Towards a Semantics of Desires

George Kiss
HCRL
The Open University
Milton Keynes
England

Han Reichgelt
Psychology Department
University of Nottingham
Nottingham
England

HCRL Technical Report No. 70
30 April 1991
Abstract

As part of an effort to define a unified formal semantics for beliefs, desires and action, this paper sketches a model theory for the axiological aspects of agent theory: hedonic states, likes, goals and values. Particular attention is paid to modelling the intensity of likes. The main intuition underlying the model theory is that the axiological aspects of agent theory can be modelled through computational generalisations of physical dynamics. Computational analogues of force, mass and potential are offered.

Introduction

An important part of agent theory appears to be the notion of desires. Several formulations of agent theory have adopted beliefs, desires and intentions as a set of basic notions (the so-called BDI models). However, to our knowledge, so far relatively little has been said explicitly in the AI literature about a theory of desires (Cohen and Levesque, 1985 and in press, Moore, 1985a; Kiss, 1988, Shoham, 1989).

This paper takes some initial steps towards the explicit formulation and formalisation of such a theory. We concentrate on axiological issues, covering hedonic states, likes, goals and values (Kiss, 1988, 1990).

Among the many issues surrounding desires, we select the question of the intensity of the attitude of liking for detailed treatment. We think that likes are not the only attitudes that have an intensity aspect. It is common to talk about the strength of beliefs too. We hope to extend our approach to those other attitudes as well in the future.

Differences between the intensities of likes are often called preferences in the literature of decision theory, economics and psychology. Preferences are usually taken as the primary, primitive, notions in the sense that preferences are directly manifested in the choices made by an agent. Likes are therefore treated as relative, comparative attitudes. Few disciplines enquire into the mechanisms that might determine such choices and it is usually assumed that it is preferences that are directly available to the agent. Absolute values of liking are usually recovered from behaviourally expressed preferences by some analytical computations from the preferences.

We would like to proceed in the opposite direction and take absolute likes as primary which in turn determine preferences. Our intuition is that an agent has representations of how far it likes various things and when faced with a choice, compares the intensities of its likes to compute a preference. This need not of course exclude mechanisms of context dependence and interaction effects.

Our longer-term research objective is to formulate a unified formal semantics for beliefs, desires and action and to lay foundations for implementation work. This short paper has limited aims. Our main concern is to refine the set of intuitions which were outlined in Kiss (1988, 1990, 1991) and sketch the model theory for a modal logic of liking. We defer the definition of the syntax and semantics of the logical language, the statement of axioms and the derivation of theorems for another paper.

The main intuition we wish to convey is that the axiological aspects of agent theory are best interpreted in terms of concepts that are computational generalisations of physical dynamics. Traditionally in physics dynamics deals with changes of state in a system...
and with the causes of these changes, usually conceptualised as forces. Modern developments have turned dynamics into a more abstract area of study, as we shall briefly sketch below.

We propose that dynamics has a natural place in agent theory, since that theory is vitally concerned with (mental) states, their properties, and with the dynamics of sequences of changes in mental state. The interpretation of knowledge and belief as a state of the agent has recently been gaining ground (Rosenschein, 1986; Halpern and Moses, 1985). There have been increasing efforts also in forging a link between knowledge and action, (Moore, 1985b; Cohen and Levesque, in press), thereby introducing a dynamic element, because of the changes caused by action. While these authors have been concerned with agent dynamics, in a sense, they have not attempted to link their logics to dynamical systems in the physical sense. In this paper we hope to continue this trend by filling in more detail about agent dynamics.

The rest of the paper is organised as follows. We first review some relevant concepts from abstract dynamics. Next, we discuss how agent-theoretic concepts can be interpreted in such terms. Finally, we formulate computational generalisations of physical concepts like potential, force, velocity, etc., and indicate how they can provide a framework in which to interpret axiological concepts in agent theory.

