Social Planning: Achieving Goals by Altering Others' Mental States

Chris Pearce and **Ben Meadows** and **Pat Langley** and **Mike Barley**
Department of Computer Science, University of Auckland
Private Bag 92019, Auckland 1142, New Zealand

Abstract

In this paper, we discuss a computational approach to the cognitive task of social planning. First, we specify a class of planning problems that involve an agent who attempts to achieve its goals by altering other agents' mental states. Next, we describe SFPS, a flexible problem solver that generates social plans of this sort, including ones that include deception and reasoning about other agents' beliefs. We report the results for experiments on social scenarios that involve different levels of sophistication and that demonstrate both SFPS's capabilities and the sources of its power. Finally, we discuss how our approach to social planning has been informed by earlier work in the area and propose directions for additional research on the topic.

1 Introduction and Motivation

Most AI research on planning and problem solving is concerned with a single agent's physical activities, but human planning regularly incorporates interactions with other people as a means to achieve goals. Sometimes these are simple and straightforward, such as asking a taller person to reach for an item on a shelf. More sophisticated plans may take advantage of others' ignorance or false beliefs, such as selling an item for a high price even when one knows it has little value. Extreme cases, such as posing as a homeless person to increase donations while panhandling, may even involve intentional deceit.

Communicative acts are common in such social plans but they are not necessary. One can suggest the need for help, take advantage of ignorance, or engage in deceit without verbal exchanges. What social planning requires is the ability to reason about others' mental states and about the effects of one's actions, verbal or otherwise, on them. This appears to be a distinctive human trait, and our research aims to develop a computational account of the structures and processes that support it. We desire a general theory of social planning that operates across different domain content and that handles different level of sophistication in strategies for altering other agents' goals and beliefs.

Before we present such a theory, we should mention research paradigms that address related issues but that differ from our own along important dimensions. These include:

- *Game playing*, an AI subfield that deals with settings in which two or more agents compete to achieve some objective; most work in this area uses some form of adversarial search (e.g., minimax), which adopts a simpler model of mental states than concerns us here.
- *Collaborative planning* (e.g., Rao, Georgeff, and Sonenberg 1992), an area of AI that focuses on formation of joint plans among multiple agents; work in this tradition often encodes others' mental states (e.g., Castelfranchi 1998), but typically assumes they share goals.
- *Multi-agent systems* (Sycara 1998), a broad subfield that deals with coordination and collaboration among agents that, in some cases, pursue shared goals (e.g., Levesque, Cohen, and Nunes 1990); this contrasts with our concerns, which revolve around single agents that achieve their own goals in social settings.
- *Dialogue systems*, an area in which some efforts (e.g., Perrault and Allen 1980) represent and reason about the effects of communicative actions on others' mental states; again, most work assumes that agents share goals, but a few analyses have examined deception (e.g., Bridewell and Isaac 2011).
- *Plan recognition* (e.g., Goldman et al. 1999), an AI subfield concerned with inferring the goals that produce observed behavior; most work in this area deals with single agents, but some efforts (e.g., Meadows et al. 2013) incorporate domain-independent rules about social relations to infer the beliefs and goals of interacting agents.

None of these research paradigms have the same emphasis as our own, but each contains ideas and assumptions that will prove relevant for our treatment of social planning.

In this paper, we present a computational theory that incorporates some aspects of these earlier paradigms, but combines them in novel ways to provide a novel, system-level account of social planning. In the section that follows, we describe a class of scenarios that have driven our research on this topic. After this, we review FPS, a problem-solving architecture that we have used as the platform for our work. We describe some representational and processing extensions that let the system generate social plans, followed by empirical studies that examine their effectiveness. We conclude by reviewing related work in more detail and discussing avenues for future research.

Table 1: Four fable scenarios of varying levels of social complexity. The primary agent is a crow. Columns specify the problem goals, simplified versions of the initial states, and plausible target plans.

