Evaluating Mixed-Initiative Systems: An Experimental Approach

Gabriella Cortellessa and Amedeo Cesta

National Research Council of Italy
Institute for Cognitive Science and Technology
Via S. Martino della Battaglia 44, I-00185 Rome, Italy
{name.surname}@istc.cnr.it

Abstract

Mixed-Initiative approaches to Planning and Scheduling are being applied in different real world domains. While several recent successful examples of such tools encourage a wider use of this solving paradigm, research in mixed-initiative interaction is still at an early stage and many important issues need to be addressed. In particular, most of the work has been dedicated to designing working prototypes and identifying relevant features of the mixed-initiative interaction, while much less attention has been given to the problem of evaluating the approach as a whole. This article is aimed at addressing some of the many diverse aspects involved in Mixed-Initiative Planning and Scheduling system evaluation, highlighting the need for a methodology to provide effective evaluation studies for this class of tools.

In this paper we consider an established research methodology in experimental psychology, and adopt it to investigate specific aspects of mixed-initiative interaction. Specifically, the experiments described in this article shed some light on three aspects: (a) understanding users' attitude when choosing between automated and mixed-initiative problem solving, (b) investigating recourse to explanation services as a means to foster the users' involvement in the solving process, and (c) investigating possible individual differences (e.g., experts vs. non-experts) in the choice of resolution strategy or access to explanation.

Introduction

Several real world domains, such as manufacturing, space, logistics and transportation have demonstrated how the use of AI planning and scheduling applications aimed at supporting human decision making are useful and convenient. Automated techniques can relieve humans from solving hard computational problems, saving their "cognitive energy" for high-level decision tasks. Nonetheless, the introduction of intelligent systems for solving complex problems has raised the issue that, in most cases, a completely automated approach is neither applicable nor suitable. As a matter of fact, automated problem solving is difficult to integrate into human-centric activities, both for technical and psychological reasons.

Although there are certainly some exceptions, total automation of decision-making is not an appropriate goal in most domains. More typically, it is the case that experienced users and automated planning/scheduling technologies bring complementary problem-solving strengths to the table, and the goal is to synergistically blend these combined strengths. Often the scale, complexity or general ill-structuredness of real domains overwhelms the solving capabilities of automated planning and scheduling technologies, and some sort of problem decomposition and reduction is required to achieve problem tractability. Likewise, human planners often have deep knowledge about a given domain which can provide useful strategic guidance, but they are hampered by the complexity of grinding out detailed plans/schedules. In such cases, successful technology application requires effective integration of user and system decision-making. In this light, the solving paradigm known in literature as *mixed-initiative approach*, (Burstein & McDermott 1996; Cohen *et al.* 1998), is receiving increasing attention and interest.

This emerging paradigm fosters human-computer cooperation during the resolution of complex problems. The approach pursues the idea of complementarity between experienced users and automated technologies and aims at integrating them to obtain a more efficient system. A combined ⟨*human, artificial solver*⟩ system can create a powerful and enhanced problem solver applicable to the resolution of difficult real world problems.

Current proposals for mixed-initiative systems are very often presented as system descriptions and developed on purpose for *ad hoc* scenarios. Less work has been devoted to understanding how it is possible to evaluate the utility of both the whole approach and its different features, and to study users' attitudes toward this new approach. In addition, while several works on the mixed-initiative paradigm claim that end-users of automated systems prefer to maintain control over the problem solving, thus appreciating mixed-initiative systems, no empirical evidence has been provided to support this statement. This paper contributes in this direction.

The work is also motivated by the observation that the main concern of scholars in the problem solving field has been the development of efficient and powerful algorithms for finding solutions to complex problems. A usually neglected issue has been the lack of effective front end design through which an end user can interact with the artificial tool. A desiderata in this respect would be to have the user

benefit from the potentialities of the automated features of the tools, taking, at the same time, an active role in the resolution process. In this light, the generation of user-oriented features and advanced services such as explanation functionalities, what-if analysis, etc. becomes crucial.

This paper applies an experimental approach to the problem of understanding users' attitude toward mixed-initiative problem solving features and investigating the importance of explanation services during problem solving. Three main issues will be considered, namely (a) users' attitude toward the mixed-initiative vs. automated solving approach; (b) users' willingness to rely on explanation as a mean to maintain control on the machine; (c) possible individual differences between experienced and inexperienced users. In general we would like to stress the need of designing effective evaluation studies for assessing the effectiveness of this class of interactive systems.

