AGENT HONESTY, COOPERATION AND BENEVOLENCE IN AN ARTIFICIAL ORGANIZATION

Kathleen Carley*
David Park
Michael Prietula*

*Carnegie Mellon University
Pittsburgh, PA 15213

December 22, 1993

Appears in proceedings of Workshop on AI and Theories of Groups & Organizations: Conceptual & Empirical Research.
Washington, DC, July 11-15, 1993.

Agent Honesty, Cooperation and Benevolence in an Artificial Organization

Kathleen Carley
David Park
Michael Prietula

Carnegie Mellon University
Pittsburgh, PA 15213

1. Introduction

Organizations are composed of individuals. However, organizational behavior is not simply a sum of the behaviors of the individuals comprising them. Organizations alter individual behavior by constraining who interacts with whom, as well as constraining the subjects motivating their interactions [March & Simon, 1958], what resources are available [Pfeffer & Salancik, 1978], providing new motivations and inducements for behavior, such as group goals [Arrow & Radner, 1979], and so forth. Nevertheless, an organization's behavior is not independent of the input of its members. Their collective actions shape organizational behavior. Thus, there are two phenomenological horizons at issue simultaneously.

One deals with the capabilities and behaviors of individuals as they coordinate to solve problems while the other is concerned with the manner in which organizations formulate solutions to the problems that they are confronted with. The relation between individual behavior and action, and the behavior of the organization to which they belong is complex. However, much of traditional organization theory fails to take individual members into account.

In contrast, artificial intelligence and cognitive psychology strive to provide an understanding of the principles underlying individual human cognitive processes. A method brought to bear in this search is the examination of theories by casting them in the form of computer programs [Simon, 1981].

This research was supported by NSF grants SES-87070005 and IRI-9111804.

One type of theory involves the specification of symbolic architectures of cognition [Newell, Rosenbloom & Laird, 1989]. We are investigating the extent to which aggregations of agents (realized as separate agents) can be used as a research mechanism for studying organizational phenomena. Thus, we are taking a "bottom up" approach to examining organizations of intelligent agents. The particular symbolic architecture we are employing is Soar [Laird, Newell & Rosenbloom, 1987].

The primary reason for using Soar is that it reflects a comparatively complete general architecture for reasoning and problem solving. This, in part, allows us to employ a substantially sufficient "theory of the agent" in the form of a computer program. The capabilities of the agent are then defined in terms of the knowledge incorporated into the Soar architecture. The capabilities of the organization are defined by the aggregate capabilities of the agents which comprise it. Some capabilites directly reflect observable organizational events (e.g., agent communication) while others are endogenous to the agents (e.g., deliberation time) and less visible, though perhaps no less important, in terms of their effect on organizational events (e.g., indirectly, through their effect on actions affecting other agents).

In this paper we describe an exploratory experiment in which we examined the effect of three agent social behaviors (Cooperativity, Reliability, Benevolence) on four measures of organizational performance: cognitive effort, physical effort, communication effort, and idle time.

Workshop on AI and Theories of Groups & Organizations: Conceptual & Empirical Research, Working Notes, Eleventh National Conference on Artificial Intelligence, Washington DC, July 11-15, 1993.

2. The Plural-Soar Architecture

In a series of prior studies [Carley, Kjaer-Hansen, Newell & Prietula, 1992; Prietula & Carley, 1993], multiple Soar agents were created to perform a task requiring the retrieval of requested items from a (virtual) warehouse. The resulting system was called Plural-Soar and was run with each Plural-Soar agent residing on a single workstation, with all agents communicating over Ethernet connections.

2.1 Soar

As noted, Soar is a symbol oriented architecture that has its historic lineage in the theoretical constructs of symbols systems the problem space formulations of Newell and Simon [Newell & Simon, 1976]. Soar, however, not only instantiates these particular constructs, but proposes simple, but universal, mechanisms for learning (chunking) and goal-driven behavior which guides learning (impasse-driven subgoaling).

