Goals in a Formal Theory of Commonsense Psychology

Jerry R. HOBBS and Andrew GORDON
University of Southern California, Marina del Rey, California

Abstract.
In the context of developing formal theories of commonsense psychology, or how peole think they think, we have developed a formal theory of goals. In it we explicate and axiomatize, among others, the goal-related notions of trying, success, failure, functionality, intactness, and importance.

Keywords. Goals, commonsense psychology, commonsense reasoning, intention

Introduction

While robots today are capable of remarkable physical action, their capabilities in simulating or interpreting human cognitive behavior is considerably more limited. In this paper, we present a step toward resolving this problem in the form of an inferential theory of human goals, authored in first order logic. This theory is part of a larger effort to formalize a broad range of commonsense inferences related to human psychology [17], and supplements previously published reports on theories of human memory and emotions [9,18]. These papers describe the empirical grounding of this work in linguistic data and in people's descriptions of their strategies [8]. The set of concepts surrounding the idea of a goal turned out to be among the most important areas of commonsense psychology. Goals are central not just in our work but traditionally in the philosophy of mind, in psychology, and in work in artificial intelligence on belief-desire-intention (BDI) models of agents.

Section 1 of this paper briefly reviews some of the vast literature in this area, especially in psychology. In Section 2 we present succinctly the notational and ontological background required to get an enterprise like this off the ground. In Sections 3 through 8 we explicate and axiomatize the concepts of goal and a number of significant concepts that depend on goals.

1. The Psychology of Goals

One of the predominant themes in our research is to underscore the difference between cognitive models and inferential theories. In developing an inferential theory of human goals, our aim is not to characterize the actual cognitive mechanisms in which goals affect human behavior. Instead, our interest is in characterizing the commonsense infer-

Corresponding Author: Jerry R. Hobbs, Information Sciences Institute, University of Southern California, Marina del Rey, California; E-mail: hobbs@isi.edu.

ences that people make when they are thinking about their own goals and those of others. We are less interested in the way people actually think than in the way people think they think. These non-scientific models of mental states and processes are historically and often pejoratively referred to as naive psychology or folk psychology. But the proper characterization of these inferential theories is itself a scientific endeavor. Our efforts may be as relevant to the engineering of human-computer systems as psychological models of human cognition. Given the relevance of cognitive models of human goals to inferential theories of goals, we highlight some trends in the psychology of goals.

Eccles and Wigfield [3] review the historical progression of theories of human motivation beginning with Atkinson's [1] expectancy-value theory, which proposes that motivation and effort are the combined result of people's expectations of success and the value that they attach to that success. Following Atkinson, Weiner [25] proposed attribution theory, which emphasized an individual's interpretations of their achievement outcomes in motivation over motivational dispositions or actual outcomes. Weiner's model defines achievement attributions, including ability, effort, task difficulty, and luck, along the three dimensions of locus of control, stability of causes, and controllability of causes. Contemporary theories within these traditions are the subject of a recent edited volume on the psychology of goals [21].

A separate line of psychology research has focused on characterizing the goals that are the target of individual's achievement attributions. Ford and Nichols [5,6] characterized 24 types of achievement goals, ranging from within-person goals of happiness and intellectual creativity to person-environment goals of social responsibility and material gain. To develop a more all-encompassing taxonomy of human goals, Chulef et al. [2] identified 135 achievement goals in previous research on developing comprehensive categories, including Murray's list of 44 variables of personality [22], Rokeach's 18 instrumental and 18 terminal values [23], and 56 goals from the study of Wicker et al. [26]. They then developed a hierarchical organization of 135 goals by having different populations of subjects sort these these goals into conceptually similar groups.

2. Notational and Ontological Background

In order to deal with a domain as complex as commonsense psychology, including goals, one must build up a great deal of conceptual and notational infrastructure and make a large number of warranted but highly controversial decisions about representation. We have done that, and we cite the relevant papers below as appropriate. In this presentation of the necessary background, we do not present the arguments in favor of our decisions; the interested reader can consult the references. Citations of the appropriate literature in these areas occur there as well. It is important to note that while we are driven in our theory construction by the need to capture concepts that occur in the data we analyzed, the theories themselves cannot follow that data slavishly, but must be constructed in as workable and elegant fashion as possible.

The concepts we introduce in the background theories and the commonsense psychology theories are often not *defined*, but are rather *characterized*, by richly axiomatizing the concept and thereby constraining the possible interpretations of the corresponding predicate. For example, we do not attempt to define `cause`, but we do encode its defeasible transitivity and make it available for expressing causal knowledge in many specific domain theories.

