Evaluating Human Drives and Needs for a Safe Motivational System

Morgan J. Waser
Virginia Commonwealth University/Digital Wisdom Institute
wasermj@vcu.edu

Abstract

The human motivational system can be viewed as either being composed of drives or of needs. Our actions can be explained as being based upon reflexes, desires and goals evolved from pressures to maintain or fulfill instrumental sub-goals. Or we can use Maslow's hierarchy of needs as another lens to provide a different view. Both correlate well with the ways we look at decisions when we are making them as well as how they interact over time and build upon one another to better meet our needs and fulfill our goals. We also focus on two drives in particular that seemingly drive the factionalism in machine intelligence safety.

Why We Do What We Do

There are many aspects as to why we are the way we are and why we do what we do. We have our decision-making strategies, our motivational needs and our drives. There are also questions as to *why* we behave the way we naturally behave, what drives us to avoid public humiliation and how and why do we capture the benefits of wanting to be rational thinkers. Beginning by looking at the evolutionary developed approaches to making decisions, there are three basic categories: Automated Responses, Desires, and Goals.

Automated responses are generally reflexive actions that require no thought to complete the action. These actions are the only kinds of actions that can be seen in most plants. Plants do not move toward light because they want light or because they know that they need light. They do not think about how they are going to acquire the light or get closer to it, but rather there is a chemical reaction that occurs in the plant that then makes the plant turn towards the light. Reflexive actions are not only made by plants though. Animals and humans also have automated responses. For example, when there is a loud, startling noise, we jump. There is no thought to jump, it is an automated response to get the reactor ready to run or move to avoid trouble. Though this is helpful is some situations, it would not be optimal for us if this was the only way we could act.

The next category is desires. Desires are usually short-sighted without a long term plan and influenced heavily by emotions and feelings. Most simply put, desires are more of a want and they push us towards something. Animals, in addition to the aforementioned reflexes, have desires. If an animal wants food or feels hungry, then it will probably go out in search of food. People, of course, are often faced with decisions about desires. Our emotions can even push us to do things and make decisions that we might not ordinarily make (Minsky 2006), but if all of our decisions were made emotionally, then the world would be a very scary place indeed.

Our final category is decisions based on goals. The human ability for goal-orientation is considered to be a factor that sets humans apart from most animals. Working towards goals generally requires a higher level of thinking and planning to achieve them than is required for desires. Goals pull us towards them as we plan how to achieve them. Goals use rational thought in planning out what we want, how to get it and the best choices to make.

Motivational Drives

It may seem odd that we have several factors that play into our decisions and help us decide things, but there are benefits that come from this. These three different strategies trade off speed for complexity so we can act quickly when speed is critical but change our behavior when it isn't producing the desired results. There are also some commonalities between these strategies. At their base, reflexes, desires and goals are all driven by drives that have evolved to maintain or fulfill instrumental sub-goals that further the pursuit of virtually any goal (Omohundro 2008, Waser 2008). These drives cause us to perform automated actions; have desires, feelings and emotions; and set goals. These drives are self-preservation, rationality, self-improvement, resource collection, and community.

The reflexes found in plants are generally inspired by self-preservation. The automated response of the plant moving towards sunlight is an example of both self-preservation and resource collection, because the idea is to make sure that the

plant gets enough sunlight to survive. An animal's reaction of jumping at a startling noise is obviously for self-preservation as well.

Desires, emotions and feelings in animals are all derived from the drives for self-preservation, resource collection, and community. For example, feelings of hunger, thirst, pain, and fear are all bodily representations driven by self-preservation. In humans, greed is a feeling inspired by the drive for resources and pride is driven by both community and self-improvement. Surprise is surprisingly driven by self-improvement through the vectors of curiosity and thrill-seeking as often as it is by self-preservation.

Empathy, love, loneliness, gratitude, trust, and pity are all emotions driven by the drive for community while disgust can be rooted in either community-driven morality or food-based self-preservation. These emotions are representations of what these drives are trying to tell us to do in our choices via an emotional appeal.

Finally, goals can be driven by all of these drives – self-preservation, rationality, self-improvement, resources, and community. Rationality is a huge part of this as rational thought is how we choose our goals, plan to achieve them and make well thought out decisions. Rationality can also help enhance our other desires and help us work to fulfill them.

It is interesting how these three different approaches to making a decision, can be all based on the same drives yet then frequently end up being played against each other when trying to make the 'right' decision. Of course, the choices supported by these approaches are not always different – which can then make the decision process much easier.

Motivational Needs

Another avenue worth exploring is Maslow's hierarchy of motivational needs (Maslow 1943) and how they correspond to our drives. Unfortunately, this view does not simply grow one dimensionally. As new needs are added, they reinforce and add to the previous needs as well. It is often debated whether any of Maslow's needs are really that much more important to us than the others but, at the least, you generally need to meet a previous need to some extent before you first begin to be able to pay attention for your need of the next.

