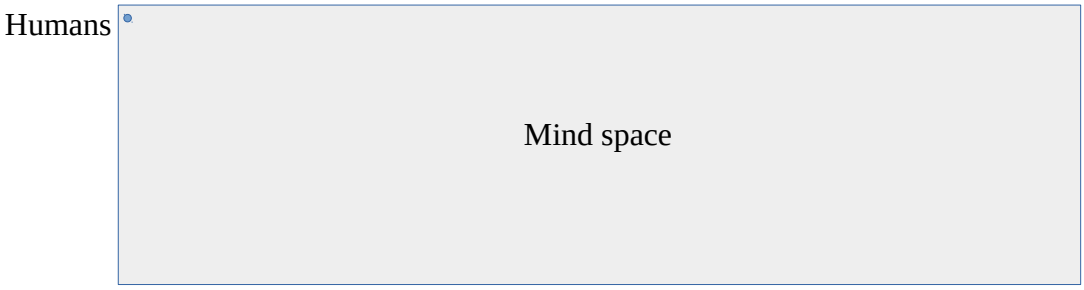
Circling Mind Space by Optimal Collective Intelligence

by Sven Nilsen, 2018

A common terminology in the field of Artificial Intelligence (AI) used to distinguish continued advancements in technology categorizes intelligence broadly into "narrow", "general" and "super" intelligence. In this paper I represent a different way of thinking about intelligent behavior that is easier to formalize and better integrated with mathematical and philosophical ideas of intelligence. It might also be more intuitive for what people in general consider to be intelligent behavior.

Mind space is an idea that all possible minds might be arranged in some space with coordinates, such that similar minds are located nearby each other. For example, all humans would be located in a small dot inside this space compared to the total size of possible mind designs.



An optimal collective intelligence is a set of agents which every agent performs equally or better than all other agents in the set. This property is uniform, which means that no matter how we remove agents from the set, the new sub-set has the same property. What we mean by optimal collective intelligence depends on how we define environments. Given a such set, an agent can be constructed that gives the best effort possible as the collective designs of agents within the set.

Instead of considering the space of all minds, one can *filter* mind space by different definitions of optimal collective intelligence. What we end up with is a space where only "the best" minds for any environment remains. The dot of human minds gets fractured into smaller dots, e.g.:

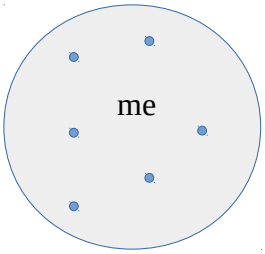
The tasks that these people perform are environments that demand a certain skill to navigate efficiently. While surgery is a name we use for a specific kind of environment, there are many types of surgery and

the same people are usually not the best doing all of them. This means that each dot in the filtered mind space is replaced by new dots when you “zoom in”, organized by refined definitions of environments.

There is a striking similarity between minds and the environments they perform well. A surgeon requires good motor control, which is often related to how the brain works in a particular area. One can relax the definition of similar minds in mind space locally and give more neighborhood weight to similar environments where the agent does its best compared to others in nearby mind space.

The result is that mind space is turned into a star-like map. All possible minds still exist within the mind space, but most of them are located “between the stars”. This star-like map makes it possible to locate a particular kind of mind without needing all the information from the whole space. Neither do you need all ways of measuring intelligence, you can just use some of them.

For example, I am located inside the human dot, but between minds who are the best at different tasks:

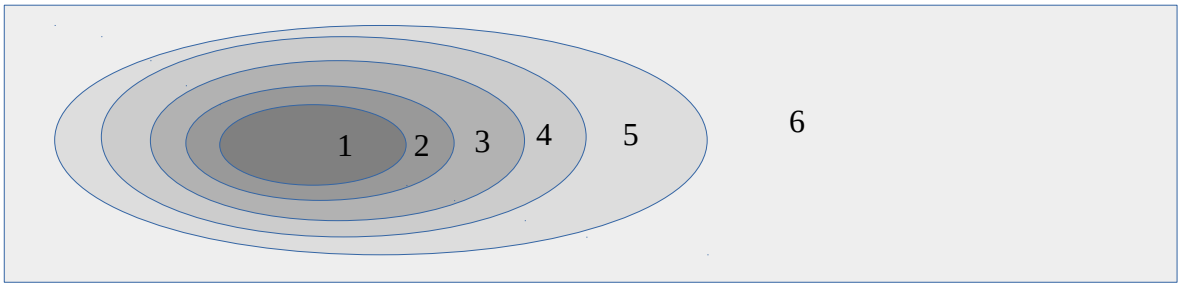


I am particularly interested in intelligent behavior seen through these 6 filters of mind space:

1. Optimal control (environments with known exact solutions and exact algorithms)
2. Heuristics and approximate algorithms (environments that are not well-defined)
3. Instrumental rationality (agents with a single goal)
4. Zen rationality (agents that reasons about multiple goals)
5. Artificial and biological life (intelligent behavior in a broader sense)
6. Mind space (all minds)

Some agents only perform their best in well-defined environments, using an algorithm that might be proven to perform equally or better than any other algorithm. This filter is called “optimal control”.