The domain of discourse is the class of possible individual entities, states, and events. They may or may not exist in the real world, and if they do, it is one of their properties, expressed as `(Rexist x)`. (Throughout this paper we use a subset of the notational conventions of Common Logic.) In a narrowly focused inquiry it is often most perspicuous to utilize specialized notations for the concepts under consideration. But our view is that in a broad-based effort like ours, this is not possible, and that it can be avoided by sufficient judicious use of reification.

For example, we treat sets as first-class individuals. Moreover, sets are taken to have "type elements", whose principal feature is that their properties are inherited by the real elements of the sets [13,16].

The term "eventuality" is used to cover both states and events [11,14]. Eventualities like other individuals can be merely possible or can really exist in the real world. We can speak of the "arguments" of eventualities or the participants in the states or events. The expression `(arg* x e)`, for example, says that `x` is a direct argument of `e` or an `arg*` of an eventuality argument of `e`. We have axiomatized a theory of time [19], and eventualities can have temporal properties. Thus, `(atTime e t)` says that eventuality `e` occurs at time `t`.

Eventualities are very finely individuated. For example, Pat's flying to Toronto and Pat's going to Toronto are two different eventualities. However, they are closely related by a relation we call `gen` (for "generates"). The principal property of this relation is that if `(gen e1 e2)` and `e1` really exists, then so does `e2`. It says that Pat's going occurs by virtue of the fact that Pat's flying occurs; the flying constitutes the going.

Eventualities can have type elements of sets as their arguments, and when they do, they are eventuality types. An `instanceOf` relation relates eventuality types and tokens. A notational convention we use is that whereas the expression `(p x)` says that predicate `p` is true of `x`, the expression `(p′ e x)` says that `e` is the eventuality of `p` being true of `x`. The relation between the primed and unprimed predicates is given by the axiom schema

```
(forall (x)
   (iff (p x)(exist (e)(and (p′ e x)(Rexist e)))))
```

We have posted our development of a number of the necessary background theories at [16]. They include, in addition to theories of sets, eventualities, and time, a theory of composite entities, or things made of other things, a theory of change of state, and a theory of causality (cf. also [15]). The key distinction in the theory of causality is between the monotonic notion of a "causal complex", which includes all the eventualities that need to happen or hold for the effect to occur, and the nonmonotonic or defeasible notion of "cause", which is the context-dependent eventuality which is viewed somehow as central in the causal complex. The principal properties of a causal complex are that if the whole complex really exists, then so does the effect, and that every eventuality in the causal complex is relevant to the effect in a sense that can be made precise. Elements of a causal complex other than the cause are said to "enable" the effect. (The use of the concept of "cause" is sometimes taken to be controversial, but it is so pervasive in commonsense reasoning it seems to us indispensible in this enterprise.)

Agents have beliefs. We take the objects of belief to be eventualities. Because eventualities are very finely individuated, there is a straightforward translation between talking of belief in an eventuality and belief in a proposition. Thus, the expression `(believe a e)` can be read as saying that agent `a` believes the proposition that eventuality e really exists. We have developed but not published our treatment of belief, but our use of the predicate here should be obvious and unproblematic.

Mutual belief figures in several places below. The chief inferences associated with mutual belief are that if a set s of agents mutually believes e, then they mutually believe they mutually believe it, and every member of s believes it.

Any treatment of commonsense knowledge requires a mechanism for defeasibility. We are assuming in our work that a system using weighted abduction would be applied to the set of axioms. We indicate the defeasibility of a rule by including the conjunct (etc) in the antecedent of implications. It is really an abbreviation of a predication unique to that axiom of the form (etc-i x y ...). It can be thought of as the negation of the abnormality predicates in circumscription [20]. It should be straightforward to translate these indications of defeasibility into the formalisms required by other adequate approaches to nonmonotonicity.

3. Goals, Subgoals, and Plans

Human beings are intentional agents. We have goals, we develop plans for achieving these goals, and we execute the plans. We monitor the executions to see if things are turning out the way we anticipated, and when they don't, we modify our plans and execute the new plans. The concept of a goal is central to this formulation.

A theory of goals and planning can be applied not just to people but also to other entities that can be conceived of as agents, such as organizations and complex artifacts. In fact, it is a not uncommon cognitive move among people to attribute such agency even to natural phenomena like trees, volcanos and hurricanes. Anything that seems to exploit and manipulate the causal structure of the world as a means toward some supposed end can be and often is viewed as a planning mechanism.

The key concept in modeling intentional behavior is that of an agent a having some eventuality type e as a goal. The expression (goal e a) says that eventuality e is a goal of agent a. Normally, e will be an eventuality type that can be satisfied by any number of specific eventuality tokens, but it is entirely possible in principle for an agent to have an eventuality token as a goal, where there is only one satisfactory way for things to work out. We won't belabor the distinction here.