Concepts of Abstract Dynamics

The main concepts of recent developments in dynamics deal with the structure of state spaces. Abstractly, the theory can be formulated in terms of functional iteration. The functions which define dynamical systems are also called mappings or maps. The main concern of the abstract theory is with the asymptotic behaviour of iterative mappings. The iteration of a function is a discrete process. If the process is continuous, the description is often given in the form of differential equations to describe the behaviour of the solution over time.

In a geometric interpretation, the iterative process maps points into points. The points correspond to the states of the process. The process is then said to go through a trajectory or orbit of points. The main concern of dynamics is to understand the nature of all trajectories of a system and to classify them as moving to a fixed point, being periodic, asymptotically periodic, etc. We shall now turn to an informal summary of some of these concepts. For more detail, see, for example, Abraham and Shaw (1981), Devaney (1986), Thompson and Stewart (1986) or Cvitanovic (1984). Cvitanovic also contains an extensive bibliography. The field is developing very rapidly under the designation of chaos theory, which is a specialised branch of dynamics.

The state space of a system is generally a topological surface (manifold) on which the possible states of the system are located. This can be just three-dimensional space, or some curved surface, for example, like a doughnut (torus).

It is normally assumed that there is a force vector field acting at all points of the state space. This vector field determines the dynamics of the system by constraining the trajectories to certain directions at each point of the state space. When typical or many trajectories of the system have been drawn, we get a phase portrait of the system.

Closed trajectories produce cyclic behaviour. Trajectories can otherwise take many shapes, like spirals, straight lines or any kind of curve.

The focus of interest is in the asymptotic behaviour of trajectories. Limit sets of state spaces are sets of points towards which the trajectories move asymptotically. Limit sets
may be solitary points, or cycles, or more complicated distributions of points. Limit sets which are solitary points, are called fixed points. 

Fixed points of functions are points \( x \) for which \( f(x) = x \). That is, the fixed points are mapped into themselves by the function. Fixed points are important in dynamics, because they correspond to equilibrium (steady) states of systems. Once a system has somehow got to a state which is a fixed point, it will not move from that state under the iteration of the function \( f \).

It is of interest to ask how a system may get to a fixed point. The simplest case is that the system may start from an initial state that is a fixed point, and there will be no further change. More interestingly, trajectories starting at other states may lead to a fixed point after a number of transitions. In such cases we say that the fixed point attracts the trajectory. The set of states from which trajectories lead to an attractive fixed point are called the basin of attraction of the fixed point. It turns out that a fixed point is attractive if the slope (derivative) of the function \( f \) is less than 1 at the fixed point. The magnitude of the slope characterizes the strength of the attractor: the greater the strength, the faster the trajectory approaches the fixed point.

A periodic point is a generalisation of the concept of the fixed point to the case when a trajectory cyclically visits a point after every \( n \) iterations of the function \( f \).

If the iteration is run backwards, trajectories would appear to diverge from an attractive fixed point. In this situation the fixed point is called a repellor. Such fixed points correspond to unstable equilibria in physical systems. Slight disturbance from the equilibrium starts the system on a trajectory leading away from the equilibrium state. Conversely, attractive fixed points correspond to stable equilibria.

**Agent Attributes and Dynamics**

We now briefly review how to interpret the agent-theoretic concepts of interest in this paper in terms of abstract dynamics.

**Compositionality.**

We assume that complex agents are architecturally compositional both structurally and behaviourally. The complex agent structure is produced by assembling simpler component elements. Complex agent behaviour is produced through the (often nonlinear) interactions between the simpler component behaviours. Concurrency, parallelism and distributed systems become important issues.

**The agent as controller.**

We assume that the agent acts as a controller with respect to the world state. The agent exerts control by taking actions. We include "doing nothing" as an agent action. Taking an evolutionary point of view, we assume that ultimately this control is in the interest of fitness for survival. Fitness for survival is dependent on the existence of certain world states, or on keeping them within permissible bounds. We assume that environmental events produce disturbances in the agent's internal state by causal effects conveyed through inputs. Agent action attempts to counteract such disturbances. An agent can control the world state either by changing its internal state or by attempting to change the external state. For example, the agent may change its beliefs or it may locomote to another location.