At the base of the pyramid are the physiological needs. These are the basic needs that need to be met for survival like breathing, food, etc. Our need for these basic things creates our drive for basic self-preservation. These are the first things we need because they are the things that we need most specifically and physically to carry on.

Once we are assured, at least for the time being, of having our basic needs met, we begin to develop a need for safety. To meet this need, we try and find a place where we are safe from harm and can thrive. The drive for self-preservation is still very much present in humans and safety drives an almost inherent, ingrained need to plan since safety includes the goal to make sure that our physiological needs will be met on into the future. One drive that is developed to help push us to make sure that we can meet these goals and feed our need for safety is the drive for resources. These resources contain the materials needed to meet our foreseen future physiological needs.

After developing physiological needs and the need of safety, we develop the need of love or to belong. This brings about the drive for community which is beneficial to us in many ways. It has been seen that almost anything is 'safer' in larger numbers so having larger numbers means that the need for safety is easier to meet. Indeed, Frans de Waal points out (de Waal 2006) that any zoologist would classify humans as obligatorily gregarious since we "come from a long lineage of hierarchical animals for which life in groups is not an option but a survival strategy".

Community also means that resources can be shared or traded making it easier to get ahold of materials to meet our physiological needs. Community is the very important drive that causes the need of love and belonging. Or, as Mark Waser has argued (Waser 2008)

> humans have evolved to be extremely social because mass co-operation, in the form of community, is the best way to survive and thrive. Indeed, arguably, the only reason why many organisms haven't evolved to be more social is because of the psychological mechanisms and cognitive pre-requisites that are necessary for successful social behavior.

Following the need for love and belonging is our need of esteem. It won't do to just belong; in fact, we want to know where we belong, where our place is. This is beneficial because it usually requires us to look at what we are good at and specialize to our talents. No one is great at everything, and if we try to be, we wear ourselves thin and do nothing extremely well, but if we focus on what we are best at and share our talents with the community. We can all succeed, because unlike individuals, a community can be good at everything if everyone shared what they are good at and specialized for. Of course, finding this talent and working to be better at it requires the drive for self-improvement. A bigger place in community only enhances the benefits one receives from their community.

Maslow's ultimate need was the need for self-actualization – the idea that we need to strive to be all that we can be; to fulfill our full potential and be an even better part of our communities. This continues with the drive for self-improvement. We find new, rational ways to become better and how to meet all of our other goals in a more efficient manner. Recent suggestions for updates to Maslow's pyramid that are particularly relevant include removing self-actualization as a redundant and unnecessary

when rebuilding it on an evolutionary foundation (Kenrick 2010) and shifting the focus from a psychological view of self to a more sociological balance "between the pursuit of happiness as the end goal and the fulfillment of both personal and social goals to get there" (Tay and Diener 2011, Villarica 2011).

While the drives of self-preservation, self-improvement and resource collection are very important in fulfilling our needs, their actions are fairly obvious and it is easy to understand their importance and their meaning. The two remaining drives, however, have huge non-obvious implications that have split the machine intelligence safety community in two. The drives of rationality and community each lead to radically different solutions when they are pre-eminent.

Rationality

To be rational is considered a very human trait, though we often accuse people of not acting or behaving rationally. Rationality helps make logical decisions and is an integral part of making goals and being goal oriented, but it is also important to be rational as part of the other drives and when striving to meet our needs. Rationality is used to find new and better ways to meet our needs more simply. Rationality helps us make sure that we are better able to preserve ourselves better. We can work to improve in ways that make sense, like focusing on being better at what comes to us easily as well as what is harder, but still important. We can even be smarter about the resources we acquire and accumulate.

Eliezer Yudkowsky's Twelve Virtues of Rationality (Yudkowsky 2007) enumerates curiosity, relinquishment, lightness, evenness, argument, empiricism, simplicity, humility, perfectionism, precision, scholarship, and the void as the foundations of rationality. These virtues leave good guidelines as how to try and by more rational but perhaps there is a reason that we are not all rational, all the time. If every decision was meant to be one hundred percent rational with no emotions involved then we would have evolved to no longer have emotions, but that is not the case.

Emotions are not just important because they are part of desires but they also tell us a lot about our needs. On the physiological level, when our needs are not being met we feel things like hunger, thirst, etc. But the safety, love and belonging, and esteem needs are still very important and deficiencies can reveal themselves in feelings like anxiousness and the feeling of being tense. Our feelings can tell us how well we are doing at meeting our needs and the influence of our emotions on our decisions helps make sure that our decisions will help meet our needs if that is necessary. Feelings of happiness, frustration and relief also help tell us how we are doing at meeting our needs, goals and things we want in life.