Agents know facts about what causes or enables what in the world, in most cases, facts of the form

```
(forall (e1 x)
   (if (p′ e1 x)
       (exist (e2)(and (q′ e2 x)(cause e1 e2)))))

(forall (e1 x)
   (if (p′ e1 x)
       (exist (e2)(and (q′ e2 x)(enable e1 e2)))))
```

That is, if e1 is the eventuality of p being true of some entities x, then there is an eventuality e2 that is the eventuality of q being true of x and e1 causes or enables e2. Or stated in a less roundabout way, p causes or enables q.

The agent uses these rules to plan to achieve goals, and also uses them to infer the goals and plans of other agents. A plan is an agent's way of manipulating the causal properties of the world to achieve goals, and these axioms express causal properties.

We will work step by step toward a characterization of the planning process. The first version of the axiom we need says that if agent a has a goal e2 and e1 causes e2, then a will also have e1 as a goal.

```
(forall (a e1 e2)
   (if (and (goal e2 a)(cause e1 e2))(goal e1 a)))
```

This is not a bad rule, and certainly is defeasibly true, but it is of course necessary for the agent to actually believe in the causality, and if the agent believes a causal relation that does not hold, e1 may nevertheless be adopted as a goal. The causal relation needn't be true.

```
(forall (a e0 e1 e2)
   (if  (and (goal e2 a)(cause′ e0 e1 e2)(believe a e0))
        (goal e1 a)))
```

We can say furthermore that the very fact that a has goal e2 causes a to have goal e1. We do this by reifying the eventuality g2 that e2 is a goal of a's, and similarly g1. (The e's in this axiom are the eventualities of having something; the g's are the eventualities of wanting it.)

```
(forall (a e0 e1 e2 g2)
   (if (and (goal′ g2 e2 a)(cause′ e0 e1 e2)(believe a e0))
       (exist (g1)(and (goal′ g1 e1 a)(cause g2 g1)))))
```

That is, if agent a wants e2 and believes e1 causes e2, that wanting will cause a to want e1. (The belief is also in g1's causal complex, but that would not normally be thought of as the cause: Why do you want e2? Because I want e1.)

Note that while the antecedent and the consequent no longer assert the real existence of having the goal (i.e., g2 and g1), if we know that g2 really exists, then the real existence of g1 follows from the properties of "cause".

Note also that the predicate goal reverses causality. For example, because flipping a light switch causes a light to go on, having the goal of the light being on causes one to want to flip the switch.

The eventuality e1 is a "subgoal" of e2 for a, and we can encode this in the axiom.

```
(forall (a e0 e1 e2 g2)
   (if (and (goal′ g2 e2 a)(cause′ e0 e1 e2)(believe a e0))
       (exist (g1)
          (and (goal′ g1 e1 a)(cause g2 g1)(subgoal e1 e2 a)))))
```

Finally, this axiom is not always true. There may be many ways to cause the goal condition to come about, and the mystery of the agent's free choice intervenes. The axiom is only defeasible. We can represent this by means of an "et cetera" proposition in the antecedent.

```
(forall (a e0 e1 e2 g2)
   (if (and (goal′ g2 e2 a)(cause′ e0 e1 e2)(believe a e0)(etc))
       (exist (g1)
          (and (goal′ g1 e1 a)(cause g2 g1)(subgoal e1 e2 a)))))
```

That is, if agent a has a goal e2 (where g2 is the eventuality of wanting e2) and a believes e1 causes e2, then defeasibly this wanting e2 will cause a to want e1 as a subgoal of e2 (where g1 is the eventuality of wanting e1).

A similar succession of axioms can be written for enablement.

In the STRIPS terminology of Fikes and Nilsson [4], the enabling conditions are the prerequisites of the plan operator, and the cause is the body.

The “subgoal” relation is a relation between two goals, and implies the agent’s belief that the subgoal is in a causal complex for the goal.

```
(forall (e1 e2 a)
   (if (subgoal e1 e2 a)
       (and (goal e2 a)(goal e1 a)
            (exist (e3 e4 e5 s)
               (and (causalComplex’ e3 s e2)(member’ e4 e2 s)
                    (and’ e5 e3 e4)(believe a e5))))))
```

In lines 5-6 of this axiom, e3 is the proposition that s is a causal complex for e2, e4 is the proposition that e2 is a member of s, e5 is the conjunction of these two propositions, and that’s what agent a believes.

