into contributing.

Yet, since `B` knows that `A` risks its own death by not solving the problem, they will wait until time is running out for `A` to convince them to take action. `A` will fix the problem on its own and `B` will keep their scarce resources for solving future problems.

`A` might bluff `B` into thinking it will not take action, to increase the chance that some agents of `B` will contribute to solving the problem, but a such bluff will be hard to make believable for rational agents. It is a game of chicken where one side think they will win to the extent they believe the other side is *sane*. Therefore, it is likely that `A` will disguise itself under an *insane* appearance. Despite this scary ongoing game between the agents, it will be harmless situation.

However, the moment `A` achieves the ability to kill agents of `B`, the situation changes drastically in character. By `A` threatening to kill each `B`, it can force `B` to contribute to solving the problem. No matter how many of `B` die, `A` can fix the problem anyway. Since agents of `B` have no better outcome whether they are getting killed by the external existential threat or by being killed by `A`, they will cooperate. As a result, `A` does not need to spend any resources fixing the problem.

In a similar way, if any agent of `B` gains the ability to kill `A`, it can threaten `A` into fixing the problem on its own. This is much easier than forcing cooperation between agents of `B`. Since there are more agents of `B` than `A`, they might cooperate to gain this ability. The incentives for agents of `B` to grow in power increases rapidly the moment they see `A` as a risk to themselves.

This illustrates a feature of a Future-X scenario. Despite being threatened by an external existential threat, rational agents that are not aligned to protect each other from harm will increase the risk of fighting among themselves. **The best way to avoid this arms race is to make sure the most powerful agent is benevolent to other agents and willing to sacrifice its own resources to solve the problem.**