We want to distinguish between a system's natural dynamics (might also be called the free dynamics) which is operating when the agent is executing the "null" action, and the constrained dynamics that results from the composition of the free dynamics with the
control dynamics produced by the non-null agent actions. The distinction is motivated by recognising that not all events in the world are produced by agent actions. On the other hand, agent actions are causing changes of state, but these changes we shall think of as either setting up starting states for the free dynamics, or as jumping between trajectories of the free dynamics. However, we defer a more detailed explanation of our model of agent action to a later paper.

Axiological aspects of agents

Axiological issues are concerned with the directional nature and asymptotic behaviour of agent dynamics. The teleological (goal-directed) nature of agent behaviour is one of the central examples of such issues. In terms of dynamic system theory, the dynamics can be described in terms of the movement of the system state towards stable equilibrium states and away from unstable equilibrium states. Teleological agent behaviour is to be identified with movement towards stable equilibria which are in this sense preferred states of the system: we shall say that the agent "likes" to be in these states. Aversive agent behaviour is to be identified with movement away from unstable equilibria which are in this sense disliked by the agent. In the terminology of dynamic system theory, these states are attractors and repellors. Unstable equilibria arise mainly through competition between attractors and represent boundaries between the basins of attraction of those attractors. Attractors and repellors determine the direction of movement, i.e. the direction of agent action. It is natural to interpret the pro- and anti-attitudes of agents with this kind of directionality. We assume that due to the physiological structuring of living organisms attractors and repellors are created in their behavioural space. By analogy, it should be possible to create attractors and repellors in non-living computational systems through appropriate construction or programming.

A related point of view is found in optimisation theory. In this approach the main underlying idea is that the states and trajectories of a dynamic system are governed by some principle that can be expressed mathematically as finding the stationary value (usually maximisation or minimisation) of an "objective" (or goal) function. There is a great deal of work on the application of such optimality principles to evolutionary, ecological, economic and behavioural processes. We wish to look upon this approach in the same spirit and regard the extrema of the objective function as specifications of the attractors and repellors of the state space. In its application to the description of behavioural or economical processes the objective function is usually called utility. Note that utility is here a descriptive aspect, revealed by the observation of behaviour. In other applications to evolutionary processes the objective function is taken to be fitness for survival. We are of course more concerned with individual agent behaviour and hence with utility in this paper. In summary, from the viewpoint of optimality theory, the agent is maximising utility.

In a utilitarian framework utility would be some function of hedonic states, i.e. pleasure and pain. One might speculate that pleasure and pain are related to fitness and have been incorporated in the architecture of organisms to make available to the individual some state variable that can be used as an indicator of fitness. Such an interpretation would not be unnatural in the case of pain as an indicator of damage and hence loss of fitness and pleasure as an indicator of health and hence of maximisation of fitness. For the time being, we adopt this utilitarian framework and assume that the agent is maximising a hedonic function.

The values of an agent correspond to global (high-dimensional) attractors and repellors of the composite dynamics. We think of values as global attractors which may never be reached or closely approached by trajectories, due to the topological structure of the state space created by the competition between them. In complex agents explicit representations of values form a value system.
A goal of an agent corresponds to a local (low-dimensional) attractor in a basin of attraction of the composite dynamics. We think of goals as attractors which are reached or closely approached by nearby trajectories.

To support our intuitions, we wish to use a mechanical analogy. According to this analogy the intensity of a desire (liking) should correspond to some abstract "force of attraction" acting on the agent, producing acceleration of state change.

Similar conceptual frameworks have already been used in mechanical engineering and in robotics (see Koditschek, 1989 for a review). In mechanics it is well known that the total energy of a dissipative system (expressed by the Hamiltonian) will monotonically decrease and will be asymptotically stable. A known technique in robot control engineering is to use feedback control which amounts to following the gradients of total energy. This technique has been used for robot arm control. Direct utilisation of the potential field has been used for path planning with obstacle avoidance in mobile robots (Barraquand and Latombe, in press).