The “subgoal” relation is transitive.

```
(forall (e1 e2 e3 a)
   (if (and (subgoal e1 e2 a)(subgoal e2 e3 a))
       (subgoal e1 e3 a)))
```

It will be useful below to state that if one believes he or she has a goal, then defeasibly he or she really does have the goal. Though not always true, we are usually pretty reliable about knowing what we want.

```
(forall (e e1 a)
   (if (and (goal’ e e1 a)(believe a e))
       (Rexist e)))
```

However, it is possible for an agent to have a goal without knowing it.

Goals do not have to be directly achievable by actions on the part of the agent, but successful plans have to bottom out in such actions or in states or events that will happen or hold at the appropriate time anyway.

4. Kinds of Goals

The most central cluster of goal-related concepts we discovered was a set of 27 concepts that characterized variations in the types of goals that people pursue, for example, “knowledge goals” as indicated by expressions such as “desire to understand, “curious”, “curiosity”, “inquisitive”, and “nosiness”. All of these types of goals can be defined in the machinery we have built up here and in our background theories. For example, a goal of knowing something, important because most actions have knowledge prerequisites, can be defined in our framework by the axiom

```
(forall (e a)
   (iff (knowledgeGoal e a)
        (exist (e1) (and (goal e1 a)(know’ e1 a e)))))
```

Similarly, you can have a goal to envision a situation in a certain manner, a goal to plan in a certain manner, or a goal to execute a plan in a certain manner.

Using our theory of time, we are also able to define goals to preserve situations, violations of such goals, goals that persist over time, goals that are achieved by a particular time, goals that are not achieved by some particular time, and goals that are never achieved. We are able to define conflicting goals and auxilliary goals that may be given up in the face of conflicts.

Individual persons are not the only kind of intentional agent. Sufficiently complex devices of various sorts can be viewed as intelligent agents. So can collectives of agents, whether collectives entirely of people, or assemblages of people and devices. These collectives can have goals. For example, General Motors has the goal of selling cars. They can devise plans for achieving goals. General Motors' plan involves manufacturing and marketing cars. These plans must bottom out in the actions of individual persons or devices, or in things that will be true at the appropriate time anyway.

For example, we can have a plan to get your car started by pushing it a short distance to the top of a hill and then letting it pick up speed on the downhill side until it is fast enough that you can pop the clutch. This plan bottoms out in my individual action of pushing on the back of the car and your individual action of pushing on the frame of the open left door. Of course these actions have to be synchronized, but these are properties of the individual actions. The plan also bottoms out in the event of the car rolling down the hill from the top. This is something that will happen anyway at the appropriate time, and doesn't have to be carried out by any member of the collective.

A shared or collaborative goal is a goal where the agent having the goal is a collective (cf. [24,12]). Moreover, the members of the collective mutually believe the collective has the goal.

```
(forall (e s e1)
   (iff (sharedGoal’ e e1 s)
        (exist (e0)
           (and (goal’ e0 e1 s)(mb s e0)(gen e e0)
                (forall (x)(if (member x s)(agent x)))))))
```

Since mutual belief implies belief and since if you believe you have a goal then you do really do have the goal, it follows that the individual members of s have e1 as a goal.

We can similarly define competitive goals and adversarial goals.

5. Trying, Succeeding, and Failing

When we try to bring about some goal, we devise at least a partial plan to achieve it, including subgoals of the original goal which are actions on our part, and we execute some of those subgoals. Moreover, our executing those actions is a direct result of our having those actions as subgoals. We can take this as a definition of "trying".

```
(forall (e a e1)
   (iff (try’ e a e1)
        (exist (e0 e2 e3 e4)
           (and (goal e1 a)(subgoal’ e3 e2 e1 a)
                (instanceOf e4 e2)(Rexist’ e0 e4)
                (agentOf a e4)(cause e3 e0)(gen e e0)))))
```

In this definition, `e` is the eventuality of an agent `a` trying to do `e1`. The eventuality (or eventuality type) `e1` is a goal of `a`'s; it's what `a` tries to do. The eventuality type `e2` is a subgoal of `e1`, and `a` is the agent in action `e4` which is an instance of `e2`. In the expression `(Rexist’ e0 e4)`, `e0` refers to the actual occurrence or execution of the action `e4` by `a`. The expression `(cause e3 e0)` says that `a`'s having the subgoal `e2` causes `a` to actually do `e2`. The expression `(gen e e0)` says that the actual doing (`e0`) constitutes the trying (`e`). That is, `a` tries to do `e1` exactly when `a` executes a subgoal `e2` of `e1` precisely because it is a subgoal.

For example, suppose you want to pass a course. An important subgoal would be to study for the final. If you study for the final precisely because it will help you pass the course, then you are trying to pass the course.