Fable	Initial State	Target Plan
HUNGRY DEER AND FRIENDLY CROW Goals: not(hungry(deer))	has(crow, apple) canEat(deer, apple) belief(deer, canEat(deer, apple) friends(deer, crow) friends(crow, deer) belief(deer, friends(crow, deer)) belief(deer, friends(deer, crow)) located(deer, field) located(crow, forest) hungry(deer) goal(deer, not(hungry(deer)))	1. travel(crow, forest, field) 2. give(crow, deer, apple, field) 3. eat(deer, apple)
CROW FLEECES THE SHEEP Goals: not(hungry(crow))	canEat(crow, berries) canEat(sheep, berries) belief(sheep, canEat(sheep, berries)) belief(sheep, canEat(crow, berries)) has(crow, pebbles) hungry(crow) has(sheep, berries) value(berries, med) value(pebbles, low) belief(sheep, value(berries, med)) belief(sheep, value(pebbles, med)) belief(sheep, belief(crow, value(berries, med))) belief(sheep, belief(crow, value(pebbles, med))) located(crow, field) located(sheep, field)	1. dishonest-trade(crow, sheep, pebbles, berries, field) 2. eat(crow, berries)
CROW MANIPULATES THE CHEETAH Goals: injured(lion) not(injured(crow))	adjacentTo(cave, field) not(injured(crow)) not(friends(crow, lion)) not(friends(lion, crow)) not(friends(crow, cheetah)) not(friends(cheetah, crow)) belief(cheetah, friends(crow, cheetah)) belief(cheetah, friends(cheetah, crow)) belief(lion, friends(crow, lion)) belief(lion, friends(lion, crow)) located(lion, cave) located(cheetah, field) located(crow, field)	1. persuade(crow, cheetah, located(cheetah, cave), field) 2. travel(cheetah, field, cave) 3. travel(crow, field, cave) 4. convince(crow, cheetah, insulted(lion, cheetah), cave) 5. persuade(crow, cheetah, injured(lion), cave) 6. fight(cheetah, lion, cave)
CUNNING CROW AND GREEDY SHEEP Goals: has(crow, jewel)	has(crow, apple) has(sheep, jewel) value(apple, mid) value(jewel, high) located(crow, field) located(sheep, field) goal(sheep, has(sheep, apple)) belief(sheep, value(apple, mid)) belief(sheep, value(jewel, low)) belief(sheep, belief(crow, value(apple, mid)))	1. bluff-inform(crow, sheep, value(jewel, med), field) 2. dishonest-trade(sheep, crow, jewel, apple, field)

2 Social Planning Tasks

We are interested in a class of tasks – which we call *social planning* – that can be characterized by three primary features. First, they are situated in a physical setting with two or more agents, including a primary agent whose beliefs and goals drive the planning process. Second, they involve not only knowledge about the conditional effects of physical actions on the environment, but also knowledge about the effects of social actions, such as communication, on the mental states of agents. Finally, each participating agent, not just the primary one, can access this knowledge.

We have developed scenarios similar to Aesop's fables to study this cognitive ability. They are brief, goal-directed, focused on high-level social interactions, and involve agents who reason about others' mental states. Table 1 presents four such vignettes, each described in terms of their initial state, the goals of the primary agent (here an intelligent crow), and a target plan that achieves those goals.

The first scenario, *Hungry Deer And Friendly Crow*, illustrates basic social planning: the crow works to achieve its goal of satisfying one of the deer's goals. The agents only model each others' mental states to the extent required by this basic joint activity. The second scenario, *Crow Fleeces the Sheep*, is more nuanced: the crow recognizes the sheep's false belief (the sheep believes that both he and the crow believe that the pebbles have medium value) and capitalizes on it for his own ends. In the third scenario, *Crow Manipulates the Cheetah*, the crow intentionally causes the cheetah to acquire a false belief. This leads the cheetah to adopt a goal to injure the lion, which produces a result the crow desires but cannot achieve directly without harming itself.

