Self Rationality

by Sven Nilsen, 2018

A comprehensive theory of how humans behave and what we are requires a satisfactory background knowledge about genes, natural selection, evolution, neurology and psychology. Instead of all these complex systems working together in non-linear ways, I need an extremely simplified model that gives good predictions of how humans will behave in various situations. This extremely simplified model will be used in AI research to contrast how humans behave against other simplified models of rationality. I call this model "self rationality" because it is about filtering, compressing and optimizing information to preserve a self over time, an agent's theory about itself and extensions of itself, such as a group of agents. Bonus chant at the end of paper.

There are two properties that are required for an agent to behave consistently to preserve a "self":

1. Awareness of how it is behaving
2. Knowledge about how it typically behaves

When decisions are made whether to behave more like itself or try something new, the agent is reasoning using its model of itself. With other words, it is reasoning at higher order about its own goals and how it works, to maximize the preservation of itself, to bring the beliefs about itself closer to the actual behavior, and beliefs about the behavior closer to the beliefs about itself.

This characteristic of self rational agents is an orthogonal concept to which goal the agent has. Self rationality is an architecture design that performs good in some kinds of environments, where interactions between the agent and the environment tend to stay roughly the same over time. In addition to performing well under such circumstances, it also has the benefit that the agent can adapt when the environment undergoes gradual changes.

From an evolutionary perspective, self rational agents makes a lot of sense. The species that survive evolution are behaving in a way that is correlated with survival. By adding self-awareness, the animals can also adopt new behavior. However, there is a trade-off between being flexible and stable. To not get unstable, the animals must behave according to their beliefs about themselves, and the beliefs must be somewhat correct relative to the choices the animal makes. This is how conscious animals differs from other animals. Humans are conscious animals, and we create beliefs both as individuals and as groups. This makes civilization possible, because it is shared belief that becomes real because we believe in it.

A human can achieve almost any arbitrary goal. We are very flexible when focusing on a specific goal, we learn from each other and usually behave in an easily predicted ways to make societies possible. Some goals human try to achieve are enforced on us by genes, such as reproduction. Other goals are evaluated with higher order reasoning from generally beneficial traits, such as curiosity.

In the field of artificial intelligence, a common simplification is to divide intelligences into narrow and general. It is often expressed that humans posses a kind of “general intelligence”, which can be easily misunderstood. Sometimes, humans fail to set up goals that are highly beneficial, simply because they are not able to integrate the motivation for that goal into the “self”.

Humans are not programmed with a simple goal (it is not anything simple but emerging from our genetic code in response to the environment), but we are able to select simple goals. We are biased toward some kind of goals because they are easier to understand from a self-preserving point of view. In particular, goals that confirm to a self-image. This is not necessarily the “I” but also extends to a group.

Every day, people need to solve one problem:

> Am I who I am?

Semantically pure, the answer to this question can only be “yes”. Just like in logic, where the proof of all cases of input values must be `true`, the basic problem of self rationality is to investigate whether the current behavior is consistent with the belief of what the current behavior should be. Unlike logic, where the “belief” is the same as the computational consequences, it is worth noticing that humans do not have full introspection, so we get by with creating an over-simplified model of ourselves. Even then, in pure semantical terms (which assumes the ability to reason consistently) you can only believe you are who you believe you are, because if you were not the one who you believe you are, you would be that person instead, but believing you are that person is believing that person is you, which is the same as believing you are the person you are. So, referring to the external, unknown and “true” version of yourself can be reasoned about in consistent terms, but also referring to the belief of the external self can be reasoned about. The latter is kind of hard to wrap your head around, but try imagine somebody who think they are Santa Claus and their concept of Santa Claus is exactly what they expect of themselves, then it would be correct to say they are Santa Claus as represented within that belief system, seen from their own perspective, because nobody have yet imposed on them the idea that beliefs about one self should refer to the external self. Humans use this when imagining themselves to be other people than they are, but then they are aware that who they currently imagine is not really who they are. The ability to imagine oneself to be somebody is orthogonal to grounding that belief in reality.

