Abstract Models of Human-AI Teams

Pat Langley and **Christopher J. MacLellan**

Georgia Institute of Technology, Atlanta, GA 30332 USA
pat.langley@gtri.gatech.edu, cmaclell@gatech.edu
http://www.isle.org/∼langley/, https://chrismaclellan.com

Background and Motivation

Although research on collaboration between humans and AI systems has a long and successful history, recent advances have produced even greater interest in the potential synergy. The most common idea is that AI assistants will let people be more productive, either by increasing their output or reducing their effort, but there is also hope that they will improve the quality of results. However, the field has seen some cases in which AI technology has led to negative consequences (Swartz, 2003). Clearly, some interaction designs for human-AI collaboration work better than others.

Naturally, we would like to understand the effectiveness of given human-AI team designs before we put them in the field. One response is to develop candidate AI agents and team designs, then to run experiments with human users that measure their joint effectiveness. But this process is time consuming and expensive, making it beyond the reach of many developers. In this paper, we propose the different approach of creating *abstract models* of human-AI team behavior and using them to explore and evaluate the space of alternative interaction designs.

Benefits of Abstract Models

Early research in AI was closely linked to cognitive psychology (Langley, 2012), with some of the first implemented systems cast by their designers as computational accounts of human behavior (e.g., Newell et al., 1960). These artifacts addressed complex tasks like theorem proving and game playing in much the same way as people but, as the field advanced, such programs remained difficult and costly to develop. Ohlsson and Jewett (1995) offered the promising alternative of creating abstract models of behavior. These were also stated as running computer programs and incorporated key ideas from AI, but they did not actually carry out tasks. For example, Ohlsson and Jewett reported a model of learning that searched through a problem space but did not represent the solver's states and goals.

We can adapt their insight to develop abstract computational models of human-AI teams. These would include agents that play the role of humans and AI members, and they would also reflect an analysis of the steps required for a complex task. Models would simulate performing these steps in sequence, including the time taken and the probability of success, but they would not actually do the steps. In this way, models would predict the time that teams require to produce results without needing to implement AI teammates or run experimental studies with human subjects.

Petri Net Models of Human-AI Teams

We have used an extension of Petri Nets (Peterson, 1977) to develop such abstract models of human-AI teams, building on the Machinations framework (Dormans, 2012). Figure 1 depicts graphically a model for the game Dice Adventure (Harpstead et al., 2023), in which three agents cooperate to achieve a shared objective. The model represents each agent as a token that flows through a sequence of tasks that involve searching for an orb. The sequences join at a coordination node for arriving at a tower, which all agents must reach before they can proceed to another level of the game.

The Machinations software not only lets one visualize this model but also simulate its behavior given different agent parameters. These can include the rate of each agent's progress and the probability it will complete steps. Running the simulation predicts the time it will take the team to achieve goals and its chance of success, letting us study how characteristics of agents on the team, whether AI or human, influence overall team performance. The Petri net's structure and agent roles are informed by a cognitive task analysis (Newell & Simon, 1972), so the framework also lets us evaluate different task decompositions and role assignments.

Of course, this approach is not limited to Dice Adventure. We can apply it to any complex problem that requires multiple steps and cooperation among agents, whether they are humans, AI systems, or a mixture. Models can be deterministic or stochastic, with the latter supporting probabilistic step completion and agent choice among different actions. The latter will require repeated runs to generate average results, but the abstract character of models means their simulation should be inexpensive even for ones that involve many agents. The framework's flexibility should make it useful across a wide range of tasks and interaction designs.

Discussion

We should note some related approaches to analyzing and modeling human performance. For instance, the GOMS framework (John & Kieras, 1996), which also bases abstract models on cognitive task analyses, has been widely used