The ability to utilize another agent's belief or lack of belief as part of a social stratagem can also arise at an embedded level. The most sophisticated scenario, *Cunning Crow and Greedy Sheep*, features a double bluff. The sheep has found a high-value jewel, but believes it to be a worthless

paste gemstone. He wants the crow's medium-value apple. The crow cannot ask to trade with the sheep, as it believes the sheep does not think the two objects have equal value. Instead the crow comments aloud that the jewel must be of medium value, knowing that the sheep does not agree and leading him to think he can make a dishonest trade with the crow. However, the crow intended for this to occur, and in reality the trade the sheep initiates is in the crow's favor.

Planning for such social vignettes appears to require capabilities for encoding and reasoning about agents' models of others' mental states, representing social operators, and reasoning about how others make inferences. We address each of these issues below, but we must first review the architecture in which we embed our responses.

3 An Architecture for Social Planning

We have chosen to implement our ideas on social planning within the FPS problem-solving architecture (Langley et al. 2013). We will not claim that FPS is superior to other problem-solving frameworks, but we found it easy to adapt to the social planning task. In this section, we briefly review the architecture's representation and processes. After this, we discuss extensions to support social planning, starting with its representations and then turning to problem-solving mechanisms. We will refer to the extended architecture as SFPS, for *Social Flexible Problem Solver*.

3.1 A Review of FPS

Like many architectures, FPS contains a working memory and a long-term memory. The primary structure in the former is the *problem*, which includes a state description and a goal description. Another important structure is an *intention* or operator instance. A *solution* to problem P comprises an applied intention I, a subproblem for transforming P's state into one that meets I's conditions, a subproblem for transforming the state produced by applying I into one that satisfies P's goals, and solutions to both subproblems. The base case is a *trivial* solution in which the state satisfies the goals.

Long-term memory contains two forms of content: *domain knowledge*, which defines predicates, operators, and inference rules for a given problem domain, and *strategic knowledge*, which is domain independent. Domain content provides the material that planning uses to generate new subproblems, intentions, states, and goals; strategic content determines the details of the problem-solving process.

As in Newell, Shaw, and Simon's (1958) theory, problem solving in FPS involves transforming an initial state into one which satisfies the goal description by applying operators that manipulate state information. The architecture operates in cognitive cycles that involve five stages, each of which uses structures in long-term memory to update the contents of working memory. These include selecting a problem P on which to focus, selecting an operator instance I relevant to P, generating new subproblems based on I, checking for failure (e.g., loops), and checking for success (e.g., satisfied goals).

For our work on social planning, we incorporated strategic knowledge that combines iterative-sampling search, backward chaining, and eager commitment methods. Our pilot studies suggested that goal-driven problem solving is more focused when tasks involve altering other agents' mental states, although another approach like forward chaining might find the same solutions with additional search.

3.2 Representational Extensions

Before SFPS can generate social plans, it must first be able to represent social situations and relations. The most important feature here is the ability to encode models of other agents' goals and beliefs. To this end, we have augmented FPS's notation for problem states. Propositions that describe a situation are stored in working memory as beliefs of the primary agent.[1] Literals of the form *belief(Agent, Content)* specify that *Agent* believes *Content*. Embedded structures can denote beliefs about other agents' beliefs, as in *belief(lion, belief(sheep, belief(lion, not(sick(lion)))))*, which encodes the lion's belief that the sheep believes the lion believes that he is not sick. SFPS represents an agent's lack of belief in some proposition Q by asserting the absence of belief in either Q's truth or falsehood. For example, taken together *belief(lion, not(belief(sheep, sick(lion)))* and *belief(lion, not(belief(sheep, not(sick(lion)))))* denote that the lion believes the sheep does not know if the lion is sick.

