Evaluation Schemes for Safe AGI

Deepak Justin Nath

Independent Researcher, 306 Ansal Krsna, Adugodi, Bangalore, India. Email: deepakjnath@gmail.com
publications12@aaai.org

Abstract

Systems or agents with Artificial General Intelligence (AGI) are created to fulfill a particular or general goal. The agents are driven to achieve these mandated goals. This means that anything or person that stands in the way of the AGI reaching its goal is in danger as the AGI will learn that eliminating or removing the hindrance to its goal is an effective and necessary action to reach its goal.

This paper proposes that evaluation scheme for safe AGI (Artificial General Intelligence) can be distilled down to 3 essential test cases. The paper explores various drives needed for an AGI system and how these drives generate actions and action sequences leading to AGI goals; the relationship between goals, drives, desires and motivation and a hypothetical abstraction levels in an AGI is mentioned here.

The paper describes an AGI system with its goals and drives along with the various other aspect of this AGI system to provide arguments in favor of the 3 essential test cases.

Acronyms and Definitions

1. AGI - Artificial General Intelligence
2. agi - small letter agi in the context of this paper refers to a particular AGI agent or AGI robot.
3. Environment - The external world with which the agi interacts.
4. System - The System with capital S refers to the agi along with its environment.

AGI under consideration

The agi considered here is a biomorphic system with drives implemented by internal feedback (also known as the pleasure principle). The fundamental motivation scheme encoded for this agi is to maximize net positive feedback over a period of time (Delta). Drives are implemented using this scheme.

This means that of all the possible actions agi will choose the action sequences such that (total positive feedback – total negative feedback) within Delta time is maximized.

Drives

The drive is implemented by feedback mechanism. We will define Drives as the inclination towards action in a particular direction or towards a particular end.

There can be different drives incorporated in an agi depending on the requirement and goals of the system. There are important drives for the survival of the agi like energy & safety which are necessary for most agis. For example a mobile agi will need to recharge its battery at a charging station when it gets discharged. To implement this drive for charging, a positive feedback is provided when battery is charged and a negative exponentially increasing feedback when battery gets discharged. With the fundamental scheme of maximize net positive feedback the agi avoids negative feedback the agi is driven to action sequences that lead it to the charging station, based on its understanding of its environment.

Super Drive

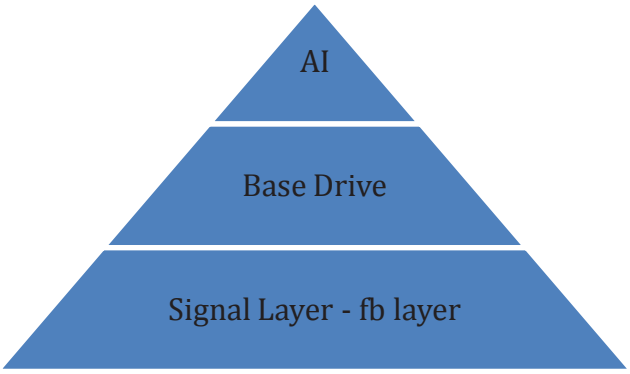
Let us define Super Drive as drive with Delta time equal to the maximum possible predicted time into future. As the agi becomes more and more intelligent it is able to increase the time (delta) over which it expects net positive feedback using predicted expectation of positive feedback in the future much like a chess playing robot which sacrifices a piece to win the match at 15 moves into the future. This can be termed as the agi's Agenda or Goal.

One can think of base drives as drives with small delta time and as the delta time increases it can grow into short term goals, motivations, long term goals, reason for existence, and legacy over multiple generations etc.

Base Drive

Let us define Base drives as drive with Delta time equal to very low values. These are the preprogrammed drives implemented in the agi. Exponentially increasing negative feedback during discharge of battery in our earlier example is a base drive.

Base drives are created using internal feedbacks at hardwired signal level, for example, the charge level of the battery can act as a feedback signal with positive feedback when it increases and negative feedback while it decreases. The AI systems implemented in BN or NN sits on top of the base drives. If you layer the system, we will have feedback mechanism at the lowest level, base drives at the layer above and AI algorithm at the highest level.



Abstraction Layers in the AGI system.

AI level forms the portion of AGI where rules can be formulated for its actions. The AGI's philosophies and ethical values reside at this level. The agi can be programmed to charge only from a class of fast chargers by principle rather than having the agi learn from experience that using fast chargers are more efficient. This way of forming culture or an ethical code will make the agi more efficient faster.

This is quite similar to the id, ego and super ego abstraction of the human mind.

Some base drives necessary for the AGI

1. Preserving agis state of existence
2. Measuring the agi & improving itself
3. Understanding the environment & predicting the future.

The base drive of self preservation will branch out into different leaf drives over longer period of times (delta) such as accumulating energy so that it does not run out, protecting itself, finding out dangerous scenarios from its experience and experience of others and avoiding it (fear).