It follows that if you try to do something e1, then you perform some action or actions in a causal complex whose effect is an instance of e1.

```
(forall (a e1)
   (if (try a e1)
       (exist (s e2 e3)
          (and (instanceOf e3 e1)(causalComplex s e3)
               (member e2 s)(agentOf a e2)(Rexist e2)))))
```

This is not sufficient as a definition of trying. The causal role of having `e2` as a subgoal is necessary in the definition of `try`. Suppose Pat wants to meet Chris. One way for that to happen is for them to run into each other someday. Now Pat is driving to the grocery store where Chris, unbeknownst to Pat, is currently shopping. Pat's driving to the grocery store is an element of a causal complex leading to the two of them meeting. But we wouldn't say that Pat's driving to the grocery store constitutes an attempt to meet Chris. That wasn't his intent in driving.

We can talk about trying and failing at a seemingly executable action, as in "He tried to lift his arm." But we are refining the granularity and viewing the action as a composite, e.g., in which the first subaction is willing the motion.

Note that this explication of trying allows for an agent to try to achieve conflicting goals. The student who studies for a final and then goes out drinking the whole night before the final may be doing just that.

To succeed at some goal is to try to do it and to have that trying cause the goal to actually occur.

```
(forall (e a e1)
   (iff (succeed' e a e1)
        (exist (e0 e2 e3)
           (and (try' e2 a e1)(instanceOf e3 e1)(cause e2 e3)
                (Rexist' e0 e3)(gen e e0)))))
```

Here, `e2` is the attempt, `e1` is ultimate goal, `e3` is the specific instance of `e1` that occurs, `e` is the success in achieving the goal, and `e0` is the actual occurrence of the instance of the goal. The expression `(gen e e0)` says that the actual occurrence `e0` constitutes the event `e` of succeeding.

Succeeding implies trying. The converse, sadly, is not true.

To fail at some goal is to try to do it and for it not to actually occur.

```
(forall (a e1)
   (iff (fail a e1)
        (and (try a e1)
             (not (exist (e2)
                     (and (instanceOf e2 e1)(Rexist e2)))))))
```

There is space between succeeding and failing. One can try to achieve something and that something can come about but not because of one's efforts. In that case, the agent has lucked out, rather than having succeeded. So a student who studied for the final but nevertheless got a 0 on it, but passed the course because the professor gave everyone an A as a political protest, wouldn't be said to have succeeded at passing the course.

6. Functionality

Complex artifacts are generally constructed for some purpose. Some real or hypothetical agent has a goal in mind, and the artifact achieves that goal. Cars, for example, have at least the purpose of moving us from place to place. The structure of complex artifacts typically reflects a plan to achieve the goal, where the various components are involved in some subgoal. For example, the steering wheel of a car is involved in the subgoal of having the car go in particular directions. We will call such a subgoal the functionality of the component. Organizations and their components can be analyzed similarly.

In fact natural objects can too, if we associate a hypothetical agent having as a goal the normal behavior the natural object tends to engage in. We can stipulate that the "goal" associated with a tree is to grow and reproduce, and we can analyze the structure of the tree as an instantiation of a plan to achieve that goal. We can then talk about the function of the various parts of the tree. We can even view volcanos, for example, as composite entities with the "goal" of erupting, and talk about the functions of its parts to this end.

In general, almost any composite entity can be associated with a goal by some hypothetical agent, and where components are causally involved with the behavior of the whole, we can view the relation between an action by the component and the behavior of the whole as a "subgoal" relation. We can then define the functionality of the component as that "subgoal" relation.

Plans to achieve goals are just a way of exploiting the causal structure of the world to achieve some end, so "goal talk" can be imported into any domain where the manipulation or exploitation of causal structure is involved.

A composite entity that goes through changes can be viewed as having as a goal some element or subset of those changes. We will define that "goal" as the functionality of the whole. We call this predicate `functionality0` because it is absolute functionality, not the functionality of a component relative to the behavior of the whole.

```
(forall (x e)
   (if (and (compositeEntity x)(changeIn’ e x))
       (iff (functionality0 e x)
            (exist (e2 e3 a))
               (and (goal’ e2 e a)(agent’ e3 a)(imply e3 e2)))))
```

Line 2 says that `x` is a composite entity and `e` is one of `x`'s behaviors. Under these conditions, `e` is a functionality of `x` if and only if there is some possible agent a whose existence would imply that `e` is a goal of `a`'s. The roundabout formulation in line 5 is a way of allowing hypothetical agents. In any positive modality in which `a` exists, `a` will have goal `e`.