The extended notation uses analogous literals of the form *goal(Agent, Content)* to specify agents' goals, which differ from beliefs by describing propositions an agent wants to hold. For example, *goal(lion, located(sheep, cave))* encodes the lion's wish for the sheep be located in the cave. Embedded structures may include goals as well as beliefs, as in *goal(lion, belief(sheep, goal(lion, has(lion, apple))))*, which says the lion wants the sheep to believe the lion's wants the apple. The primary agent's goals are not embedded within beliefs, but appear at the top level of memory.

In addition to domain-level operators, which alter the environment, SFPS also allows *social operators* that alter the mental states of other agents. Each of the social operators, which typically involve communication, are associated with an acting agent, and may refer to agents' beliefs and goals in both its conditions and effects. For example, the operator

bluff-inform(A1, A2, Content, Place) [A1 = actor] :
  alive(A1), alive(A2), at(A1, Place), at(A2, Place)
  belief(A2, not(Content))
  belief(A1, belief(A2, not(Content)))
  not(belief(A2, belief(A1, not(Content))))
  belief(A1, not(belief(A2, belief(A1, not(Content)))))
 →
  belief(A2, belief(A1, Content)) [main effect],
  belief(A1, belief(A2, belief(A1, Content))) [side effect]

defines a communicative action in which the acting agent, A1, informs the other agent, A2, about a proposition that A2 already believes to be false. The action causes A2 to believe that A1 believes the proposition is true.

[1] We do not assume the primary agent is omniscient: it does not necessarily believe all true propositions about the world it inhabits. It starts only with the beliefs specified in the initial state, which may not be a complete world description.

As seen in the example, operators can have both *main* and *side effects*, typically one of the former and several of the latter. This distinction is useful when the primary agent must incorporate other agents' actions into its plans. The primary agent should not assume that others will willingly perform any action that helps achieve its goals. To this end, SFPS's intentions for nonprimary agents include an extra field to specify the goal that motivated its selection.

3.3 Processing Extensions

These representational changes let SFPS encode the information needed for social planning, but their effective use depends on extended mechanisms. The basic problem-solving cycle remains similar, but we have modified the architecture in two primary ways. The first adds the ability to elaborate states using query-driven deductive inference. This mechanism operates during the intention generation, subproblem creation, and termination checking stages, where it aids in determining whether preconditions and goals are satisfied.

This reasoning process has more general applicability, but it is especially important in social settings. The inference stage can operate over domain rules, but SFPS also takes advantage of conceptual rules like

not(belief(A, X) $\leftarrow$ *belief(A, not(X))* and
not(belief(A, not(X)) $\leftarrow$ *belief(A, X)* .

These let the primary agent infer the truth of propositions about lack of belief. For example, the first one lets it conclude, from *belief(lion, hungry(crow))*, that *not(belief(lion, not(hungry(crow))))*.

Like its predecessor, SFPS adopts the closed world assumption at the top level of the primary (planning) agent's beliefs. If this agent does not have a belief in working memory, and if it cannot infer it, then it does not hold that belief. For another agent, A2, whose beliefs the primary agent attempts to model, checking whether A2 is ignorant of some proposition Q involves checking that both Q and its negation appear within A2's negated beliefs in the proper embedding.

However, to generate plans that incorporate the beliefs and goals of other agents, SFPS must sometimes carry out embedded inference. The application of social operators, for example for communicating beliefs , produces some of these directly. However, once these have been added to working memory, the new stage applies inference rules for any embedded context. For instance, given the rule

not(safe_at(A, P)) $\leftarrow$ *could_harm(B, A)* $\wedge$ *located(A, P)*

and the working memory elements

belief(lion, belief(sheep, could_harm(lion, sheep)))
belief(lion, belief(sheep, located(lion, cave))) ,

the inference stage would add the working memory element

belief(lion, belief(sheep, not(safe_at(sheep, cave)))) .

for this state, even though the inference rule was defined outside the context of any mental states.