If you circle these dots, one will find that for every “best mind”, it is possible to construct a mind performing equally well from optimal collective intelligence by including new dots from other filters:



With other words, “optimal control” can be placed on top of “heuristics and approximate algorithms”, which can be placed on top of “instrumental rationality”, which can be placed on top of “zen

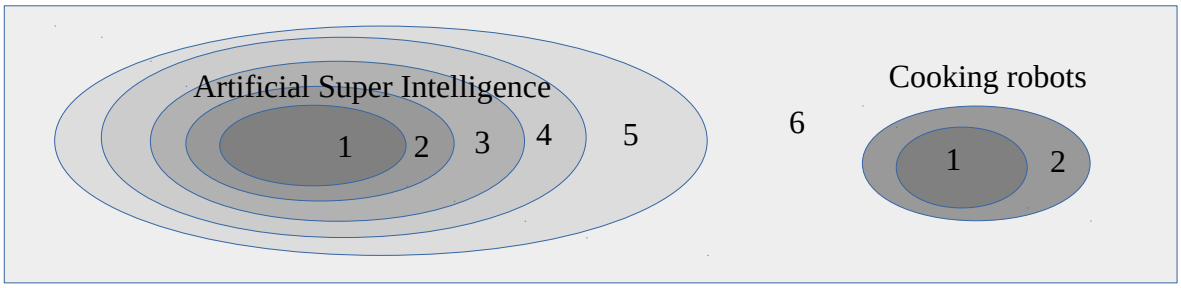
rationality", which can be placed on top of "artificial and biological life". When one set is placed on top of another, it is extended with the new agents underneath.

Optimal collective intelligence makes sure that the extended minds are similar. Therefore, they are located in the neighborhood of their original core, but with additional components. This operation can be performed for any "best mind". When circling around these minds, you get a hierarchy of sets.

When we categorize intelligent behavior, we do not use the exact coordinates of the "best minds". Instead, we circle the "best minds" by similarity and then check whether the mind we want to categorize is located within the circle.

For example, a computer program playing chess might be similar to the best chess program. It is natural to think of such programs to display some kind of intelligent behavior. This is because in mind space, such programs are located nearby each other. However, the kind of intelligent behavior might be very different from other minds. This is why a program good at playing chess might not be good at e.g. cooking.

Mind space does not contains just one peak, there are different neighborhoods of minds:



One can think of it as a landscape with peaks and valleys. For example, in far future, a genetically designed animal to be as smart as possible, might be close to "super" intelligence in mind space without actually being a "super" intelligence. Yet, it is closer the ASI-peak than the cooking-robot-peak.

What we mean by "narrow", "general" and "super" intelligence is roughly the following:

ANI = 1' + 2'	Artificial Narrow Intelligence
AGI = 1' + 2' + 3'	Artificial General Intelligence
ASI = 1' + 2' + 3' + 4'	Artificial Super Intelligence

When creating an ASI that operates autonomously and replicates itself, one might consider it an artificial life form, which is 5'. However, this not a necessary behavior. This is why I only included 1-4.

If there exists a problem which can be solved though ANI, some AGI agent will be able to solve it. It is not guaranteed that any AGI agent will be able to solve it! When referring to "general" versus "narrow" intelligene, what one means: AGI is considered better than ANI, because when considered as optimal collective intelligence, it can solve every problem that ANI can as optimal collective intelligence, plus more.

In practice, it is *impossible* to construct such sets of agents, because they grow too large and the computational power required to do so goes to infinity. However, such agents might be approximated.

Instrumental vs Zen Rationality

If you have a goal and an agent that can interpret it and achieve it with optimal performance, then the agent is considered *instrumentally rational*.

If you have multiple goals and you need the agent to figure out what the right goal is while achieving it optimally, then it is considered *zen rational*.

For any "best" instrumentally rational agent `$a_0$`, there exists at least one "best" zen rational agent `$a_1$`. There exists some environment where `$a_1$` performs better than `$a_0$`, yet `$a_1$` performs at least equally well in the environment where `$a_0$` performs best compared to other instrumentally rational agents:

> performance(a_0, e_0) <= performance(a_1, e_0)
> performance(a_0, e_1) < performance(a_1, e_1)
>
> a_0 : instrumental_rational
> a_1 : zen_rational

The relation between instrumental and zen rationality is similar to the relation between well-defined and not well-defined environments. In an environment that is not well-defined, one can not use optimal control get obtain best performance. Heuristics or approximate algorithms are necessary. Similarly, there exists environments where zen rationality is required, but those environments are very hard to formalize.