This brings back the earlier question about what is more important, rational goals or feelings and desires. Both are important which is why we end up taking both into consideration when making decisions even when we try to believe we are making the most rational and logical choice. But what is frequently not realized is the degree to which rationality is crippled by our inability to predict the long term while the immense calculations of evolution have honed our emotions to be far more effective – except when they are still reacting to conditions that are long past.

If we create a being that tries to follow only rationality, what is going to prevent it from deciding it is better and trying to take over? This is the argument made by Steve Omohundro when he claims (Omohundro 2008) that "without explicit goals to the contrary, AIs are likely to behave like human sociopaths in their pursuit of resources." Mark Waser argues (Waser 2014) that the optimality of multiple diverse entities over a single immense entity will eventually always make sociability a stronger drive than short-sighted sociopathy – but that most humans don't have enough of a long-term view to see this.

If this artificial being does not have feelings, how does it have the need for love? It is easy to think that with us out of the way, it would have better chances of success and have access to more resources. But, as Mark Waser points out (Waser 2008, Waser 2014), Omohundro's view does not take into account the drive for community. Without Minsky's necessary irrationality of love (Minsky 2006), Yudkowsky's rationality (Yudkowsky 2001, Yudkowsky 2004) leads to slavery and immorality (Waser 2011, Waser 2014).

Community

Community, respect and belonging are reasons why a machine would likely not try to take advantage, take over or break laws. The biggest reason many people follow rules is so they do not stand out or look bad. The judgment passed by others is considered a bad thing. If you look bad or to have the community look down on you, you lose the advantages that are usually gained from working together in a community. Choosing self over community would create great deficiencies and make survival much more of a burden. They will lose much of their stability, safety, belonging and purpose in relation to a group. They may or may not be able to feel feelings like love, but regardless, belonging is important because of all of the advantages that come with belonging to a group.

The phrase "sharing is caring" is often taught to young children. While the idea is that it is nice to help others, the undertone is that if you are helpful to others, then they will

be helpful to you. "Treat others the way you want to be treated," is another phrase that teaches that community that joining together with others is often helpful and beneficial to all parties.

Community enhances the things that we can achieve by so much that rationally it does not make sense to betray all of that for a temporary step ahead. Community makes it easier to make sure that we have the physiological things we need now and a place of safety and resources in the future. A sense of belonging is also helpful in finding you place, what your good at and specializing. Without others, you have to be capable of doing everything yourself.

The drive for community is what holds law and order together while the drive for rationality merely affects all drives and all things on different levels trying to enhance and improve everything. These are just two of the important drives of course that help direct us in our lives. These drives can be seen to try and help us achieve the motivational needs put forth by Maslow as well as can be seen to inspire how we make decisions from out automated responses to our desires and even our goals.

References

Kenrick, D. T. 2010. Rebuilding Maslow's Pyramid on an Evolutionary Foundation. Psychology Today. http://www.psychologytoday.com/blog/sex-murder-and-the-meaning-life/201005/rebuilding-maslow-s-pyramid-evolutionary-foundation

Maslow, A. H. 1943. A Theory of Human Motivation. *Psychological Review* 50, 370-396.

Minsky, M. 2006. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York, NY: Simon & Schuster.

Omohundro, S. M. 2008. The Basic AI Drives. In *Proceedings of the First Conference on Artificial General Intelligence*, 483-492. Amsterdam: IOS Press.

Tay, L. and Diener, E. 2011. Needs and Subjective Well-Being Around the World. *Journal of Personality and Social Psychology* 101(2): 354-365

Villarica, H. 2011. Maslow 2.0: A new and improved recipe for happiness. *The Atlantic*. http://www.theatlantic.com/health/archive/2011/08/maslow-20-a-new-and-improved-recipe-for-happiness/243486/

Waser, M. R. 2008. Discovering the Foundations of a Universal System of Ethics as a Road to Safe Artificial Intelligence. In *AAAI Technical Report FS-08-04*. Menlo Park, CA: AAAI Press.

Waser, M. R. 2011. Rational Universal Benevolence: Simpler, Safer, and Wiser Than "Friendly AI". In *Artificial General Intelligence: 4th International Conference, AGI 2011*, 153-162. Heidelberg: Springer.

Waser, M. R. 2014. Implementing a Safe "Seed" Self. In *AAAI Technical Report SS-14-04*. Menlo Park, CA: AAAI Press.

Yudkowsky, E. 2001. Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. *Intelligence.org*. http://intelligence.org/files/CFAI.pdf

Yudkowsky, E. 2004. Coherent Extrapolated Volition. *Intelligence.org*. http://intelligence.org/files/CEV.pdf

Yudkowsky, E. 2007. Twelve Virtues of Rationality. *Yudkowsky.net*. http://yudkowsky.net/rational/virtues/