In our mechanical analogy too, the forces would be derived from a potential field and the agent is assumed to follow the gradients of the potential. From the point of view of optimisation theory, the objective function is used as the potential. The description of such a potential therefore amounts to the specification of a goal which is the asymptotically stable equilibrium state of the agent. We can also represent a value system in this analogy as additional potentials, with opposite sign, superimposed on the potential created by the goal. In the robot navigational applications such potentials are used to represent obstacles to be avoided while moving towards the goal state. In our analogy these obstacles correspond to elements of the value system, expressed as "prohibitions". The analogy is reasonable in the light of value systems often being expressed in the form of prohibitions (laws, regulations, etc). Presumably positive values (obligations) could always be re-expressed in a negated form.

Finally, the usefulness of knowledge for an agent is, of course, in guiding action towards a goal. In process dynamics terms the agent needs knowledge in order to tell what trajectory to follow. However, we shall pursue the interpretation of epistemic concepts in a dynamics framework in another paper.

**Model theory**

In this section we review some of the fundamental concepts that we need for our model theory: space, time, state, and process (trajectory). Our formulation draws on and extends previous work by Rosenschein and Kaelbling (1986) on agents as situated automata and by Halpern and Moses (1990) on knowledge in distributed systems. Our main concern is the formal characterisation of a process (or trajectory). For this, we need formal notions of time, space and state. We describe each in turn briefly.

*Time* is analysed as consisting of a set of instants $T$ and a total ordering relation $<$ over $T$.

*Space* will be regarded as a set of locations $L$. We shall not assume any specific topology over $L$, but wish to partition $L$ into subsets, which we shall call *systems*. In particular we shall want to distinguish the set of locations that constitute an agent from the set of locations that constitute the agent's environment. Sometimes we shall call these the set of internal and external locations, respectively.

*States* are defined as functions from locations to data values. We assume that for every location $l$ in $L$, there is a set of data values $D_l$ that this location can take. We distinguish between global and local states as follows. Global states are functions which assign to every location $l$ a data value from the appropriate set $D_l$. Given a set of
locations \( L \), we define \( GS_L \) to be the set of possible global states. Clearly, if the number of locations is \( n \) then \( GS_L \) can be regarded as an \( n \)-dimensional space. If \( g \) is a global state, then \( g(l) \) denotes the data value assigned to location \( l \) by \( g \).

A local state is a function which assigns appropriate data values to a subset \( Loc \) of the set of locations. The set of all possible local states over \( Loc \) is denoted \( LS^{Loc} \).

Processes are defined as temporal sequences of states. Since the concept of state is tied to that of space through locations taking on data values, it is natural to regard processes as occupying a spatio-temporal region. Following Rosenschein and Kaelbling (1986), we capture these intuitions in two steps. First, at each instant in time a process can be regarded as occupying a set of locations. Second, each occupied location takes on a specific data value. We thus have two functions. The first is a function from \( T \) to subsets of \( L \) determining the occupied locations, while the second associates data values with these locations. We can thus generalise the notion of state to processes. The state of a process at time \( t \) is determined by the set of locations occupied at \( t \) and their data values.

Just as we did with states, we distinguish global and local processes. Global processes are temporal sequences of global states. Thus, global processes occupy, and assign data values to, all locations at every instant in time. Halpern and Moses call such a global process a "run" of a system. A global process or run can be regarded as one possible way the world can unfold over time, or a "possible world". Formally, a run is a function from \( T \) into \( GS_L \). We denote the set of all runs by \( R \) and an individual run by \( r \). Then \( r(t) \) gives the state of the run \( r \) at \( t \), and \( r(t)(l) \) gives the data value of location \( l \) assigned by the run \( r \) to \( l \).