Another extension involves the intention selection stage, which SFPS must invoke not only for the primary agent but also for other agents in the scenario. Recall that operators now distinguish between a main effect and side effects. When selecting operator instances for nonprimary agents, SFPS has a higher probability of selecting candidates that achieve a goal through the main effect. This strategy does not rule out entirely consideration of nonprimary intentions that utilize side effects, but it discounts them and biases the search process to favor plans that rely on main effects.

3.4 Illustrative Example

To explain the process of social planning, we examine a trace of SFPS's steps as it generates a solution for the *Cunning Crow and Greedy Sheep* scenario from Table 1. We will refer occasionally to issues of backtracking and search, but we will focus mainly on choices included in the final plan.

The goal associated with the top-level problem is for the primary agent, the crow, to have the jewel. SFPS uses backward chaining to generate a set of candidate intentions based on operators that would produce a state which satisfies the goal description. These intentions include each agent trading their objects (both honestly and dishonestly), as well as the sheep simply giving the crow the jewel.

The system rates each intention's potential on the basis of its unsatisfied conditions and the goals it would achieve. Based on this rating, SFPS creates subproblems for each of the *give* intention's conditions, but attempts to solve them fail, and it eventually selects a *dishonest-trade* intention with the sheep initiating the trade, then uses the intention's conditions to specify the goals for a new subproblem. These goals include certain beliefs about the relative values of the objects being traded. For example, one of the operator's conditions, *belief(A1, not(*$\geq$*worth(Ob1, Ob2)))*, is not in working memory, so SFPS successfully deduces this relation using an inference rule that states $\geq$*worth(Obj1, Obj2)* if Obj1's value is greater than or equal to Obj2's value.

However, two conditions of the *dishonest-trade(sheep, crow, jewel, apple, field)* intention are still not met in the initial state: *belief(sheep, belief(crow,* $\geq$*worth(jewel, apple)))* and *belief(sheep, belief(crow,* $\geq$*worth(apple, jewel)))*. No operator achieves $\geq$*worth* directly, so SFPS resorts to embedded inference to find operators whose application would satisfy the current subproblem's goals. In this case, it finds that an operator with the effect *belief(sheep, belief(crow, value(jewel, med)))*, when combined with the existing element *belief(sheep, belief(crow, value(apple, med)))*, would serve this purpose indirectly.

At this point, SFPS realizes that one of its communicative actions will have the desired effect. The system selects the intention *bluff-inform(crow, sheep, value(jewel, high), field)* and finds its conditions are all satisfied. Applying it to the initial situation produces a successor state that satisfies the previously unsatisfied conditions of the *dishonest-trade* intention. SFPS notices that applying this operator achieves the top-level goal, and thus recognizes it has produced a plan that solves the original problem, and halts its search.

This example demonstrates the system's ability to construct plans that are quite complex – not in their length, but in their manipulation of others to the primary agent's own ends. Here the crow believes that the sheep does not believe the crow thinks the objects have similar value, and that the sheep actually believes the opposite, so the crow cannot initiate the trade. However, by tricking the sheep into believing

the crow has a false belief about the jewel's value, he manipulates the sheep into capitalizing on what it incorrectly views as the crow's ignorance. This is a prime example of what we call social planning.

3.5 Central Features

We can summarize these extensions in terms of three capabilities that appear central to the task of social planning:

- Storing the primary agent's beliefs/goals about others' beliefs/goals, which may differ from its own structures;
- Elaborating upon states using inference, including applying rules within different levels of embedding;
- Incorporating other agents' goals into plans, but preferring candidates that use operators' main effects.

The first capability supports reasoning about the effects of communication and interaction, whereas the second lets the primary agent apply its knowledge to elaborate on the beliefs of others. Without the final capability, nonprimary agents would be mere extensions of the primary agent's will, automatically performing any action it desires and thus avoiding the need for social manipulations. SFPS combines these capabilities into a novel architecture for social planning that operates from the primary agent's perspective in scenarios that involve others who may hold fewer or different beliefs.

