Minimal Goal-Directed Systems

Joscha Bach
Harvard Program for Evolutionary Dynamics
Cambridge, MA 02138

joscha@bach.ai

Abstract

How is it possible for a physical system to have goals? In the way we commonly use the notion, goals have qualities that are different from attractor states that characterize the dynamics of a system's progression. Goals are part of the characterization of at least *minimally intentional* systems. Thus, our question cannot be extricated from: What is the minimal descriptive frame that allows us to describe a system as representing, choosing, committing to and pursuing goals? And when we can answer that: what processes and dynamics are necessary and sufficient for the genesis of such a system?

Introduction

If you are reading this, you are clearly a system that should fulfill our requirements for having goals. Before identifying properties of minimal goal directed systems, let us briefly build an intuition of what having a goal entails for a human being, such as you. Humans are social primates living in a complex dynamic environment, and their survival in this environment necessitates complex regulation. Changes in the environment may disturb the parameters within which the human organism can function: for instance, heat and cold may push the body temperature outside of the range of metabolic homeostasis, falls from great height might decelerate it too rapidly to let it maintain cohesion in its cellular organization, and so on. To build and maintain its structure, the human organism needs to consume *negentropy* in many forms from its environment. Informally, to create and maintain order against a continuos onslaught of disturbances, an organism must consume order in its environment, and in that, it competes with many other organisms. You are quite literally the set of structural principles that has outcompeted all other structural principles in consuming the negentropy in your volume of spacetime.

Regulation starts with *feedback* systems, such as the ones in our brainstem, regulating heart rate, body temperature and breathing patterns. The controlled consumption of negentropy and avoidance of disturbances requires *second order regulation*, i.e. regulation that affects lower level feedback loops, for instance to switch behaviors on and off. *Pleasure* is a signal that tells the human nervous system to continue a behavior that is currently fulfilling one of its needs, while *displeasure* asks it to stop the current frustration of a need. To generate these signals, an organism must measure the need itself, and the changes in it, so it can correlate them to its behavior. Pleasure and displeasure are measures of *preference* between states.

The regulation of future behavior is the domain of *impulses*. Impulses are the result of the association of need changes with observable patterns in the environment that increase the likelihood to afford the satisfaction or frustration of the needs, and actions that change the probability of the occurrence of these patterns. The human neocortex is tasked (among other things) with abstracting these patterns into *situations*—encodings over observable patterns that allow to classify different patterns into sets, according to the need they can satisfy. As a result, we can form situation representations as abstract and complex as "restaurant", "fight", or "graduation". These situations can be arranged into sequences of events. If the probability of actualizing events can be influenced by actions, the events and transitions between them form *causal models*. Much of what we commonly call *intelligence* can be characterized as the detection of causal structure in a domain.

If a situation is part of a causal model in such a way that the individual believes that it can increase the probability of its occurrence by effecting suitable changes in the environment, and the individual is able to assign a preference to that situation, because it affords a consumption that directly or indirectly affects a need, then this situation is called a *motive*. A motive becomes a goal by an act of commitment: by changing the internal regulation of the human so that the motive is actively pursued via enacting changes of the environment that are believed to increase the probability of the manifestation of that situation. This act of commitment is what we call a *decision*. What are the minimal requirements to construct a system that is capable of causal modeling, assigning preferences, and making decisions? It turns out that we can capture the relevant elements of the description above in terms of information processing. Both the goal-directed system and its environment can be described as *computational machines*, and goals are a class of representational states within causal models created by certain computational *observers*.

The computationalist perspective

The computational perspective is perhaps the most significant contribution that the 20th century made to philosophy. It begins with the insight that the atoms of our epistemology should not be something as unfathomable as "phenomena", but discernible differences, i.e. *information*, which can be expressed as states, which are vectors of bits. As Claude Shannon discovered, systems can be characterized by their *information entropy* [1948], the number of states a system may be in, based on what we know about the system.

When systems change, we can describe that as changes in their state vector. Let us call the function that orders the states (i.e. derives one state from its predecessor) the *transition function*. This function may be deterministic or probabilistic, discrete or continuous. A *computational system* is one that can be described as a non-random succession of states, based on a single transition function that captures the regularities in all state transitions. Alonzo Church and Alan Turing famously discovered a class of systems that can all compute the same, infinitely large set of *(Turing) computable functions* [Church 1936]. Every computer that allows us to implement a *Turing Machine* [Turing 1936], and that can itself be implemented on a Turing Machine, can *effectively* compute all the same functions. (However, not all functions can be equally *efficiently* computed on all these systems.)