Local processes occupy only a subset of \( L \) at each instant in time and are thus a sequence of local states. Local processes can also be thought of as subprocesses of a global process. Formally, the spatial region occupied by a local process is a function \( \pi \) from \( T \) and \( R \) into \( \text{Powerset}(L) \). The state of the process is then given by a function \( s(\pi, r, t) \), which is a set of \(<\text{location}, \text{data-value}>\) pairs: \( \{<l, r(t)(l)> | l \in \pi(r, t)\} \). This notation emphasizes that the data values of the local process depend on the run of which it is a subprocess.

The foregoing define processes in a very general way. For many applications simpler special cases are sufficient and are conceptually easier to handle. For the purposes of the rest of this paper we introduce fixed-location processes, which occupy the same locations at every instant in time. Thus, fixed-location processes do not move spatially and the only change that takes place at successive instants of time is that the fixed locations take on different data values.

In the agent-theoretic use of this model we shall analyse complex agents into sets of (fixed-location) processes. This is the usual picture of a distributed, concurrent computing system. The component processes in such a system interact with each other through constraint relationships, implemented as message or signal passing. Connectionist architectures can also be interpreted in such a model; here the messages are values, usually in the real number or boolean data domains. Thus our framework is very general and can be used for analysing a wide variety of system architectures.

We now show how this model theory can accommodate the notion of an "accessible world", which we will need for the construction of a modal logic of desires. A process \( \pi \) only occupies a subset of the set of all locations at time \( t \). It is therefore possible for different runs to assign the same state to \( \pi \) at \( t \). We shall call such runs alternative runs with respect to \( \pi \) at time \( t \). Thus, the runs \( r \) and \( r' \) are alternative runs with respect to
process $\pi$ at time $t$ if $s(\pi,r,t) = s(\pi,r',t)$. If we identify a process with an agent situated in the world, the alternative runs can be regarded as different states of affairs which are indistinguishable as far as the state of the agent is concerned. This construction gives us a way to interpret the epistemic operator in our logic.

**State Transition Functions and State Space**

In order to introduce dynamics into the model theory, we introduce state transition functions. Let $L$ be the set of locations. Then a state transition function for $L$ can be defined in general as a function which maps every state of $L$ into a new state of $L$. Formally, a state transition function $f$ for a set of locations $L$ is a function from $GS_L$ into $GS_L$. Clearly transition functions over $L$ define the behaviour of the global process.

In specific cases, the new states of locations in $L$ may not depend on the previous states of all the locations in $L$, but only on a subset of them. For certain architectures, like cellular automata, the transition function is defined for each location separately as a function of only the immediate neighbours of that location. For logic circuits too, the transition function of a location is usually only a function of a small subset of specified locations.

Each state $s$ of the agent process is a substate of a state of one of the possible runs (trajectories) of the global process. The set of possible runs as been denoted $R$. Trajectories in $R$ are generated by the iterated application of a transition function $f$, $s_{i+1} = f(s_i)$. The transition function $f$ represents the changes brought about by the agent’s actions, including doing nothing. Recall that the changes may be either internal or external to the agent.

We shall be concerned with nonlinear functions $f$ which have attractive limit sets. In the case of a limit set which is just a single point, we have a fixed point $s_{\text{fixed}}$, such that $f(s_{\text{fixed}}) = s_{\text{fixed}}$. Here $s_{\text{fixed}}$ is an attractor. For all attractors there is an open set, called the domain of attraction $D$, such that for all states $s \in D$ the iterated application of $f$ eventually carries the state into $s_{\text{fixed}}$.

We distinguish global attractors from local attractors. The transition function may assign the same state to a subset of locations at different times. We call such a local fixed state a local attractor. By analogy to global domains of attractions we can define a local domain of attraction in a straightforward manner.