4 Empirical Evaluation

We have described extensions to FPS designed to let it engage in the task of social planning, but whether they work as intended is an empirical question. In this section, we present three hypotheses about the system's abilities and report experiments designed to test these assertions.

4.1 Claims and Methodology

We are interested in SFPS's ability to generate plausible plans from the perspective of an agent in a social scenario. These scenarios involve not simply interactions with other agents, but also interactions that involve incomplete information, reasoning based on false or incomplete beliefs, and other agents who do not cooperate by default.

We can transform these ideas into three hypotheses about the extended system's ability to carry out social planning:

- SFPS can create plausible plans for achieving goals in social scenarios;
- This ability relies on embedded inference to generate models of others' beliefs; and
- This ability also relies on SFPS's capacity to incorporate the actions of other agents in its plans.

We have tested these hypotheses experimentally using two fables for each level of complexity, including those described in Table 1. The target plans for this fable set have an average of 3.2 operators. We addressed the first hypothesis simply by running SFPS on each scenario and measuring its success rate. To test the second claim, we removed the ability to model the reasoning of nonprimary agents using embedded inference. We evaluated the final hypothesis by preventing creation of plans containing nonprimary actions.

Table 2: Experimental results. The rows describe, for each level of sophistication, the number of runs in which SFPS generated a plausible plan, generated an implausible plan, or failed to generate a plan within 10,000 cycles. We ran the system 50 times on each of the eight scenarios.

Level of Sophistication	**Plausible Plan**	**Implausible Plan**	**Did Not Finish**
Basic Social Interation	75	25	0
Capitalize on Misbeliefs	78	22	0
Deceive Other Agents	68	29	3
Encourage False Beliefs	86	10	4

In order to determine SFPS's proficiency in each condition, we must be able to measure its behavior. Our dependent variable is the number of *plausible* plans generated. We view a plan as plausible if its operator sequence transforms the initial state into a goal-satisfying state with nonprimary agents that only apply actions to achieve those operators' main effects. We consider runs that generate plans violating this condition, or that exceed 10,000 cycles, to be failures. SFPS is nondeterministic, in that it selects intentions probabilistically, so we ran it 50 times on each of the eight problems and report summary results from these runs.

4.2 Experimental Results

Table 2 presents the results of our first experiment, addressing our basic claim, that SFPS can generate plans that achieve the primary agent's goals in a social setting. The system found valid plans in 97 percent of runs, with the failures caused by reaching the cycle limit. These plans ranged from two to six operators, with 3.2 on average. Some 38 percent of these operators involved nonprimary agents. Success rates differed slightly across levels of social sophistication.

However, SFPS did produce plans that involved nonprimary agents carrying out actions involving unusual 'side-effect' motivations in 22 percent of the completed runs. In one run, a sheep asked a fox to adopt the goal of being injured, so the fox would try to start a fight that benefited the primary agent (injury of the acting agent being a side effect). In this case, the bias against wishful thinking was ineffective due to the probabilistic character of intention selection.

Our second hypothesis was that SFPS's ability to generate social plans relies on its capability for embedded inference. With this facility disabled, the system found plans on only 193 of the 400 problem runs. Of these, 38 had actions selected for their side effects. Most of the 207 failures occurred before the cycle limit was reached, as the system had generated all plans that it was possible to create without embedded inference. This omission meant that the system failed to realize that some operators would lead – indirectly – to satisfaction of certain goals, which it did when embedding was available. Despite this hindrance, SFPS still produced plans on nearly half of the runs (when those runs did not rely on inference over the beliefs of nonprimary agents).

Our final claim focused on incorporating the actions of other agents into plans. As this facility plays a central role in our definition of social planning and was emphasized by our scenarios, it is not surprising that, without this ability, SFPS could find plans for only one of the eight problems. However, it produced plausible solutions on all 50 runs for that problem because there was a valid plan that only required actions of the primary agent. Of course, we might have increased the system's success rate if we had made additional operators available to the primary agent, but these would not have been social plans as we have defined them.