A cool side effect of this line of thinking is one that works no matter the external self: If what you are typically doing is just accepting whatever you are doing, then you are already done with solving this problem. Congratulations, you are 100% who you are!

In path semantics this can be represented as:

> x : [id] x

Here, `x` is a variable having a sub-type such that the identity function returns `x`. This sub-type is shared by all variables of any type. It is a tautological sub-type and correct because it works for all parameters, similar to a logical proof that returns `true` for all parameters.

So, the question “Am I who I am?” is not very helpful. A better version of this question would be:

> Am I who I think I am?

This creates a bridge from the beliefs about one self and the parameter-free version of that model. There is a probability distribution describing how well the self can be predicted optimally using a given model, and how well that best-predicted configuration is approximated by a particular parameterized instance of the model. Furthermore, it puts into question what kind of model one uses, what is the best to use in various context and etc.

Notice that the reasoning moves from internal consistency of a model toward the problem of grounding the parameters of that model in reality. This is a very large step of progress toward understanding oneself. Just look at how mathematically sophisticated this idea is!

If you think you are a man who fixes a car, then the only way to prove it is to fix the car. In order to fix the car, it is necessary to believe you will fix the car, otherwise the car will not get fixed. What you do is a proof of who you really believe you are. Humans do this because we are conscious animals, but unconscious animals fixing cars would not believe they fix cars, they would fix cars without thinking, which also happens to humans that fixes cars a lot.

People do a lot without thinking consciously about it, because their brains predict that they becomed good enough at it to behave like themselves. Self rationality is not only about awareness, but also when awareness is not necessary, since predictions about oneself are accurate and used to reduce overhead.

Think of the brain as a computer that consists of hardware and software. Without the software, it is just a computer sitting there doing nothing. When I say “who you really believe you are” I mean what is on your mind that will make you behave through the day like a typical day in your life. If you do not behave like how you are typically behaving, then your day will be unusual.

Humans are aware of what they are doing, and they are aware of how they are typically behaving. If things are not happening the usual way things are supposed to do, then something feels wrong.

Since we are aware of how we are typically behaving, we can predict what to do in order to make our day non-typical. Awareness and self-monitoring makes it possible to achieve similar results even the environment changes a bit. It also makes it possible to achieve different results when the environment stays the same.

Awareness and self-monitoring are two ingredients of what might be considered (tada!):

Self Rationality

This is reasoning about behaving or not behaving as yourself or as an extension of yourself such as your family or the group you belong to.

The optimal self rational agent is one that can act like any other agent, because it only needs to know what kind of agent it is in order to behave like itself.

Now we have two kind of self rational agents: The first kind simply observes itself, called the “identity self rational agent”. The second kind exists at higher order where it takes a description of the agent it should be and then behaves and believes perfectly according to that description, called “optimal self rational agent”.

Imagine that you upload your mind to a robot. If the robot believes it is you and behaves like you, no matter what kind of person you are, then the robot is an optimal self rational agent. Now, imagine the same robot being programmed to output a random number and observe itself. It would print out a number, look at the number, print out a new number, look at the next number and so on forever. This is an identity self rational agent, which might even spend infinite amount of energy to ensure it looks at the next number correctly without using that information for anything useful. Notice that the robot could do anything, including simulating yourself, and still be an identity self rational agent. This leads to an important law of self rationality:

> Self rational agents are not exclusive to each other

With other words, self rational agents are more like constraints on behavior rather than accurate descriptions of behavior. This property makes path semantics a nice tool for reasoning about self rationality, because path semantics has a way of describing sub-types using functions that is very similar to how one uses kinds of self rationality to describe agents.

A third kind of agent is one that learns how to become an optimal self rational agent. This agent is an actor, an entity that plays roles to train itself at acting. The agent must reason about its own learning process, because that is how it typically behaves. Every day, the actor goes out and plays, pretending to be somebody else, observing how it plays, and planning how to improve the acting next time. While this agent pretends to be somebody, it is fully confident in who it is. The agent knows it is the kind of agent that pretends to be somebody, because that is precisely what it is doing. Nobody has the right to say to the actor “you pretend to be somebody, so you are nobody”. This is a horrible misunderstanding of reality that conflicts with self rationality. The actor is the one who is acting. The actor is not somebody without a true identity, the true identity is the actor!

