AGI: CAUSALITY

Mykola Rabchevskiy

This chapter will continue our discussion of the "inner workings" of decision making in systems dealing with the natural physical environment, which we began in chapter DECISION-MAKING FUNDAMENTALS.

Since we are talking about ***General*** AI, the reasoning should operate in pretty general terms. The process of functioning of an intelligent system in a natural environment looks in general as a ***sequence of events separated by intervals of smooth changes in all parameters*** that characterize the environment and the state of the system itself in the representation of an intelligent system (the current ***model of the situation***). At the time of the event, smoothness is ***disrupted***. The reason may be:

- A ***signal*** from a specialized sensor about a particular event (the light went out, the robot stumbled upon something).
- Changing the ***structure of the situation model***: a new object is ***detected*** in the environment; the tracked object is ***classified*** in a different way than before; tracked object ***ceases to be visible.***
- ***Start or end*** of an ***action*** initiated by the system (the command to retract the landing gear is given; the chassis is locked in the retracted position)

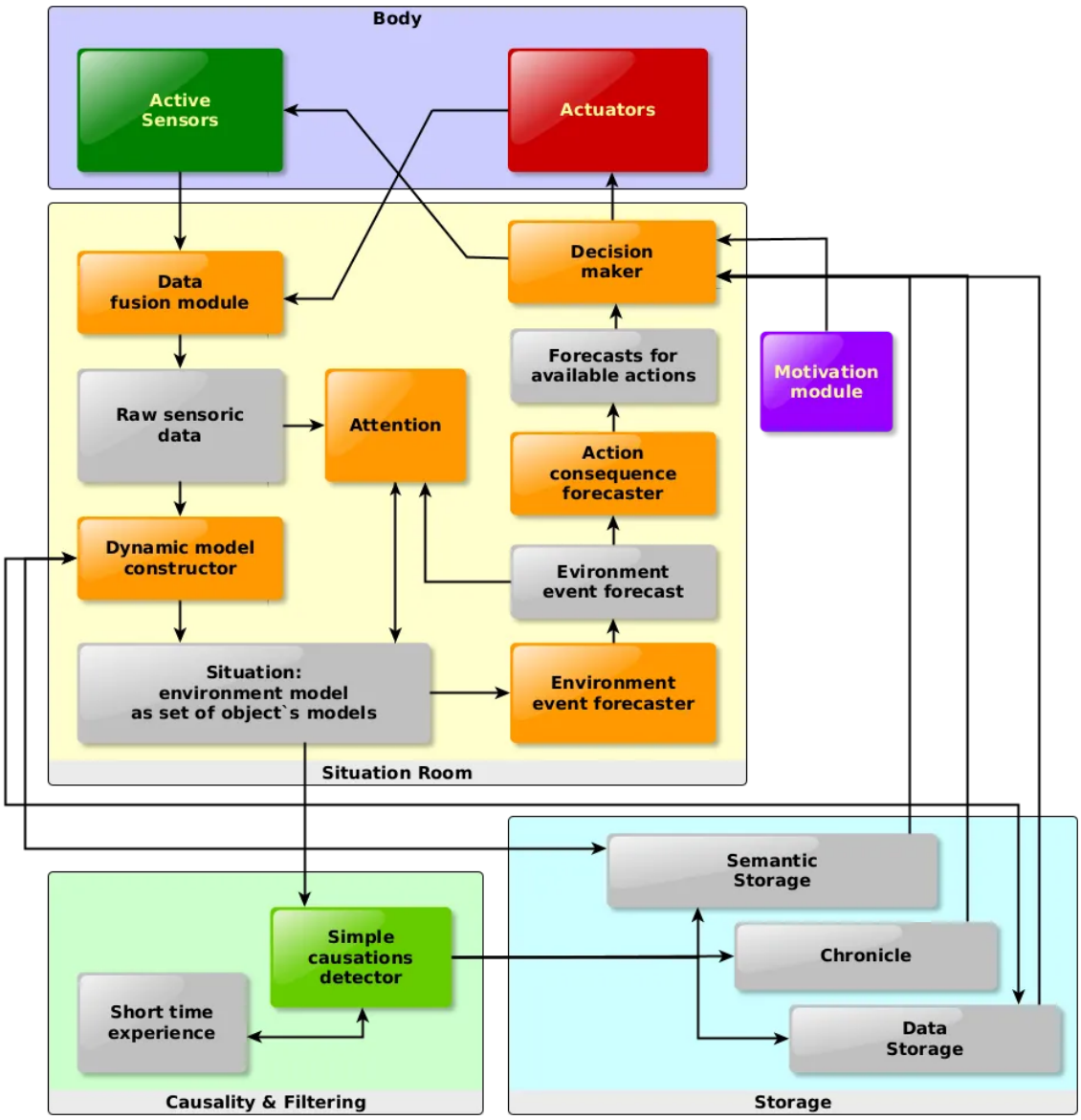
The occurrence of an event entails modifying the structure of the model (description) of the situation. The sensor data is used to update the parameters with the same structure in the interval between events.

Since decision-making is based on a forecast, it is necessary to forecast both ***changes in the parameters*** of the current model of the situation and ***forecast those events that are caused by changes in parameters***. For example, the distance between an object in space and a robot changes; by predicting the change in distance as a function of time, it is possible to predict a collision (including an ***estimate of the time*** of a potential collision).