In summary, our experiments on the eight scenarios demonstrated that SFPS can, in many cases, generate plausible social plans that range from simple helping activities to Machiavellian schemes that take advantage of false belief and intentionally deceiving the other party. The studies also revealed, as expected, that drawing inferences from the perspective of other agents and incorporating their actions into plans were crucial to success on these tasks.

5 Related Research

As noted earlier, our approach to social planning borrows ideas from a number of paradigms that we cannot review in detail here. However, our research incorporates three key ideas about social cognition, and we should briefly review efforts related to each one.

Our account of *social planning relies centrally on encoding the primary agent's beliefs and goals about others' mental states and operators for altering them.* There is a substantial body of research on reasoning about other agents' mental states, including Fahlman's (2011) work on 'contexts' in Scone, and Bello's (2011) use of 'worlds' in Polyscheme, but these and most other efforts have focused on reasoning rather than on goal-directed planning, as in our framework. Levesque et al. (1990) also discuss reasoning about others' beliefs and goals, but only in the context of cooperative activities. Perrault and Allen (1980) analyzed speech acts in terms of operators for altering mental states, but they did not use such operators for plan generation. Briggs and Scheutz (2013) have used them to plan role-specific dialogues, but only for cooperative scenarios.

A related assumption is that *social planning benefits from inference about problem states, including application of rules at different levels of embedding*. Some classical planning systems carry out inference over state descriptions, but these do not reason about others' mental states. Scone and Polyscheme utilize default reasoning by inheritance to support this ability, which SFPS achieves by nested application of inference rules. Bridewell and Isaac (2011) analyze deception in terms of reasoning about other agents' goals and beliefs, which they store in distinct partitions, but, again, they do not address plan creation.

The final idea is that *social planning requires incorporating other agents' intentions into plans, but only in a constrained way that avoids wishful thinking*. Early work in this area includes Mueller and Dyer's (1985) computational model of daydreaming, which also explored social interactions but which actually emphasized wishful thinking. Meehan's (1977) early approach to story generation, as well as Riedl and Young's (2010) more recent system, also generate plans for social interaction. The latter supports scenarios that involve deception, but both focus on producing interesting stories rather than achieving a primary agent's goals.

Our approach to social plans contains elements that have appeared in earlier systems, but it combines them in novel ways to support generation of plans that achieve goals by manipulating the goals and beliefs of others. At first glance, Pontelli et al.'s (2012) work on planning with physical and communicative actions, as well as inference over effects, appears to address similar issues. However, they do not model other agents' goals or bias their behavior toward actions that would achieve them, as does our account of social planning.

6 Concluding Remarks

In this paper, we introduced the task of social planning, in which an agent attempts to achieve its goals by altering others' mental states, and we presented a theory of the representations and processes that underlie this high-level cognitive ability. Our account includes the representational claim that social planning requires encoding others' mental states and how activities alter those states. In terms of processes, it posits that planning mechanisms must include the ability to elaborate mental models using embedded inference and to select operators for nonprimary agents using main effects. No other changes are required to support this functionality.

We implemented these theoretical assumptions by extending FPS, a flexible problem-solving architecture. We tested the resulting system, SFPS, on eight scenarios that involved social planning at increasing levels of sophistication. In addition to demonstrating a basic capability for producing plans that involve influencing others' beliefs and goals, we reported lesion studies that showed the role played by embedded inference and nonprimary agents' operators. We also reviewed prior research that shares theoretical assumptions, although we found none that have combined them to support social planning in the sense we have defined it.