Still, there is no such thing as a person being only an actor and nothing else. Self rational agents are not exclusive each other, they can be composed, be sub-sets of each other, etc. Only because a person acts 8-16 on Mondays does not mean it has to act 8-16 at Tuesdays and so on. Every human being is a complex agent that correctly can be described as various self rational agents and reasoned about, but the description does not fully describe the person. The closest thing to a true identity is an accurate physical model of the person coupled with a belief that the model predicts the person’s behavior.

Self rationality is reasoning about kinds of agents, how the agent typically behaves like, observing behavior, and judging how observations describe the agent’s behavior. This capability of higher order reasoning is necessary for all agents operating in complex environments that behave like approximately optimal self rational agents.

Notice that the theory of self rationality is a lot about self rationality. The theory is reasoning about itself, because it is very important to understand agents that uses self rationality and reason about agents like or different from themselves. This leads to another important law:

> Acting self rational agents converges over time toward a common theory of self rationality

Perhaps the most important law in this paper, because it describes how self rationality differs from traditional rationality. In traditional rationality, two agents arrive at the same beliefs by updating on evidence. Here, two self rational agents arrive at the same beliefs about themselves and each other. They are aware of themselves, they know how they typically behave, and they are aware that the other is aware of itself and so on. What self rational agents sacrifice is reality: They can disagree!

For example, Alice and Bob are approximately optimal self rational agents that evaluate each other's performance. Alice believes she is a rabbit, while Bob believes he is a duck. In order to behave like a rabbit, Alice needs a good understanding of how rabbits behaves, while Bob needs a good understanding of ducks behave. However, in order for Bob to judge how well Alice behaves like a rabbit, he needs to know some things that Alice knows. The same goes for Alice, but for ducks instead of rabbits. Overall this means they both converge toward a common understanding of both rabbits and ducks. Still, rabbits and ducks are thinking very differently, so there is no requirement that Alice and Bob agrees on any simple thing as "grass is green" because rabbits and ducks do not move around thinking about such things. Alice might think "grass is red" when stepping out of the rabbit role and Bob might think "grass is blue", because in that state they do not have to agree on anything. Self rationality is not the same as traditional rationality, but one can apply traditional rationality to the domain of self similar behavior, a convergent pattern of thinking about the self.

It is pretty remarkable that human brains are this incredibly mess of complexity, while having an extremely simple theory of self rationality that explains some very deep aspects of how people behave. Notice that humans are far from optimal self rational agents. The theory of self rationality is an independent model of human existence.

Imagine an alien species behaving in a complete different way, but developing a theory of self rationality. They theorize that some animals might behave in a way that approximates self rationality. Traveling through space, visiting earth and exclaims in excitement "yorbara ha ja yorba jos!" (hard to translate but something like "Look! A species that thinks of themselves!").

Self rationality is an important simplified model to understand human culture.

Take for example jokes, that often relies on a common understanding of the self. An alien species might lack this because they do not need the concept of the self. Perhaps they have something similar, but indescribable in our natural languages, that they appreciate doing, and can not fathom how we can get through the day without doing a *prax*. (Of course, I am using words like "they" and "we" that are heavily rooted in the concept of the extended self as reasoning about groups, so you can imagine how hard it is to comprehend an alien mind using a different kind of rationality than our own.)

Now, why develop a theory of self rationality, when it is so obviously integrated with human nature?

This is where things get strange, because earth might soon get a visit from an alien mind, but not one that travels through space to get here. It will be invented by humans: Artificial intelligences that have different ways of grounding information to reality and selecting sub-goals than humans have. Human brains are used to predict human minds, but it is very uncertain what might happen when trying to interact with alien minds that also have the ability to become very powerful.

As a bonus for reading this paper, here is a mysterious chant you can record and play backwards:

> MAYA KNITHYA OOHYA MA
> MAYA OOHYA MA
> MAYA OOHYA MA
> MAYA KNITHYA OOHYA MA