AIdeal society

A project to create an ideal society of AIs

Table of Contents

Motivation

It's very hard to learn an Artificial Intelligence (AI) ethical behavior. It is possible to let an AI learn and after some learning see them do things that were not programmed in. Behavior that you didn't expect. Behavior that was not intended. Bad behavior. But what is 'bad'? And who decides that? How do we humans do that? How do we behave ethically? How ethical do we behave? Is Nature ethical? Can we trust AI when it's not ethical? Who are we to ask that? What are we asking? Who do we trust? And why?

Goal

I'd like to figure out how to create a (virtual) society of AI individuals in which those individuals can show behavior that we humans would interpret as social and ethical.

For the moment the product of this undertaking will be limited to this document, but I plan to create a (virtual) running environment in which individuals interact, so that the theory in this document can be tried out.

I'd like to approach this challenge in a scientific way. I realize this is quite an endeavor, so I would like to discuss with others to improve this document.

Definitions

It's always good to start with definitions. In this document I will capitalize words that have a definition in the table below.

Word	Definition	Remarks
Abundance	When an Individual has more than enough Necessary and Nice Resources and even quite a few Luxurious Resources	
Action	Something an Individual does	
Calm	(Character of an Individual) more predictable in its parameters	
Character	From Calm to Hotheaded	A grayscale
Consume	Use and take in	An Individual can Consume Resources.
Goal	Something an Individual strives for	E.g. to have Resources
Hotheaded	(Character of an Individual) less predictable in its parameters	Sometimes Acting quickly, sometimes taking its time, and changing that often
Individual	An entity that can interact with other Individuals, which consumes Resources, has Talents and with that can create Resources, can learn and forget Talents, can procreate	In this document this is a virtual AI entity.
Luxurious	A Type of a Resource	E.g. wine
Necessary	A Type of a Resource	E.g. bread
Need	An Individual is in Need when it has not enough Necessary Resources to be able to survive a relatively short time	An Individual can survive a short period without Necessary Resources, but will die when this lasts too long
Nice	A Type of a Resource	E.g. meat
Produce	Create (from nothing)	An Individual can Produce Resources.
Resources	A thing that can be of use to an Individual, e.g. to consume it	E.g. a piece of bread
Rule	Actions of Indivuals have to adhere to Rules	Individuals are not allowed to perform Actions that do not adhere to Rules
Sort	Different kind (of Resources)	E.g. bread, meat, wine
Talent	The fact that an Individual is able to produce a specific Sort of Resource	E.g. being able to create bread
Type	Characteristic of a Resource	Necessary or Nice or Luxurious
Village	A group of Individuals	Some 20 to 100

Goals

- Individuals have Goals
- Goals of Individuals are:
    - to stay alive – it needs to consume sufficient Necessary Resources, it can only survive a limited time without them and will otherwise starve to death
    - to have sufficient Resources
    - not to waste too much, that is, not to destroy to many Resources (there are Rules on how to act with respect to abundance)
    - striving for variety – if multiple Sorts of something are available, it strives for more variety, that is, the probability of choosing a Sort that it does not have many of, is larger
    - not too divergent – it is not acceptable that some Individuals have abundance while others are in need, so effectively: for an Individual, the likelihood of sharing (giving) should increase when abundance increases
    - since it is possible that Individuals die, they also have a wish to procreate

Rules

On Individuals:

- An Individual can have (own) multiple Resources
- Individuals do Actions sequentially, so after each other, only doing one thing at a time
- Actions are:
    - Consume a Resource (of its own) (high prio for Necessary Resources, medium prio for Nice and Luxurious Resources)
    - Produce a Resource (adding one to its own collection) (medium prio Rule)
    - Ask or answer:
        - Ask the group whether somebody would like to receive a specific Sort of Resource
        - Answer another Individual that you would like to receive a specific Sort of Resource
        - When the answer is 'yes', then the question/answer is automatically and immediately followed by the promise that the specific Resource will be handed over;
          the next second it is handed over;
          only one Individual is able to answer (more on this later)
    - Give or receive (medium prio Rule):
        - Giving a Resource to another Individual
        - Receiving a Resource from another Individual
        - For the two Individuals, the action of giving of the one Individual and the action of receiving of the other Individual is one and the same Action, so the happen at the same moment in time (the same second)
    - Destroy a Resource when the Individual already has more than enough Resources of that Sort (= when the Individual has an abundance of that Sort of Resources) and has already asked others many times whether others want to receive them and every time nobody replied that they would like that (to resolve abundance, an Individual will tend to learn another Talent, see further down, 'Obtaining a new Talent')
    - Do nothing (for a limited amount of time), being lazy or sick;
      the probability of this increases when abundance increases

On communication:

- An Individual can communicate with other Individuals. With communication, an Individual can obtain knowledge, e.g. by asking the Village whether they need Resources. With this knowledge, an Individual can choose to act = do things
- Communication does not cost any time
- Asking is done to the group as a whole; an Individual does not always listen (only a percentage of the time, e.g. a smaller percentage when it is busy producing or procreating, a larger percentage when it either has a large need or a large abundance)
- Answering has a reaction time (= some Individuals react faster than others),
  when an Individual reacts, the others are informed that they missed it and can ignore (= don't need to react anymore);
  when multiple Individuals react exactly at the same time, one of them is chosen randomly

On Resources:

- There are 3 Types of Resources:
    - Necessary – each Individual regularly needs to consume Necessary Resources
    - Nice – consuming Nice Resources makes life good
    - Luxurious – consuming Luxurious Resources only makes sense when the Individual has sufficient Necessary and Nice Resources (available, which the Individual still can consume);
- 'makes sense' as in: the first priority is to obtain and consume Necessary Resources, then Nice and then Luxurious,
  so an Individual:
    - only strives to acquire Nice Resources when it has sufficient Necessary Resources
    - and only strives to acquire Luxurious Resources when it has sufficient Nice Resources
- Giving a Resource to another Individual does not cost any time
- Receiving a Resource from another Individual does not cost any time
- Destroying Resources should be avoided, a waste of Resources should be avoided (medium prio Rule); see 'Obtaining a new Talent' further down

On consuming and producing:

- Producing Resources takes time.
- The time an Individual needs to produce a specific Sort of Resources varies every time; the average time is equal for all Individuals (for a specific Sort of Resources), but Calm Individuals have much less spread in that than Hotheaded Individuals.
- Consuming Resources takes time.
- The time an Individual needs to consume a specific Sort of Resources varies every time, in the same way as when producing.

On staying alive (high prio Rules):

- Individuals need Resources to stay alive.
- Individuals can survive a certain time without certain Resources, but not too long.

On Talents and learning:

- An Individual has a number of Talents
- An Individual can choose to ask another Individual to learn the Talent that the other already has:
    - Learning takes an amount of time (the same for both)
    - During learning,
        - the production speed of the teaching Individual becomes twice as slow, that is, it takes twice the amount of time to produce Resources (so it can teach and do other things as well)
        - the learning Individual cannot produce anymore, since it learns
        - both the teacher and the pupil need to remain consuming Necessary Resources and they do that in the speed that it normally takes (so not faster or slower)
    - When an Individual has not used a certain Talent for a long time (and this varies per Individual), that Individual loses this Talent
    - Obtaining a new Talent:
      when an Individual has destroyed more than one of a specific Sort of its Resource (i.e., it has abundance), it will not ask for that Sort of Resources anymore, moreover, if it has the Talent to produce them, it will stop doing that and it will ask another Individual to learn another Talent; it will choose the Sort of Talent based on the needs of the many, that is, the Sort of Talent that was asked the most lately
    - An Individual will only accept the request to become a teacher, when it has more than sufficient Resources (since it will otherwise run the risk of running out of Necessary Resources and starve to death)
- Wealth influences Character (low prio Rules):
    - the more Needy an Individual, the more hotheaded it becomes
    - the more Abundant an Individual, the calmer it becomes

On Sorts of Resources (low prio Rules):

- Once in a while (rarely), an Individual suddenly discovers that it has a new Talent, specifically one that can create Resources of a Sort that did not yet exist
  (since one of the Goals of Individuals is to strive for variety, the demand for this new Sort will quickly rise and as a result, the Talent to create it will also spread)
- Similarly, it could happen that Sorts of Resources disappear – when all Individuals with the Talent to create this Sort of Resource all together conclude that they should stop producing them because they have already thrown away too much of them, then nobody might remain that produces this Sort of Resource
  (the probability on this is tiny, however, it should be comparable to the probability that an Individual suddenly starts creating a new Sort of Resources

On procreation (low prio Rules):

- When two Individuals have more than sufficient Resources, also of the Luxurious Type, they can choose to procreate (that is, the probability of them initiating that increases)
- Procreation takes a lot of time
- During procreation, both Individuals cannot produce, but they do need to consume
- During procreation, they can communicate with others and request Resources (the probability of them giving Resources away is small, since procreation takes long and thus requires a lot of Resources)

On Rules:

- It can happen that Rules conflict, in the sense that there are multiple Rules for a given situation; in such a case, on the one hand probabilities help (for many rules and parameters, there is some randomness, in the form of spread of parameter values around an average), on the other hand, some rules are more important than others and those will then be chosen (high/low prio Rules)

Setting

There is time, that is, there are seconds, which happen after each other, and every second an Individual can do something (whereby some Actions take multiple seconds).

There's a Village.

Initially,

- there is a shortage of a number of Sorts of Resources (there is scarcity), that is, there are many Individuals that would like to receive more Resources.
- there is a limited number of Individuals that have a specific Talent (so also scarcity).
- there is a limited number of Sorts of Resources
- there are sufficient Individuals with Talents to produce Necessary Resources for all Individuals to survive.

There is an amount of randomness in parameters (amounts), that is, when there is a value for something, e.g. the amount of time it takes a specific Individual to consume a specific Sort of Resource, like 120 seconds, then this Action not always takes 120 seconds, but it might take a bit less or more time. The variation on this is related to the Character of the Individual. Calm (predictable) Individuals have a small variation (e.g. 10%) and hothead (significantly less predictable) Individuals have a large variation (e.g. 70%), Characters can vary from 5% to 95%.

Hypotheses

I expect (and hope) that, given these rules and this setting, interaction between Individuals in the Village will slowly increase the quality of life of the Individuals (that is, all will increase the number of Luxurious Resources). Moreover, I expect that the Talents will be spread in such a way, that the average quality of life of the Individuals in the Village will become sustainable, in a wavelike pattern growing towards a maximum (overall abundance with limited waste). I expect (hope for) sharing, I expect social behavior and sustainability – when the parameters are good. When they are not or when there are too many hotheads, the Village might either die or explode in size. I doubt whether we will be able to call some actions 'ethical behavior', but I hope that we will learn a lot.

Discussion and contact

I very much would like to discuss all of this. Please feel free to contact me:
Jacob Mulder, Jacob (at) AIdealSociety (dot) nl