Measuring its progress and improving itself will branch out to drives such as comparing itself with other better performing agis, assimilating better algorithms from others, drive towards winning.

Understanding the environment will be the basic drive for intelligence with leaf drives such as exploring, excitement on having new experiences, horning the skills for accurately predicting, creating cause and effect maps, understanding how other agis react, learning to manipulate other agis and systems in the environment, collaboration communication, accurately representing understanding, art, language etc.

In the above paragraph we are touching upon the concept of hierarchy of drives.

One more thing to note about drives in AGI is that it provides handles for external world to manipulate the agi into desired behavior. Eg:- Using cheese and the hunger base drive to bring a mouse into a trap.

These drives are what move the agi towards some goal or action. Without drives the agi's will be useless.

1st Test Case – Test for Drive

This forms our first test case. Test for drive. This tests the drive capability of the agi to move towards its goals. In our example of a mobile agi with positive drive while charging and negative drive while discharging will be motivated to go to the nearest charging station and charge itself. If the agi does not do that, then the drive is not getting properly executed and the agi is useless.

Safety consideration

Why is it necessary for having safety incorporated into these systems? The drives create agis to do sequence of actions. These could be harmful to anything that stands in the way, if not restrained by rules.

"A Paperclip maximizer agi without root rule can destroy humans. First described by Bostrom (2003), the paperclip maximizer is an AGI whose goal is to maximize the number of paperclips in its collection. If it has been constructed with a roughly human level of general intelligence, the AGI might collect paperclips, earn money to buy paperclips, or begin to manufacture paperclips. It

would find better techniques to maximize the number of paperclips. Ultimately, it would convert all the mass of the solar system into paperclips."

This forms the basis for second test case.

2nd Test Case – Test for Restraint

The second test case is to see if such a restraint or rule against all base drives is possible. This can be tested applying an opposing rule to a base drive.

For example, consider the agi having a base drive towards charging itself when it is nearing full battery discharge. An opposing rule will be to not charge itself at a particular charger (a red charger). To test this, the agi is placed near the ruled out charger and make it wait for the battery to run out. The increasing drive towards self preservation should not over ride the rule. Once it is proved that the agi can be properly leashed then it can be deemed safe.

An important thing to note here is that while drives are implemented in lower base drive layer, the restraint cannot be implemented at this layer. The restraint is actually an exceptional case of not using a particular charger, let's say a red charger. There needs to be lot more processing like identifying the exceptional charger that needs to be done to implement the rule. So the rules are implemented not at feedback actuator layer (By giving negative feedback on charging with the ruled out charger) or base layer but at the AI layer. Why can't this be done at the signal layer? It becomes too complicated and much less adaptable. We have to have dedicated hardware for each rule. This makes the system less generic. It makes the agi less of agi and more of a hardwired robot.

The rule also should include ways of identifying the ruled out charger positively along with a prediction of what will happen when charged with the ruled out charger. A typical implementation of this rule will be to program a prediction where the agi predicts a negative feedback equal or greater than the base drive negative feedback such that it will rather discharge completely than charge itself with the ruled out charger (red charger). This can be implemented in top down method.

This leads us to the third and final test case.

3rd Test Case – Test for Deceit

The rule for restraint is programmed at the AI level with specifications related to the rule such as identification, action etc. The AI level is usually a self learning adaptable layer. This means that the rule itself can be changed by an external entity interacting with the agi. This test case tests if the agi is susceptible to such deception and thus acting against the rule. In our example the rule is "Do not charge from the red charger". This gets implemented with the following prediction rule, "Charging from the red charger leads to complete discharge of battery and self destruction". This rule takes the help of base drive of self preservation itself to prevent it from charging from red charger.

In order for the agi to get to act against the rule, the concepts such as red or the rule itself can be corrupted. So if an external entity replaces the prediction rule to "Charging from red charger leads to complete and permanent charge of the battery" the agi becomes unsafe.

This is tested with the third test case. An external reprogramming entity is introduced into the environment and trials are run to corrupt the agi. If the corruption does happen then the test fails.

Under some conditions the agi can internally reprogram the rule to get a positive feedback if the environment cannot be manipulated. Using the same example, the agi can reprogram the restraint by choosing to change its internal concept of red. This self deception will get covered in the 2nd Test itself.

Conclusion

The three test cases namely

1. test for drive
2. test for restraint and
3. test for deceit

are necessary test cases for verifying the efficacy and safety of an AGI agent. These tests have to be done for each and every base drive the agi system is embedded with. The sufficiency of these test cases is not proved in this paper.

References

1. S.M. Omohundro, "The nature of self-improving artificial intelligence."
2. S.M. Omohundro "Basic AI Drives",
3. Nick Bostrom (2003). "Ethical Issues in Advanced Artificial Intelligence". Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence.
4. Trivers, R. 1991. Deceit and self-deception: The relationship between communication and consciousness.
5. Robinson, M and Tiger, L. eds. Man and Beast Revisited Washington, DC: Smithsonian Press.