All Soar behavior is characterized as a *search* through *problem spaces* in service of satisfying *goals*. Soar acts through a series of *decision cycles*, based on the current state of working memory (collections of symbol structures), permanent memory (if-then production rules that potentially match symbol structure patterns in working memory and, when matched, propose preferences to modify those structures), and a preference memory (which collects preferences proposed by activated permanent memory productions).

Soar behavior is goal-driven and Soar productions are crafted with that in mind; that is, Soar productions always have a particular goal context residing in their left-hand side. Thus, decision cycles in Soar reflect attempts to find operators (in a particular problem space) that may be applied to achieve a particular goal. If a satisfactory operator cannot be found (for a variety of reasons), an *impasse* occurs. When an impasse occurs (there may be several types), the Soar architecture automatically generates a new goal to resolve the impasse. This results in the engagement of a new (or different) problem space and the knowledge (as operators) contained in it. This *automatic subgoaling* may proceed recursively until sufficient knowledge is engaged (or discovered) to satisfy (or perhaps reject) the original motivating goal.

The basic activity unit in Soar is the *decision cycle*. A decision cycle reflects a single deliberative cognitive act. Though decision cycles may potentially vary in the "clock time" required to execute them, they can serve as nominally calibrated metrics for comparison. Within a decision cycle, Soar examines the contents of working memory and explores the space of potential next states by determining the implications of all relevant (permanent) knowledge. As either contradictory or competing states are possible from this method, the Soar architecture employs special preference semantics for determining appropriate action.

Soar, however, is a symbol level system; that is, it provides the fundamental architectural mechanisms for symbolic processing. In order to have Soar "to a task," it is necessary to code task knowledge in the form of Soar productions. In our case, the collection of Soar productions is called Plural-Soar.[1]

2.2 Plural-Soar

Plural-Soar itself is a group of agents that perform a warehouse task. The task involves agents proceeding to a particular location (an Order_Stack), selecting an order (possibly waiting in line with other agents), determining where that item may be in the warehouse

[1]Plural-Soar actually has additional Lisp code to permit inter-agent communication and virtual task representation on the network.

(given a row of identifiable Item_Stacks). In Plural-Soar, agents may have memories of the contents of the Item_Stacks they have encountered as well as the ability to broadcast requests (and responses) of particular item locations. The knowledge defined in each Agent specifies possible actions to engage (e.g., ask a question, go to a location) as well as knowledge that determines how to resolve ambiguous choices of actions.

In this paper we extend Plural-Soar by adding elements to the agents in order to begin to bring social components of organizations to Plural-Soar.

2.3 Extending Plural-Soar

Work relating these two perspectives -- individual problem solving and organizational problem solving -- is still in its infancy. One approach toward this end explores group problem solving by several agents working together in an environment in which they may interact "socially." Carley and Newell [1990] define a social agent in terms of the attributes that such an agent must possess and present a matrix using two dimensions. The first dimension defines a decreasing information processing ability: an omniscient agent (relative to the task), a rational agent, a boundedly rational agent, a cognitive agent, and an emotional-cognitive agent. The second dimension describes the possible increasingly detailed knowledge of a (rich) social environment: non-social, multiple agents, interactive multiple agents, social structures, social goals, and cultural history. The two dimensions delineates the potential relations between artificial agents of greater and lesser social ability. This, in a sense, outlines a set of increasingly inclusive categories that progressively define what a social agent must be and must be able to do.

Currently, most AI systems fall short of the model social agent, being neither limited enough in cognitive capability nor rich enough in social knowledge. Individual AI agents typically work in isolation to solve problems of varying complexity. Given the Carley-Newell model, if we are to understand organizations as collections interacting of intelligent agents, we need to use agents that are nearer the model social agent within the sort of problem solving environment first used with Plural-Soar.