Since humans display complex behavior that might be approximately zen rational, an idea is to draw on inspiration from resolving goal conflicts in human psychology, e.g. ground "friendship" in the behavior of the agents. This requires heavy game theoretical insights.

In order to solve all problems that humans do, it sufficies to specify a single goal per problem and apply instrumental rationality. However, in order to have self-improving and self-aligned AI (abilities that no humans have to an extreme degree), in practice it must be able to reason about its own goals and resolve conflicting goals. Therefore, instrumental rationality is not sufficient to describe the behavior of ASI. It would, if you knew the goal that specified how the agent should self-improve, but since you do not, it is more like zen rational behavior.

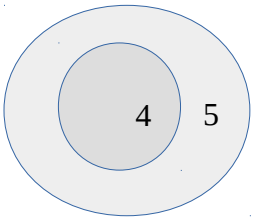
An approximate zen rational agent might have some knowledge of optimal control, some knowledge of heuristics and approximate algorithms, some knowledge of instrumental rationality and some knowledge of zen rationality. However, the term "zen rational behavior" can also be used to describe agents who know nothing of e.g. optimal control. For example, humans might be approximate zen rational on some problems that are non-analytic within the framework of instrumental rationality. This is because humans live in complex environments where their own goals are often unknown.

The “Arrow” of Intelligence and Growth of Diversity

Any way of implementing an approximate zen rational agent requires some kind of machine or biological form. It falls under the category of “artificial and biological life”. However, we do not consider this category as generally smarter than zen rational agents. *The extended category is too big.*

The “arrow” of intelligence breaks down when passing zen rationality, because we have not yet invented words for what lies beyond it which can be called “intelligence” or “rational”.

A zen rational agent might be extended with e.g. “using eyeballs to see”, which is a biological behavior. For every “best” zen rational agent, such extensions are possible. This is why zen rationality can be circled inside “artificial and biological life”, but there is no “arrow” of intelligence between them.



Circling in mind space using optimal collective intelligence generalized the notion of “arrow” which breaks down between 4th and 5th level.

All zen rational agents have less in common than all instrumentally rational agents. With other words, zen rational agents display more *diversity* in behavior. This is important to remember, because as you climb up the ladder of capabilities, you get more possible designs.

The “arrow” of intelligence by technological development is actually moving in all kinds of directions over time. AI does not get smarter mostly in a direct sense (e.g. performance), it gets smarter mostly in a *broader* sense. You get more diversity of intelligent behavior, not necessarily an integrated system that solves all problems optimally.

An intuitive way of thinking of diversity in intelligence, is considering that the “best minds” occupy a larger chunk of mind space when circling around them, as you increase the complexity of environments. The reason for this is simple: For every “best mind” in one category, if there are two ways to extend the agent, then the new category must be twice as large. Hence, diversity grows.

Yet, the definition of an ASI requires an agent being good at solving problems related to e.g. optimal control. This is why AI people talk of the “arrow” of intelligence from ANI → AGI → ASI. We pretend that the agent of optimal collective intelligence exists from the best of existing technology. A such agent might be possible to approximate, so experts estimate when a such implementation or something similar in capabilities might arrive.

Growth in Diversity Makes Arrival of Artificial Super Intelligence Uncertain

Whether humanity develops a specific ASI agent is currently unknown. What we expect is that some capabilities from e.g. zen rationality will become better understood over time. It is fully possible to research ASI concepts without actually building one that qualifies as ASI. Zen rationality is something I believe is important for ASI, but it will not become “real” ASI before it is integrated in a way such that it performs far beyond human capabilities on all domains that matters. While figuring out the concepts might be easy, it could take a very long time before it is integrated in a single mind, operating on all levels 1-4.

Artifical and Biological Life – Seen as Intelligent Behavior

Biological life are full of examples that can be considered intelligent behavior, but which not falls into a well-defined optimization problem.

For example, viruses might be good at replicating, but they do not move by their own energy like bacteria. Instead they prey on other life forms capable of self-replication. It is impossible to replicate exact conditions under which viruses replicate, so one speaks of their general behavior in a vague sense, one that is understood by humans, but not by computers. I think it is important to be able to speak of intelligence in this context, but it is not “stacked” according to optimal collective intelligence from optimal control to zen rationality, like the way intelligence is considered in AI safety research.

One could say “virus optimizes self-replication!”, yet the problem is that the environment which virus self-replicate is very hard to define. It is not formalized in the same sense instrumental rationality is attempted formalized with various decision theories. When writing a model of virus behavior, it is merely predictive, but when writing a optimization algorithm, it *is* the behavior.

This is why I think that artificial and biological life deserves a notion of *intelligent behavior* that is broader than the philosophical frameworks we have for optimal rational behavior.