If a component has some property or engages in some behavior that is in a causal complex for the functionality of the composite entity as a whole, we can say that property or behavior is the functionality of the component with respect to the `functionality0` of the whole.

```
(forall (e1 y e x)
   (iff (functionality e1 y e x)
        (exist (s)
           (and (functionality0 e x)(causalComplex s e1)
                (member e1 s)(arg* y e1)(componentOf y x)))))
```

The expression (functionality e1 y e x) says that e1 is the functionality of y with respect to behavior e by composite entity x, of which y is a component. Because of the close connection between the subgoal relation and causal complexes, functionality can be viewed as a close analog of the subgoal relation.

A component is “intact” if it is able to fulfill its functionality. That is, there are no properties of the component that would cause the functionality not to occur.

```
(forall (x)
   (iff (intact x)
        (forall (e1 y e)
           (if (functionality e1 y e x)
               (not (exist (e2 e3)
                       (and (arg* y e2)(not’ e3 e1)
                            (cause e2 e3))))))))
```

7. Thriving

It is formally convenient to assume that agents have one plan that they are always developing, executing, monitoring and revising, and that that plan is in the service of a single goal. We will call this goal “Thriving”.

```
(forall (a)
   (if (agent a)
       (exist (e)(and (goal e a)(thrive’ e a)))))
```

More specific goals arise out of the planning process using the agents’ beliefs about what will cause them to thrive.

The main reason for positing this top-level goal is that now instead of worrying about the mysterious process by which an agent comes to have goals, we can address the planning problems of what eventualities the agent believes cause other eventualities, including the eventuality of thriving, and of what alternative subgoals the agent should choose to achieve particular goals. We are still left with the problem of when one goal should be given priority over another, but this is now a plan construction issue.

We will not attempt to say what constitutes thriving in general, because there are huge differences among cultures and individuals. For most of us, thriving includes staying alive, breathing, and eating, as well as having pleasurable experiences. But many agents decide they thrive best when their social group thrives, and that may involve agents sacrificing themselves. This is a common view in all cultures, as seen in suicide bombers, soldiers going into battle in defense of their country, and people risking death to rescue accident victims. So thriving does not necessarily imply surviving.

Similarly, a man may decide that he is in so much pain that the best way to thrive is to kill himself. In contrast, a religious ascetic may decide that the best way to achieve the long-term goal of eternal life is to live in pain.

A good theory of commonsense psychology should not attempt to define thriving, but it should provide the materials out of which the beliefs of various cultures and individuals can be stated in a formal manner.

8. Importance

In [16] we define scales in terms of partial orderings. Many scales, including the scale of importance, cannot be defined precisely, but constraints can be placed on their partial ordering. That is what we will do here.

A concept, entity or eventuality is more or less important to an agent depending on its relation to the agent's goals. The "more important" relation is thus a partial ordering that depends on the agent.

```
(forall (x1 x2 a)
   (if (moreImportant x1 x2 a)(and (neq x1 x2)(agent a))))
```

The expression `(moreImportant x1 x2 a)` says that something `x1` is more important than something else `x2` to agent `a`. We place no constaints on the things `x1` and `x2` whose importance is being measured. They can be anything.

The "more important" relation is, at least defeasibly, transitive.

```
(forall (x1 x2 x3 a)
   (if (and (moreImportant x1 x2 a)(moreImportant x2 x3 a))
       (moreImportant x1 x3 a)))
```

An agent proceeds through the world by continually developing, executing and modifying a plan to achieve the top-level goal "To Thrive", All of the agent's actions can be seen as subgoals in this plan; when the actions are dysfunctional, we can see them as part of a plan based on false beliefs about what will result in thriving. A plan can be thought of as a tree-like structure representing the `subgoal` relation. The higher a goal is in a plan, the more important it is, because of the greater amount of replanning that has to be done if the goal is not to be achieved. So the first constraint we can place on the importance scale is that it is consistent with the subgoal relation.

However, this is a bit tricky to specify because an eventuality can be a subgoal of a number of different higher-level goals in the same plan, and we do not want to say an eventuality is of little importance simply because one of its supergoals is of little importance. So we first need to define the notions of an "upper bound supergoal" and a "least upper bound supergoal". An eventuality `e1` is an upper bound supergoal of `e2` if it is a supergoal of all of `e2`'s immediate supergoals. More precisely, any supergoal of `e2`'s must either be `e1`, be a subgoal of `e1`, or be a supergoal of `e1`. It will be convenient to define the upper bound for a set of subgoals.

```
(forall (e1 s a)
   (iff (ubSupergoal e1 s a)
        (and (agent a)(goal e1 a)
             (forall (e2) (if (member e2 s)(subgoal e2 e1 a)))
             (forall (e2 e)
                 (if (and (member e2 s)(subgoal e2 e a))
                     (or (subgoal e e1 a)(eq e e1)
                         (subgoal e1 e a)))))))
```

The expression `(ubSupergoal e1 s a)` says that `e1` is an upper bound supergoal of all the goals of agent `a` in set `s`. Lines 3-4 specify the conditions on the arguments of the predicate. The predicate holds if and only if any eventuality `e` which is a supergoal of a member `e1` of `s` is either a subgoal of `e1`, `e1` itself, or a supergoal of `e1`.