3. A Minimal Social Agent

In expanding Plural-Soar to build a minimal social agent, the first step was to incorporate a *social memory* into the agent. This social memory took the form of a structure defined in working memory in which symbolic descriptions of simple, though specific, properties of other agents were stored. The social memory was used by the agent to judge the to judge the reliability of information obtained from other agents.

If an agent requested information from other agents on the stack location of a specific item, any responses from other agents were remembered. When the agent eventual found the item, it noted the stack location and re-examined the information obtained from all the other agents regarding item location and updated its social memories on the *reliability* of the agents.

An agent's social memory of another agent's reliability took on one of three values: reliable, possibly reliable, and unreliable. If an agent's information was accurate, then the representation in social memory was "upgraded" (if currently unreliable or possibly reliable) or "confirmed" (if currently reliable). Conversely, if an agent's information was inaccurate, then the representation of the reliability of the agent was "downgraded" in a similar fashion.

If information from any agent has been judged as "unreliable" in two consecutive

evaluations, then the informing agent itself is judged as "unreliable" and any further information is rejected without review.

In the original Plural-Soar design, communication from other agents was accepted unconditionally. In this version, communication from other agents are accepted or rejected on the basis of social memory. If an agent received conflicting information from different agents, then the agent that has the highest reliability rating is selected. If there is still a tie, then other types of preferences are brought into play, such as preferring the nearest location.

4. A Minimal Social Agent Experiment

In order to test the extremes of a social situation, we incorporated three social characteristics into the social agent: honesty, cooperativity, and benevolence. Each characteristic had to opposing values (i.e., honesty vs. lying, cooperative vs. selfish, benevolent vs. nonbenevolent).

4.1 Agent Properties

An *honest* agent will always answer to the best of its knowledge. A lying agent always gives a false location to a query if it knows the true location (whether from memory or from observing it directly). In this study, all organizations are uniform in honesty (but were varied on the other characteristics); therefore, all agents in an organization are either liars or honest.

The warehouse problem was expanded in order to reduce the probability of misattribution of agents. This could occur when an (honest) agent passes location information that is no longer accurate due to an intervening agent moving the item in question in order to remove another item. Similarly, misattribution of a liar as an honest agent could occur if the item in question is moved onto the actual location before the asking agent arrives there. Whereas the problem used in Plural-Soar has 15 items over 10 stacks, the expanded problem used here has 15 single item orders and 20 item stacks with 3 items each. Thus, there are 30 additional filler items in the expanded problem, so that when an interfering item is moved, it will be filler (i.e., non-requested) item.

A *cooperative* agent would choose to help others by answering other agents' questions before it helps itself (by moving itself or an item to fill an order). *Selfish* agents would help themselves (e.g., by searching, removing an item to the conveyor belt) before helping anyone else (by answering a question).

Finally, a *benevolent* agent reflected the degree to which it "forgave" an agent that provided wrong information.

A benevolent agent would upgrade its opinion (i.e., social memory) of previously classified liars (i.e., those that have provided two unreliable pieces of information in a row), giving them the benefit of the doubt after rejecting one message. The next communication would be accepted if it is true and the sending agent's reliability rating would be upgraded according to the previously defined algorithm. In principle, a given agent may be classified differently over many different trials. On the other hand, a nonbenevolent agent was unforgiving in the sense that once it identified another agent as a unreliable, it never again changes that rating — the classification of another agent as a liar is an absorbing state.

These traits were selected because they characterize individual behavior that is observable within an organization by other agents and that is likely to affect organizational outcomes. In an organization where communication is a beneficial component of the task, agents must rely on the word of other agents and it becomes imperative that agents know

and act on whether or not other agents are reliable sources of information. Furthermore, varying the extent to which an agent judges another agent's behavior harshly or not, or how a listening agent adjusts its own behavior in the presence of a request from the group, offers minor, but valid, variations on the theme.