Plan of the paper.   In the remainder of the paper we first summarize the state-of-the-art in mixed initiative systems and highlight some research aspects that have not received but deserve attention. We then describe our work, which inherits features from experimental research in psychology and human-computer interaction. Following the structure of psychological evaluation methodologies, we set up an experimental apparatus, design experiments formulating hypothesis, gather data, and finally interpret them. A final section discussing the practical implications entailed by our approach ends the paper.

Overview on Mixed-Initiative Systems

Mixed-initiative systems for solving planning, scheduling and in general complex combinatorial problems are becoming more and more pervasive in many application areas such as space missions, rescue, air campaign or vehicle routing. In recent years, several systems have been proposed for mixed-initiative problem solving which try to integrate in a unique system the complementary abilities of humans and machines.

MAPGEN (Ai-Chang *et al.* 2004) represents a successful example of a mixed-initiative system used to address a real world problem. The system uses a constraint-based temporal planner as the main automated solver and assists the Mars Exploration Rover mission control center in generating the activity plans. The design of the interactive part has been instrumental for the introduction of the tool in the real mission. COMIREM (Smith, Hildum, & Crimm 2005), is a general purpose tool for continuous planning and resource management under complex temporal and spatial constraints. It implements a user-centered approach to scheduling and resource allocation, providing users with a variety of tools for mixed-initiative resource allocation, feasibility checking, resource tracking and conflict resolution. Both MAPGEN and COMIREM are based on an interactive incremental approach that allows users to posts their decisions and see immediately the effects. In this context, conflict analysis and explanation services become fundamental "tools" for collaborative problem solving. Also *what-if* analysis capabilities are useful tools for guiding the search process and compar-

ing different partial solutions. TRIPS (Thinking, Reasoning, and Intelligent Problem Solving) (Ferguson & Allen 1998) provides an example of integrated system developed to support spoken-language dialogue for the collaborative resolution of planning problems. PASSAT (Plan Authoring System Based on Sketches, Advice and Template) (Myers *et al.* 2003) introduces the concepts of plan sketches and supports a user in the collaborative refinement of plan sketches, ultimately leading to a satisfactory solution.

A more specific system is the one developed at the Mitsubishi Electric Research Laboratory (Anderson *et al.* 2000). The system proposes an effective and interactive schema called *human-guided simple search* devoted to the solution of a well-known and difficult combinatorial optimization problem, namely the *capacitated vehicle routing with time-windows*. The human-guided search paradigm allows users to explore taking into account trade-offs among possible solutions, thus aiding the process of choosing a solution based on the user's understanding of the domain. Users can manually modify solutions, backtrack to previous solutions, and invoke a portfolio of search algorithms. More significantly, users can guide and focus the search through a visual metaphor that has been found effective on a wide variety of problems.

Broadly speaking, all of the above systems follow general principles for enabling collaborative problem solving schemes between the system and the user. First, they make solution models and decisions user-understandable, that is, they communicate elements of their internal models and solutions in user-comprehensible terms (for example, by using simple forms of explanation functionalities). Second, they allow different levels of user participation, that is, a solving process can range from a monolithic run of a single algorithm to a fine-grained decomposition in a set of incremental steps. Furthermore, they provide tools (e.g. what-if analysis, conflict resolution mechanisms, etc.) which promote the interactive and incremental construction of solutions.

This brief overview points to the issue we feel is missing in the current research trend in mixed-initiative problem solving, namely that of evaluation. Because of their composite nature, the design, implementation, and above all the evaluation and measurement of their effectiveness and utility, is an arduous and stimulating challenge. Two factors contribute to this challenge. First, the diversity and complexity of the two entities involved in the resolution process, i.e., the human user and the artificial solver. On one hand, humans perform unpredictable and sophisticated reasoning; on the other, artificial solvers are technically complex and adopt solving strategies which are very different from those employed by humans. Secondly, the environment from which the problem to be solved is drawn is usually uncontrollable and uncertain. Together, these factors complicate the task of designing precise and effective evaluation studies. For this reason, the design and use of well-founded methodologies for the validation of mixed-initiative planning and scheduling tools is fundamental for a successful deployment of this kind of systems in the real world.

Which research topics for mixed-initiative systems?

Usually, the synthesis of implemented systems that effectively exploit a certain methodology is instrumental for the establishment of any specific research area. Moreover, after the successful deployment of such systems, it is of great importance to consolidate the theory behind these systems, as well as to identify open problems and indicate possible road-maps for solving them.

The concept of mixed-initiative systems has been recognized as useful, and many specialized events have been dedicated to it. A study that identifies a set of issues involved in a mixed-initiative effort is presented in (Bresina *et al.* 2005), which explicitly lists a number of subtopics put forward by the MAPGEN experience.