**Hedonic Functions**

In order to use the model theory for the interpretation of desires, we introduce hedonic functions. Intuitively, the hedonic function specifies the amount of pleasure or pain an agent experiences in some state. The hedonic function $h^\pi$ of a process $\pi$ maps states into hedonic data values. The domain of hedonic data values $H$ is a partially ordered set, containing a distinguished element $\text{neutral}$, corresponding to a neutral hedonic data value. All other hedonic data values are either hedonically greater than or smaller than $\text{neutral}$. The hedonic relational operator will be denoted by $<_{\text{Hed}}$. We shall assume that the hedonic state of an agent depends only on the local state of the agent. As indicated below, we will interpret the hedonic state as a computational analogue of potential field, with a potential of zero corresponding to the hedonic $\text{neutral}$.

The hedonic state results from the superposition of attractive and repelling potentials at the point corresponding to the current state. These potentials are produced by the agent's value system and by the current goal. The contributions of individual attractors
and repellors can be separately computed as \( h^\pi(s, s_{\text{fixed}_i}) \), where \( s_{\text{fixed}_i} \) is the fixed point state corresponding to attractor \( i \).

As is usual in utilitarian agent theories, we assume that an agent acts in order to maximise its hedonic state. The computation the agent executes to determine its action is therefore the optimisation of the hedonic function. Maxima of the hedonic function correspond to limit sets in the state space of the agent.

**Computational Analogues of Force, Mass and Potential**

We assume that for each point \( s \) in \( GS^L \) we can define the distance \( d(s, s_{\text{fixed}}) \) as the distance between the point \( s \) and the fixed point \( s_{\text{fixed}} \). If we regard \( GS^L \) as an \( n \)-dimensional space, then this could be euclidean distance. If we interpret each function iteration of the transition function \( f \) as a unit of time then we can define the velocity at \( s \) as the distance \( d(s, f(s)) \), since this will be the distance travelled in unit time. The definition of a force vector \( F \) acting at the point \( s \) follows the mechanical analogy and is the product of acceleration and mass. It is natural to interpret mass in our computational domain as some measure of the size of the state \( s \). For example, we can take the number of locations occupied by the agent process, \( n \), as this measure. Then,

\[
F(s, s_{\text{fixed}}) = n \cdot d(s, f(f(s))) - d(s, f(s)).
\]

The analogy can then be even further extended by defining force, as in physics, as the gradient of a potential, \( F = \text{grad} \, H \).

The joint effect of two or more fixed points at \( s \) can then be reflected by vector addition of the forces acting at \( s \). Let us denote two such forces by \( F_i \) and \( F_j \) for two different fixed points. The joint effect is then \( F = F_i \oplus F_j \) where \( \oplus \) denotes vector addition.

We can now assess the relative strengths of two attractors by comparing the magnitudes of the two forces and say that \( \text{Greater}(s_{\text{fixed}_i}, s_{\text{fixed}_j}) \) if \( |F_i| > |F_j| \).

The intuitive agent theoretic interpretation of these concepts is then as follows. As stated before, the potential is interpreted as the hedonic state. Components of the potential correspond to the values of the agent. The forces correspond to the intensity of liking. The concept of relative intensity, or preference, is based on the comparison of forces. We model the activity of the agent as following gradients in a potential field produced by the superposition of all the forces, i.e. values, acting on the agent. Gradient following corresponds to hedonic maximisation.

**Conclusions**

We have described some intuitions about the interpretation of axiological aspects of agent theory in terms of concepts from physical dynamics. The first steps have been taken towards formalisation by sketching a model theory. This model theory can be used straightforwardly for the construction of a logical language in which to reason about an agent’s hedonic state, likes, goals and values. We believe that the model theory can also be used for an integrated interpretation of axiological, epistemic and praxiological aspects of agent theory.

As indicated in Kiss (1991), such a model theory can also offer a link between concerns for formalisation and concerns for implementation strategies. As shown by Rosenschein’s work on situated automata theory and the implementation language
REX, there is a complementary relationship between a mathematical model and a physical phenomenon, both of which can be taken as alternative interpretations of a logic. When this is the case, the logic can be used for reasoning about a design, the mathematical model provides the semantics of that reasoning, while the physical phenomena (or their computational analogues) can be used for the implementation of the design. Our motivation for this work is thus not just formalisation, but implementation as well.

References