4.2 Design of the Study

Several measures were adapted from Plural-Soar. The time measure is achieved by assuming one *decision cycle* (Soar's basic event unit) was sufficient to complete a fundamental deliberative act, whether that act was moving an item, moving the agent itself, asking a question, consulting the agent's stack memory, or any other (physical or cognitive) act available to the agent. These decision cycles were summed for each agent in an organization and the maximum of these sums is taken to be the *total time* the organization spends in solving the problem.

Cognitive effort is measured as the number of decision cycles that an agent incurred in the course of solving its part of the problem.

In addition to these measures, the number of agent movements and the number of times an agent moved an item were tracked. Two measures are derived from these: physical and communication effort. *Physical effort* is the sum of the number of agent movements, item movements, item removals, and orders taken by an agent. It is thus an indication of how hard an agent is working to manipulate the shared warehouse environment.

Communication effort is the sum of the number of times an agent asked a question, answered a question, evaluated an answer, generated an answer, considered whom it should ask for information, updated its belief rating of another agent, and chose to accept or reject an answer from another agent. It gives a sense for the amount of cognitive effort that is expended in generating, processing, transmitting, receiving, and using communications with other agents. Thus, it serves as a measure of the effort involved in dealing with the social environment within the warehouse.

Wait Time indicates the amount of idle time the agent spent simply waiting for other events to happen, usually while in a queue behind another agent.

For each of the eight cells in the 2 x 2 x 2 design, five different organizations were simulated. These differed in the number of agents comprising the organization: one to five agents. Runs for each of the 40 simulations ranged from 45 minutes to several hours. All runs were conducted on linked Decstations 3500 and 5000 workstations. Analysis of network timing variations revealed no significantly stochastic effects on model dynamics, so fluctuations and timing variations were judged to be considered random error distributed normally. All agents were written in the Lisp-based Soar5 environment.

4.3 Results

In terms of *total time* to complete the task, the trend of diminishing returns to scale (also found in Carley et al., 1992) is recapitulated by organizations of sizes up to 4 agents — the more agents in the organization, more quicker things get done. However, most of the 5 agent organizations show a clear upward move producing a U-shaped curve. This result occurs because as agents become more complex, larger organizations put a greater cognitive load on each agent, and force it to deal with added communication and social considerations. In the warehouse, as the number of agents increases, the task for each agent decreases. Nevertheless, the social nature of the agents eventually forces the total time spent on the problem

to increase as the size of the organization increases.

The single exception to this trend is the honest, noncooperative, benevolent organization type. Of the eight types of organizations, this is the one that would tend to put the least cognitive strain on its members due to social and communication effects. Selfish (i.e., noncooperative) agents are the least social. They minimize communication. When faced with reliable answers, the answer evaluation process is streamlined because in this case, there is no need to evaluate answers from several other agents along the (expected) truthfulness dimension. If everyone told the truth all the time, the level of deliberative effort regarding social reliability would be greatly reduced if not eliminated. Similarly, in an organization of truth-tellers, the answer generation process for each agent is streamlined because the agents do not have to expend the extra effort to generate a lie about an item's location.

The trend toward a U-shaped curve is also found when evaluating *cognitive effort*. The values among 5-member organizations (where the social effects should be greatest) for honest organizations tend to fall below those for organizations of liars. This would be expected if the communication and social cognitive strains were responsible for the upward trend. The total cognitive effort for 1-agent organizations of all types is the same. This tends to validate the notion of social effects in the simulation — in the absence of a social environment, there is no one with whom the agent can interact. Thus, the capabilities and characteristics that are intended to deal with social problems are never invoked and differences in social styles should be inconsequential, as indeed, they are.

Diminishing returns to scale in *physical effort* that increasing organization size produces were found. Recall that physical effort is a measure of only the "physical" actions that agents take in interacting with their shared warehouse environment. The results of all eight organizations are essentially identical. Thus, all of the variation observed between organization types must be due to factors other than the physical actions agents took in solving the warehouse problem. Thus, the differences might be thought of as a form of social overhead which the different organization types must negotiate in the process of solving the common warehouse problem. The fact that measures of physical effort are invariant illustrates the orthogonality of these social characteristics and the purely physical considerations involved in completing the warehouse task, lending further credence to the measure definition.