But of course, there are computational systems that have lesser or greater power than Turing Machines. For instance, if we limit the state vector and the state space to finite numbers, we obtain a *finite state machine*, which can only compute a finite set of functions. Using continuous transition functions, we can define a class of *hypercomputers* that can solve certain geometric problems with infinite precision. We can even define *a-causal hypercomputers* that use a transition function that derives part or all of the current state from one of its successor states: a kind of "time machine" that can tell us next week's lottery numbers today.

The characterization of systems as computational machines makes the perspective of mechanistic philosophy much more precise and useful, by replacing the intuition of a universe built from moving parts that are interacting in regular ways with a universe that changes its state by regular transitions. Julien Offray de LaMettrie anticipated that change in his famous pamphlet "L'homme machine" as early as [1748], when he argued that humans are best understood as machines, and that these machines should be seen as mathematical and transcendental.

When be begin to see the physical universe as computational, rather than mathematical, we realize that mathematics is the domain of all formal languages, most of which is uncomputable, whereas computation is the domain of systems that can be implemented, and thus may exist. Mathematics is the realm of all possible specifications, computation is the realm of implementations.

Causal systems

We usually define the universe using its causal envelope, i.e. everything that contributes in some way to the evolution of the universe's state vector is part of it. However, an observer that is situated in a universe does not have access to its whole state vector, and thus it will have to limit its observations to a small subset of the bits of the universe state, from which it may make predictions concerning the future state of another small subset of bits. As a result, observers are usually concerned with open systems, i.e. subsystems that must be described conditionally, depending on the influence or non-interference of unknown or independently modeled parts of the environment. A system in which the distribution of expected states changes (deterministically or probabilistically) as a result of such variables inflicted by other systems outside the current one is what we may call a *causal system* [cf. Pearl 2000]. The variable that yields the conditional difference in the state transitions is called the cause. In the extreme case, when we have a causal system that captures the entire state space and all possible transitions and conditions, we have an *algorithm*.

Causal systems are usually implemented by a *substrate*, that is, by the dynamics of other, underlying causal systems, which provide the means to store the information determining the state, and the process to execute the transition function. The relationship between a systemic description of the substrate dynamics and the causal structure its implements is called *supervenience*, or *supervenient emergence* The causal system is *stable* as long as the underlying dynamics do not compromise its transition function (otherwise, it will turn into something different). Because causal systems are rarely completely closed off from all possible influences of their substrate, they are usually incomplete characterizations of physical systems. However, if an observer finds itself in an entirely stable causal system, that observer will be causally insulated from all underlying dynamics, and cannot learn anything about the properties of the substrate beyond the fact that it provides the necessary and sufficient conditions for computing the observable state changes.

Causal systems, while being computational, do not necessarily have to be computable. They can often be characterized by emergent noncomputable mathematical dynamics that are nevertheless (approximately) computationally implemented by the substrate layer. For instance, the causal evolution of waves on the surface of a body of water can be described by partial differential equations, which are hypercomputational, but it does not follow that our universe can actually compute them with arbitrary precision. Instead, the waves are emergent statistical properties of the aggregate local interactions of very large numbers of water molecules, which in turn are emergent properties of the aggregate interactions of elementary particles, and so on. The properties and computational dynamics of water waves do not capture the implementation of the physical universe, but the best specification of a causal system found by an observer that is not coupled to particles, but to their aggregate properties, such as local changes in the height of the water surface.

Observers

Let us define a *minimal general observer* as a causal system (not necessarily one that is computable by a finite state machine) that has access to an environment, which is itself a bit vector with a regular transition function of some kind. The access is given by a subset of the environment bit vector (the *interface*) that is at the same time a subset of the observer bit vector, in such a way that changes in the shared bits do not immediately lead to a change in its transition function, i.e. destabilize the observer.

The observer must implement a function that leads to change in its internal state beyond the interface. The portion of the information in the interface that contributes to a change in the observer beyond the interface is the *observation*. The interface are only those shared bits for which a function is implemented in the observer that satisfies observer stability and causal influence on its internal state.

Using our definition, observers do not have to be humans. A camera will qualify as an observer, too. So will an ant, or a falling leaf that bends as a result of a gust of wind, or rock that increases its temperature while the sun shines on its surface.

Observers are characterized by the causal structure that we impose in our description of their coupling, their state space and their transition function.

Observers can be be described algorithmically, for instance using the idea of *Solomonoff induction* [1960]: Given a computational observer that is coupled to its environment with a vector of bits, the best model that the observer can possibly find about this environment is the shortest program among all the programs that best predict the current observation from all past observations, for all observations. We may use also non-algorithmic characterizations of our mental processes, for instance dynamical systems [cf. van Gelder 1998], which do not even have to be computable. The implementation of our memories and perceptions, and the causal structure of the transition function that lets us access and compare them, is only implemented as an approximation. The degree to which we can make observations of our environment is limited by the accuracy of our own implementation.