A goal `e1` is a least upper bound supergoal if it is an upper bound supergoal and a subgoal of all other upper bound supergoals.

```
(forall (e1 s a)
   (iff (lubSupergoal e1 s a)
        (and (ubSupergoal e1 s a)
             (forall (e)
                (if (ubSupergoal e s a)
                    (or (eq e e1)(subgoal e1 e)))))))
```

Because every goal is ultimately in the service of the top-level goal “To Thrive”, every goal has a least upper bound supergoal.

Now we can say that if eventuality `e1` dominates eventuality `e2` on every path in the agent’s plan that includes `e2`, then `e1` is more important than `e2`. Every reason for wanting `e2` is in the service of `e1`.

```
(forall (s e1 e2 a)
   (if (and (singleton s e2)(lubSupergoal e1 s a))
       (moreImportant e1 e2 a)))
```

More generally,we can say that the least upper bound supergoal of a set of goals is more important than the whole set, since all the members of the set are in the service of the supergoal.

```
(forall (s e a)
   (if (lubSupergoal e s a)(moreImportant e s a)))
```

An agent’s goals are important. So are eventualities that affect the agent’s goals. Importance doesn’t care about polarity; if passing a course is important to you, so is not passing the course. Thus, we define an eventuality as “goal-relevant” to an agent if its existence implies the existence or nonexistence of one of the agent’s goals.

```
(forall (e a)
   (iff (goalRelevant e a)
        (exist (e1)
           (and (goal e1 a)
                (or (imply e e1)
                    (exist (e2)(and (not’ e2 e1)(imply e e2))))))))
```

The “goal consequences” of an eventuality are those goals of the agent’s whose existence or nonexistence is implied by by the eventuality.

```
(forall (s e a)
   (iff (goalConsequences s e a)
        (forall (e1)
           (iff (member e1 s)
                (and (goal e1 a)
                     (or (imply e e1)
                         (exist (e2)
                            (and (not’ e2 e1)(imply e e2))))))))
```

Then we can say the importance of an eventuality depends on the importance of its goal consequences. The first of the following axioms says that if something `x` is more important than the goal consequences of eventuality `e`, then it is more important than `e`. The second axiom says the opposite.

```
(forall (x s e a)
   (if (and (moreImportant x s a)(goalConsequences s e a))
       (moreImportant x e a)))

(forall (x s e a)
   (if (and (moreImportant s x a)(goalConsequences s e a))
       (moreImportant e x a)))
```

In a more complete theory of importance, we would relate the importance of an eventuality to its effect on the likelihood of the agent's goals obtaining or not.

The importance of an entity depends on the importance of its properties and of the events it participates in. Thus, we define the set of "goal-relevant properties".

```
(forall (s x a)
   (iff (grProps s x a)
        (forall (e)
           (iff (member e s)
                (and (arg* x e)(goalRelevant e a))))))
```

The expression (grProps s x a) says that the set s of properties of x are relevant to a goal of a's.

The next two axioms say that the importance of an entity depends on the importance of its goal-relevant properties.

```
(forall (s x a)
   (if (and (moreImportant x1 s a)(grProps s x2 a))
       (moreImportant x1 x2 a)))

(forall (s x a)
   (if (and (moreImportant s x1 a)(grProps s x2 a))
       (moreImportant x2 x1 a)))
```

To summarize, x1 is more important than x2 to a if x2 is, or affects something that is, or has properties that affect something that is, in the service of x1. Note that there may be other properties constraining the moreImportant relation, but this one at least is among the most significant.

9. Summary

In this paper we have developed a formal treatment of an agent's goals. We have shown how a variety of goal-relevant concepts that arise in natural language and in strategic planning can be explicated in terms of this. Moreover, we have provided at least a first pass at axiomatically characterizing such concepts as trying, succeeding, failing, functionality, intactness, and importance. These concepts are critical in modeling human behavior and also in the development of more human-like computational agents.

Acknowledgments

This research was supported in part by the Office of Naval Research, under Contract No. N00014-09-1-1029 and Contract No. NOOO14-09-C-0613.