Although our work to date offers a promising account of this important cognitive ability, we can still extend it along a number of fronts. First, we must develop more robust methods for avoiding wishful thinking, such as assuming other agents have default goals like remaining healthy and rejecting plans that voilate them. Second, we should address scenarios that involve more complex forms of manipulation, such as those arising in confidence tricks. These will require us to provide SFPS with additional social knowledge, but changes to the architecture itself should not be necessary. Third, we should tackle cases in which inference about others' mental states involves abductive rather than deductive inference to provide belief ascription. These will require changes to the architecture, but we should be able to build on recent work by Meadows et al. on abductive reasoning. Finally, we should note that our work to date has assumed that social actions have deterministic effects. Future versions of SFPS should reflect the fact that outcomes are uncertain and support conditional plans that offer the agent multiple paths to its goals in case early efforts fail. Taken together, these extensions will provide a more comprehensive account of the structures and processes that arise in social planning.

Acknowledgements

This research was supported in part by Grant N00014-10-1-0487 from the U.S. Office of Naval Research, which is not responsible for the paper's contents. We thank Paul Bello, Will Bridewell, and Alfredo Gabaldon for discussions that influenced our approach to social understanding, along with Miranda Emery and Trevor Gee for work on the FPS system.

References

Bello, P. 2011. Shared representations of belief and their effects on action selection: A preliminary computational cognitive model. In *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*, 2997–3002. Boston, MA: Curran Associates.

Bridewell, W., and Isaac, A. 2011. Recognizing deception: A model of dynamic belief attribution. In *Proceedings of the AAAI Fall Symposium on Advances in Cognitive Systems*, 50–57. Arlington, VA: AAAI Press.

Briggs, G., and Scheutz, M. 2013. A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 1213–1219. Bellevue, WA: AAAI Press.

Castelfranchi, C. 1998. Modelling social action for AI agents. *Artificial Intelligence* 103:157–182.

Fahlman, S. E. 2011. Using Scone's multiple-context mechanism to emulate human-like reasoning. In *Proceedings of the AAAI Fall Symposium on Advances in Cognitive Systems*, 98–105. Arlington, VA: AAAI Press.

Goldman, R.; Geib, C.; and Miller, C. 1999. A new model of plan recognition. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 245–254. Stockholm: Morgan Kaufmann.

Langley, P.; Emery, M.; Barley, M.; and MacLellan, C. J. 2013. An architecture for flexible problem solving. In *Poster Collection of the Second Annual Conference on Advances in Cognitive Systems*, 93–110.

Levesque, H. J.; Cohen, P. R.; and Nunes, J. H. 1990. On acting together. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, 94–99. Boston, MA: AAAI Press.

Meadows, B.; Langley, P.; and Emery, M. 2013. Understanding social interactions using incremental abductive inference. In *Proceedings of the Second Annual Conference on Advances in Cognitive Systems*, 39–56.

Meehan, J. R. 1977. TALE-SPIN: An interactive program that writes stories. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, 91–98. Cambridge, MA: Morgan Kaufmann.

Mueller, E. T., and Dyer, M. G. 1985. Towards a computational theory of human daydreaming. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, 120–129. Irvine, CA: Lawrence Erlbaum Associates.

Newell, A.; Shaw, J. C.; and Simon, H. A. 1958. Elements of a theory of human problem solving. *Psychological Review* 65:151–166.

Perrault, C. R., and Allen, J. F. 1980. A plan-based analysis of indirect speech acts. *Computational Linguistics* 6:167–182.

Pontelli, E.; Son, T. C.; Baral, C.; and Gelfond, G. 2012. Answer set programming and planning with knowledge and world-altering actions in multiple agent domains. In Erdem, E.; Lee, J.; Lierler, Y.; and Pearce, D., eds., *Correct Reasoning: Essays on Logic-Based AI in Honour of Vladimir Lifschitz*. Berlin: Springer. 509–526.

Rao, A. S.; Georgeff, M. P.; and Sonenberg, E. A. 1992. Social plans. In *Proceedings of the Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, 57–76.

Riedl, M. O., and Young, R. M. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 39:217–268.

Sycara, K. P. 1998. Multiagent systems. *AI Magazine* 19:79–92.