In a similar, but more fundamental way, any physical system that is coupled with another system it observes might itself only be realized in approximation. The degree to which a system can determine the state of its environment is not only given by the information that is available to the observer, but by stability and degree of realization of the state and the transition function of the observer itself.

Reversible computational substrates

I think that foundational physics ultimately explores the idea that there is a fundamental, causally closed layer to the universe. However, if we discovered such a layer, we may not say anything about its substrate beyond the fact that it can produce the basic dynamics that give rise to the observable universe, because a causally closed layer will insulate observers from everything below it. The machine that implements the universe: Aristotle's *Prime Mover*— the inevitable principle that moves the universe along without moving itself—is forever out our reach.

However, the laws of conservation that we empirically observe in our universe might indicate a fundamental property of its computations: the transition function that operates on its states could be *reversible*.

A reversible computer cannot destroy information, which means that none if its operations can delete bits. Conversely, our digital personal computers can delete bits: the result of an *AND* or an *OR* operator will not allow us to deduce the bits that went into the operation.

A reversible process is not necessarily a symmetric process. For instance, imagine a set of idealized, frictionless billiard balls, lined up in their well-ordered initial position on an infinite playing field. After we confer a strong initial impulse to them, they will enter a short phase of occasional collisions, which will end after they have spread out so far that no further collisions are possible. During their collisions, the balls confer information (momentum) to each other, which influences their trajectories, so that collided balls may take different courses at different speeds than the others. The asymmetry between the interactive phase and the solitary phase, and between "virgin" balls and collided balls is not in conflict with the reversibility of the process, but a function of the starting state and the openness of the playing field. The former may be reversed by deflecting all balls back into the field at the same time, and the latter by restoring the original distribution of momentums by inverting the collisions.

If we limit the playing field with a boundary, the balls will be deflected, and they will eventually reach an equilibrium dynamic. (If the field is discretized, i.e. the balls can only occupy a finite number of positions, the equilibrium dynamic will eventually enter an infinite loop.)

A reversible universe is deterministic: if the collisions and momentum exchanges of the balls are probabilistic, it will be impossible to perfectly restore the starting state by inverting all collisions.

It is trivially easy to implement a reversible computer on top of an irreversible one, simply by outlawing all operations that might delete a bit, or by keeping an “undo” history of those that do. It is also possible to build an irreversible computer on top of a reversible substrate, but each time a bit is deleted, it will have to be stashed away somewhere. Thus, the longer an irreversible system runs, the more indelible “garbage bits” it will accumulate.

Observers and other emergent causal systems have an interesting property: they are necessarily irreversible computers. To be stable, their implementation must involve mechanisms that keep their transition function (i.e. the characteristic dynamics of the system) unchanged in the face of small disturbances by the substrate. Observers are *ergodic systems*, their underlying regulation will keep the implementing causal structure in a limited parameter range that is independent of the system’s history. For example, our body temperature should be largely independent from yesterday’s room temperature. Any regulation that lets the system forget past disturbances must effectively delete bits, i.e. reduce its entropy and because these garbage bits cannot infinitely accumulate in the system itself without destabilizing it, they will have to be ejected from it, and thus require an environment that exhibits at least temporary openness.

This does not only apply to organisms, but to digital personal computers as well, which are built to exhibit the same functionality at different battery levels, different environmental temperatures and so on. It even applies to planetary orbits, or the formation of celestial bodies from interstellar gas. All these systems exhibit stability against disturbances, and consequently they depend on a disequilibrium, in which their environment can absorb the bits they need to delete.

The ability of a system to absorb entropy is called negentropy, and the existence of any stable causal system depends on the availability of a suitable negentropy gradient in its environment, and because the available negentropy is a function of a starting state of the evolution of a reversible system, causal systems implemented in a reversible substrate will have only a temporary existence.

Life

Different causal systems may compete for the same sources of negentropy. If negentropy gradients cannot be collapsed by simple systems, they open an opportunity for complex systems that may apply higher degrees of control through a more complex transition function. Using these opportunities may require systems to actively explore the space of transition functions, i.e. to turn into new systems with behaviors that are better adapted to their environment. The smallest universal machine we know to be capable of extracting negentropy over a wide range of environments is the cell. The cell is in fact very large, a structure that typically extends over 24 magnitudes above the Planck scale, and it implements all necessary functionality for maintaining homeostasis, replication, complex information processing, and executing an evolutionary search that lets its descendants adapt to different opportunities for the extraction of negentropy.