References

[1] Atkinson, J. (1964) *An introduction to motivation*, Princeton, NJ: Van Nostrand.

[2] Chulef, A., Read, S., and Walsh, D. (2001) A hierarchical taxonomy of human goals. *Motivation and Emotion* 25(3):191-232.

[3] Eccles, J. and Wigfield, A. (2002) Motivation beliefs, values, and goals. *Annual Review of Psychology* 2002(53):109-32.

[4] Fikes, R., and Nilsson N. (1971) STRIPS: A new approach to the application of theorem proving to problem solving, *Artificial Intelligence*, 2: 189-208.

[5] Ford, M. (1992) *Human motivation: Goals, emotions, and personal agency beliefs*. Newbury Park, CA: Sage.

[6] Ford, M. and Nichols, C. (1987) A taxonomy of human goals and some possible application. In M. Ford and D. Ford (Eds.) *Humans as self-constructing living systems: Putting the framework to work*, pp.289-311. Hillsdale, NJ: Erlbaum.

[7] Gordon, A. (2002) The Theory of Mind in Strategy Representations. *Proceedings*, Twenty-fourth Annual Meeting of the Cognitive Science Society (CogSci-2002), George Mason University, Aug 7-10. Mahwah, NJ: Lawrence Erlbaum Associates.

[8] Gordon, A. (2004) *Strategy Representation: An Analysis of Planning Knowledge*. Mahwah, NJ: Lawrence Erlbaum Associates.

[9] Gordon, A. and Hobbs, J. (2003) Coverage and Competency in Formal Theories: A Commonsense Theory of Memory. *Proceedings*, 2003 AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning, Stanford University, March 24-26, 2003.

[10] Gordon, A., Kazemzadeh, A., Nair, A., and Petrova, M. (2003) Recognizing Expressions of Commonsense Psychology in English Text. *Proceedings*, 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003) Sapporo, Japan, July 7-12, 2003.

[11] Hobbs, J. (1985) Ontological promiscuity, *Proceedings*, 23rd Annual Meeting of the Association for Computational Linguistics, Chicago, Illinois: 61-69.

[12] Hobbs, J. (1990) Artificial intelligence and Collective intentionality, in P. Cohen, J. Morgan, and M. Pollack, *Intentions in Communication*, Cambridge: MIT Press: 445-460.

[13] Hobbs, J. (1995) Monotone decreasing quantifiers in a scope-free logical form, in K. van Deemter and S. Peters (Eds.), *Semantic Ambiguity and Underspecification*, CSLI Lecture Notes No. 55, Stanford, California: 55-76.

[14] Hobbs, J. (2003) The logical notation: Ontological promiscuity, `http://www.isi.edu/ hobbs/disinf-tc.html`

[15] Hobbs, J. (2005) Toward a useful concept of causality for lexical semantics, *Journal of Semantics*, 22: 181-209.

[16] Hobbs, J. (2005) Encoding commonsense knowledge, `http://www.isi.edu/ hobbs/csk.html`.

[17] Hobbs, J. and Gordon, A. (2005) Encoding Knowledge of Commonsense Psychology. 7th International Symposium on Logical Formalizations of Commonsense Reasoning. May 22-24, 2005, Corfu, Greece.

[18] Hobbs, J. and Gordon, A. (2008) The Deep Lexical Semantics of Emotions. Workshop on Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology (EMOT-08), 6th International conference on Language Resources and Evaluation (LREC-08), Marrakech, Morocco, May 27, 2008.

[19] Hobbs, J. and Pan F. (2004) An ontology of time for the Semantic Web, *ACM Transactions on Asian Language Information Processing*, 3(1): 66-85.

[20] McCarthy, John, (1980) “Circumscription: A Form of Nonmonotonic Reasoning”, *Artificial Intelligence*, 13: 27-39.

[21] Moskowitz, G. and Grant, H. (Eds.) (2009) *The psychology of goals*. New York: The Guilford Press.

[22] Murray, H. (1938) *Explorations in personality*. New York: Oxford University Press.

[23] Rokeach, M. (1973) *The nature of human values*. New York: Free Press.

[24] Searle, J. (1990) Collective intentions and actions, in P. Cohen, J. Morgan, and M. Pollack, *Intentions in Communication*, Cambridge: MIT Press: 401-416.

[25] Weiner, B. (1985) An attributional theory of achievement motivation and emotion. *Psychological Review* 92:548-73.

[26] Wicker, F., Lambert, F., Richardson, F., and Kahler, J. (1984) Categorical goal hierarchies and classification of human motives. *Journal of Personality* 52(3):285-305.