Although physical effort is invariant across organization types, the details of organizational dynamics are not. As agents in different organizations follow different patterns of movement and communication, different patterns of waiting will emerge as well. This result again recapitulates the results of Plural-Soar well. The general trend of *wait times* is to increase monotonically with increasing organization size. As organization size increases, so too would the idle time that each agent on average spends in queues waiting to manipulate a stack.

Communication effort was found to increase as organization size increases. This lends further support to the assertion that social dynamics are largely responsible for the upward tending tail in the U-shaped curve of time and total cognitive effort. These results suggest that the U-shape responses for total cognitive effort and time to completion is a result of the linear combination of physical effort, which displays diminishing returns to scale, and waiting time and communication effort, both of which tend to increase monotonically.

5. Conclusions

In this paper we have used computer simulation to examine specific components of a virtual organization. This is an instance of incorporating simulation in general for theory development [Carley & Prietula, in press]. The main results largely confirm and extend those found by Carley et al. [1992] for Plural-Soar. Specifically, the orthogonality of social effort and physical effort was clearly demonstrated. The former arises from the interaction between agents while the latter is connected solely with directly solving the common problem in the "physical space" of the warehouse. In addition, the U-shaped curve for *total cognitive effort* reported by Carley et al. [1992] for their most sophisticated agent organizations was also found in the present work. This U-shape was shown to be the result of a combination of three curves, one monotonically decreasing, the others monotonically increasing with organization size. The former curve was for *physical effort*, which became less on average for the organization as the number of agents working conjointly to solve the problem increased. The second and third curves were for *communication effort* and *wait time*. Each of these measures increased with increasing numbers of agents, since as the organization grew in size, not only did each agent have to devote more effort to coordinating its own activities with the others, but the chances that each agent would have to wait behind another agent at a stack increased as well. Thus, the U-shaped curve was shown to result from the combination of *physical effort* directed at solving the common physical problem and social overhead, comprising *communication effort* and *waiting time*.

6. References

Arrow, K.J. & Radner, R. [1979]. Allocation of resources in large teams. *Econometrica*, 47, 361-85.

Carley, K., Kjaer-Hasen, J., Newell, A. & Prietula, M. [1992]. Plural-Soar: A proglegomenon to artificial agents and organizational behavior. In M. Masuch & M. Waglien (Eds.), *Artificial intelligence in organization and management theory*. New York, NY: North-Holland.

Carley, K. & Newell, A. [1990]. *On the nature of the social agent*. Paper presented at the American Sociological Association annual meeting, Washington, DC.

Carley, K. & Prietula, M. (Eds.). [in press]. *Computational organization theory*. Hillsdale, NJ: Lawrence Erlbaum.

Laird, J., Newell, A. & Rosenbloom, P. [1987]. Soar: An architecture for general intelligence. *Artificial Intelligence*, 33, 1-64.

Newell, A., Laird, J. & Rosenbloom, P. [1989]. Symbolic architectures for cognition. In M. Posner (Ed.), *Foundations of Cognitive Science*. Cambridge, MA: MIT Press.

Newell, A. & Simon, H. [1976]. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3), 113-126.

March, J. & Simon, H. [1958]. *Organizations*. New York, NY: Wiley.

Pfeffer, J. & Salancik, G. [1978]. *The external control of organizations: A resource dependency perspective*. New York, NY: Harper & Row.

Prietula, M. & Carley, K. [1993]. *Computational organization theory: Autonomous agents and emergent behavior*. Paper submitted for publication.

Simon, H. [1981]. Studying human intelligence by creating artificial intelligence. *American Scientist*, 69(3), 300-309.