All life is made of cells, and while cellular replication and evolutionary adaption make it an incredibly robust causal structure, the genesis of the first cell from simple organic chemistry seems to have been an event so unlikely (or is at least so poorly understood) that it may have had to be helped along by a generous number of cosmic dice throws. The last universal common ancestor of all known cells (*LUCA*) is thought to have existed 3.8 Billion years ago, and it already possessed a complexity similar to today’s single cellular organisms [Weiss et al 2016].

The conditional information processing and specialization of cells enabled the evolution of multicellular organisms: stable causal systems on the next level of supervenience, implemented by large numbers of coordinated cells, with more complex information processing facilitated by specialized cells.

In turn, the evolutionary competition between organisms gives rise to coordinated groups of separate organisms, implementing an even higher level of supervenience.

The nervous systems of multicellular organisms will regulate the disturbances imposed on the coordinated group of cells by their environment, and in accordance with the *Good Regulator Theorem* [Conan and Ashby 1970], the plasticity of large multicellular information processing allows for regulation that is so flexible that it will tend to form models of this environment. Hence, multicellular organisms with a sufficiently complex nervous system can usually be characterized as *agents*, a notion to describe goal-directed computational systems, which was developed in the context of Artificial Intelligence in the 1980ies [cf. Bratman 1987]. The same is often even true for coordinated groups of agents, i.e. organization of agents can implement supervenient causal structures that realize all criteria of agency by themselves.

Goals

Colloquially speaking, an agent is a causal system that is an observer with internal states that encode *beliefs*, *desires* and *intentions* (*BDI*) and that is capable of affecting its environment according to its intentions.

A *goal* is a world state represented by an observer, in such a way that the observer realizes a causal mechanism that will regulate the environment to increase the probability of occurrence of that state, according to the model of the observer.

Thus, a goal has complex prerequisites, it does not just entail the tendency of a system to increase the probability of an reaching a state! Rather, goals are *intentional states*, and they do not concern present conditions and actions, but predicted ones. A goal presupposes at least the following components in an observer:

1. A mechanism that maps observations to suitably generalized encodings (world states), which we may call the *encoding function*.
2. A *prediction function* that allows to anticipate future world states in a causal structure, i.e. in a model that extrapolates which operations on the environment will influence the probability of reaching a given world state. (The encoding function and the prediction function implement the generation and manipulation of *beliefs*.)
3. A *preference function* that imposes an ordering on the anticipated possible alternative world states. (The preference function implements *desires*.)
4. A *decision function* that changes the agent's state (*intention*) so that it anticipates to perform operations that sufficiently increase the likelihood of the occurrence of a selected, preferred world state, according to the prediction function.

As we see, this understanding of goals does not only capture our intuition of goal-directed behavior in humans. Rather, it does so for computational systems in general: agents are the minimal causal structures that are capable of holding goals in a a meaningful sense. While agents are characterizations of particular physical systems, most physical systems cannot meaningfully described as exhibiting goal directed behavior, in the absence of implementing functions for encoding, prediction, preferences and decision making.

To explain how this computational notion of agency can arise in a physical universe, I have taken you on an brief excursion from information based epistemology, the metaphysics of computational machines, an understanding of causal systems and computational observers, through computable worlds, irreversible dynamics in reversible worlds and machines that evolve to defy entropy, to finally characterize observing agents that do not just model the world but are also capable of holding preferences and to make decisions.

References

Bratman, M. (1987). Intentions, Plans and Practical Reason. Harvard University Press

Shannon, C.E. (1948), "A Mathematical Theory of Communication", Bell System Technical Journal, 27, 379–423 & 623–656, July & October, 1948

Church, A. (1936). "A Note on the Entscheidungsproblem". Journal of Symbolic Logic, 1, 40-41

Conant, R.C., Ashby, W.R. (1970). "Every good regulator of a system must be a model of that system", International Journal for Systems Science, 1970, vol 1, No 2, 89–97

van Gelder, T. (1998). "The dynamical hypothesis in cognitive science." Behavioral and Brain Sciences 21, 615–665

LaMettrie, J. O. (1748): "L'Homme Machine"

Pearl, J. (2000). "Causality, Models, Reasoning and Inference". Cambridge University Press

Solomonoff, R. J. (1964). A formal theory of inductive inference. Information and Control, 7:1—22, 224—254

Turing, A.M. (1936). "On Computable Numbers, with an Application to the Entscheidungs problem". Proceedings of the London Mathematical Society

Weiss, M.C., Sousa, F.L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., Martin, W.F. (2016), "The physiology and habitat of the last universal common ancestor." Nature Microbiology, Jul 25, 2016, 1(9)