To implement such a forecast, it is required to know a ***distance versus time*** and the ***cause-and-effect relationship between a change in a parameter (distance) and a consequence (collision)***. The parameter value versus time can be constructed as an ***approximation*** of the function using a set of measured values (with refinement as the data accumulates); ***extrapolation*** gives an appropriate ***prediction***. It is essential that the ***approximation can be applied to any quantitative parameter***, regardless of the type of

sensors, mission, and so on - ***complete generality takes place***. On the contrary, the second requirement, ***knowing the causal relationship between the values of continuous parameters and discrete events***, obviously ***depends on the substance of the parameter and the meaning of the event***. Corresponding causal relationships can be predetermined, laid down initially following the mission of the system. However, the requirement for AGI to permanently self-learn implies the ***ability to detect this kind of cause-and-effect patterns*** during the operation of an AGI-controlled object. Appropriate behavior of an AGI-controlled object in a complex natural environment (for example, a robot in a crowd of people, or autonomously driven vehicle surrounded by many others, etc.) requires an analysis of relations between many types of events and a large number of parameters of the situation model. The broader the set of parameters describing the situation, and the more varied the real situations in the environment of the system's activity, the more likely it is to detect dependencies that are not provided for by the initial set of knowledge embedded in the system before its use.

In turn, knowledge of such causal relationships allows you to generate more adequate forecasts and make smarter decisions. Quoted in NUTS AND BOLTS OF THE DECISION MAKING the diagram after adding the module for detecting the discussed causal relationships becomes as follows:



The implementation of the detection of cause-and-effect relationships on a straightforward test problem of detecting a connection between the position of a particular object in the environment and an event of a collision with it is described below. Despite its simplicity, the task demonstrates the essential aspects of the approach.

A specific sensor measures the position of a moving object in the external environment at the request of the system and returns four values of the parameters describing the situation: the Cartesian coordinates of the object **x** and **y**, and the polar coordinates of the object, the distance **r** and azimuth **a**. It is required to determine which of the parameters (and which values) plays the cause of the collision event. Naturally, we know in advance the correct answer (which is required for the test problem): the cause of the collision is the decrease in the distance to a specific value. The method should show the expected result (distance as a cause) and the fact that the other three parameters are not suitable for the role of a collision predictor.

Many would probably suggest analyzing the ***correlation between the values of the parameter and the collision event*** as a natural way of finding the desired relationship and considering the parameter for which the correlation is greatest in absolute value as the cause of the collision. This approach, however, immediately reveals a few problems. First, the sought-for ***dependence itself is totally determined***; the use of correlation creates a ***false impression of the probabilistic nature of the dependence*** of the event on the cause. Second, in practice, it is not enough to know which parameter is associated with an event - we need to know which parameter ***values*** mean a collision and which do not; that is, ***some additional analysis is required***. And, finally, the most essential thing: thorough research shows that it is possible to choose such values of the parameters of the environment objects (not the parameters of the situation!), at which the ***correlations for each the estimated parameters of the situation are close to zero***, which makes it impossible to detect the desired causal relationship, so ***using correlation does not guarantee success***.

A workable alternative is an approach based on the analysis methods of cause and effect relationships, detailed in the book "**The book of why**" by **Judea Pearl & Dana Mackenzie**. In short, the essence boils down to ***formulating a hypothesis*** that a particular factor is the cause of events, with the ***subsequent verification of this hypothesis*** by analyzing the accumulated data (purposeful experimentation can be used to collect data). In our case, the data are the coordinates of random points uniformly distributed in space; each tuple of 4 parameters **(x, y, r, a)** is supplemented with information about whether there is a collision in this particular case or not. This simulates the collection of data received from the sensors. The hypothesis is applied in turn to all four parameters after each new sample is added to the dataset; the assessment result may be "***the hypothesis contradicts the facts***" or "***the hypothesis does not contradict the facts***". ***As the volume of accumulated facts grows, the number of hypotheses that do not contradict the facts falls to one desired hypothesis***; in parallel, an ***interval of distance values*** is formed, going beyond which leads to a collision. Numerical values for analysis are ***discretized***; the size of the discretization quantum does not play a fundamental role.

The source code for a Python program demonstrating the approach is available on GitHub.

In this case, the causal relationship is deterministic, and tests show this; in other situations, the sought dependence may be probabilistic. The approach can be extended to more complex cases when an event depends on several parameters.

An ***essential difference between this approach and methods based on statistical analysis is the need to formulate hypotheses***; the more complex the real reason, the more complex the sought hypothesis turns out to be. The ***complexity of hypotheses*** generated for subsequent analysis can be used as characteristics that determine the ***level of intelligence***.

Formulating hypotheses is in itself a very ***informal task*** - which explains the difficulty of identifying true causes. Suppose two parameters, distance, and azimuth, are removed from the model described above, leaving only Cartesian coordinates. In that case, it will be necessary to construct a hypothesis (along with possible others) that the criterion for predicting a collision is the combination **x*x + y*y**.

For the formation of hypotheses, an approach can be used in which an enumeration of ***combinations of several basic functions*** is used to generate hypotheses; a set of functions and an algorithm for enumerating compositions from them becomes part of the "***congenital***" knowledge of the AGI system.

Revealing causal relationships also helps solve another problem faced by developing AGI systems that can operate in natural conditions. Proper behavior in a changing environment requires the accumulation of experience, which serves as the basis for decision-making. The original source of information is sensors, which generate a ***massive amount of data*** in real systems. The number of sensors in a modern car is in the ***hundreds***, in airplanes - in ***thousands***, similarly in the case of autonomous security robots. Memorizing all this data is both impossible and irrational. Accordingly, the system should ***filter data by separating essential information from the one that should not be remembered***. Knowledge of cause-and-effect relationships makes it possible to ***separate data that affect the course of events from those that do not***. For example, if the system tracks the position of many objects, then it is reasonable to remember information only about those objects whose movements are dangerous or beneficial.

Subscribe to AGI engineering

By Mykola Rabchevskiy · Launched 2 years ago

AGI: fundamentals, architecture, implementation, source code