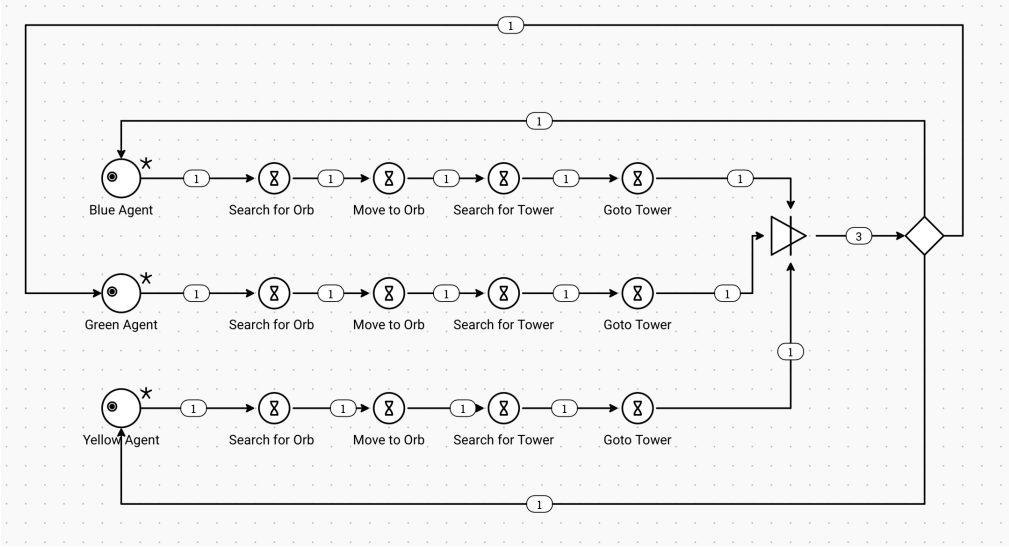

Figure 1: An abstract task model for the cooperative game Dice Adventure.

to predict behavior in human-computer interaction. To our knowledge, it has not been used to model coordination in human-AI teams. In contrast, Vignati et al. (2022) have used *joint activity graphs* to describe interdependence in multi-agent teams, but they used it to inform agent behavior rather than to predict overall team performance. Of course, we have built directly on Dormans' (2012) work, but he invented the Machinations formalism to specify designs for multi-person games, not to model team effectiveness.

Our research on abstract models of human-AI teams remains in its early stages. In future work, we plan to test the paradigm on additional collaborative tasks to demonstrate its generality and to study its ability to handle problems that require hundreds of teammates. We will also extend the framework to handle settings in which agents must communicate, with a special case being organizations in which some members have authority over others. Finally, we plan to examine tradeoffs between different performance criteria, such as time taken to complete tasks and quality of the result.

In summary, abstract computational models of human-AI collaboration hold considerable promise for predicting team performance before one makes the effort of constructing AI agents or running experiments with human subjects. Petri nets offer a convenient formalism for describing team activities at an abstract level and Machinations provides a useful software environment for specifying and simulating them. Together, they should let researchers explore the space of interaction designs for human-AI teams and evaluate alternatives in a cost-effective manner.

References

Dormans, J. (2012). *Engineering emergence: Applied theory for game design*. Doctoral dissertation, University of Amsterdam, Amsterdam, Netherlands.

Harpstead, E., Stowers, K., Lawley, L., Zhang, Q., & Maclellan, C. (2023). Speculative game design of asymmetric cooperative games to study human-machine teaming. *Proceedings of the Eighteenth International Conference on the Foundations of Digital Games*, 1–4.

John, B. E., & Kieras, D. E. (1996). The GOMS family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on Computer-Human Interaction*, *3*, 320–351.

Langley, P. (2012). Intelligent behavior in humans and machines. *Advances in Cognitive Systems*, *2*, 3–12.

Newell, A., Shaw, J. C., & Simon, H. A. (1960). Report on a general problem-solving program for a computer. *Proceedings of the International Conference on Information Processing* (pp. 256–264). UNESCO House, France: UNESCO.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S., & Jewett, J. J. (1995). Abstract computer models: Toward a new method for theorizing about adaptive agents. *Proceedings of the Eighth European Conference on Machine learning*, 33–52. Heraclion, Crete: Springer.

Peterson, J. L. (1977). Petri nets. *ACM Computing Surveys*, *9*, 223–252.

Swartz, L. (2003). *Why people hate the paperclip: Labels, appearance, behavior, and social responses to user interface agents*. Honors Thesis, Symbolic Systems Program, Stanford University, Stanford, CA.

Vignati, M., Johnson, M., Bunch, L., Carff, J., & Duran, D. (2022). Interdependence design principles in practice. *Frontiers in Physics*, *10*, 969544.