It is worth noting that many of the issues to be investigated belong to two possible categories: (a) improvement of the underlying problem solving technology to better serve the mixed-initiative interaction; (b) empowering the user with services that enhance their involvement and their active role. Examples of this first type are the effectiveness of specific features of the underlying technology (e.g., extracting information from the temporal network that represents the current solution in constraint-based technology, tweaking user preferences in connection with the same representation, etc.). An example of the second type is the need for automated synthesis of explanations for the users. Some work is appearing with initial interesting results, mostly based on constraint based representations (e.g., (Wallace & Freuder 2001; Jussien & Ouis 2001; Smith *et al.* 2005)).

A point we consider particularly relevant is the identification of a precise methodology not only for the design but also for the evaluation of mixed-initiative systems. This limitation has been recently recognized, but little work has produced explicit results. A first interesting exception is the paper (Kirkpatrick, Dilkina, & Havens 2005), where a framework for designing and evaluating mixed-initiative systems is presented. Through this framework, several general requirements of an optimization mixed-initiative system are listed. According to the authors, these requirements can help to establish valid evaluation criteria. Very interesting is also the work presented in (Hayes, Larson, & Ravinder 2005), where a specific mixed-initiative system is described and an evaluation procedure is shown to measure the influence of the mixed-initiative approach on the problem solving performance. The present work aims to contribute further on this specific issue.

Evaluating mixed-initiative systems

This paper proposes steps for systematic evaluation that rely upon a ground methodology to quantitatively analyze different features of mixed-initiative systems.

Generally speaking, the evaluation of mixed-initiative systems entails two main aspects:

– *Measuring the problem solving performance*, that is evaluating the problem solving performance of the pair ⟨*human, artificial solver*⟩. This type of evaluation aims at demonstrating the advantages of the mixed-initiative approach for improving problem solving performance. For example, in (Anderson *et al.* 2000) experiments have

shown that human guidance on search can improve the performance of the exhaustive search algorithm on the targeted problem.

– *Measuring quality of interaction*, that is evaluating different aspects related to the users' requirements and judgment on the system (e.g., users' preferences on interaction styles, interactive features, level of trust of the system, clarity of presentation, usability, etc). These aspects, which are more strictly related to the human component, are fundamental for a successful integration of human and artificial solver during problem solving, especially if the system is intended to be used in real contexts.

Our work aims to highlight how the design of interactive tools need to take into consideration users' needs. To this end, we assume that the theoretical and methodological approach in psychological research (see for example (Goodwin 2005)) could be a valid means to better understand the issues involved. Our work relies on this methodology for evaluating various aspects of mixed-initiative planning and scheduling problem solving. In particular, we have set up a rather rigorous experimental approach and used it to look for answers to two open questions. The first, very general question, relates to the validity of the whole solving approach, that is the study of users' attitudes toward the mixed-initiative approach in comparison with the use of a completely automated resolution. The second question is more specific. It is related to the emerging topic of generating automatic explanations. As already mentioned, mixed-initiative systems imply a continuous communication between users and machines. Explaining the system's reasoning and choices is considered an important feature for these systems, and the problem of generating user-oriented explanation is receiving much attention. Our experiments are aimed at providing empirical evidence for assessing the willingness of real users' to rely on explanation during mixed-initiative problem solving. The choice of this second feature with respect to many others in mixed-initiative research is indeed biased by our current research interests.

Setting up an empirical study

The design of the experimental study has focused on features of a baseline mixed-initiative system named COMIREM (Smith, Hildum, & Crimm 2005). This is a mixed-initiative problem solver devoted to the resolution of planning and scheduling problems. In accordance with the mixed-initiative approach, the ambitious idea behind COMIREM is to capture the different skills that a user and an automated system can apply to the resolution process, by providing both powerful automated algorithms to efficiently solve problems and interactive facilities to keep the human solvers in the loop. The choice of COMIREM is motivated by the presence of the set of features we were interested in. Specifically, the tool allows a user to choose between an automated solving strategy and an incremental solution refinement procedure, and is endowed with an interesting set of simple explanation features.

In order to extend the experimental evaluation to a large number of participants while maintaining some meaningful variables of the experiment under tight control, our exper-

iment was designed on a simplified version of COMIREM. Specifically, the system layout was conceived in order to limit, to some extent, the richness of information of its interaction front end. The version of COMIREM chosen for the experiment solves scheduling problem instances in a TV broadcasting station domain. This domain was chosen because it is rather intuitive also for non expert users, whose participation in the experiments was pivotal.

Once the experimental context was set, we formulated the motivating questions for the analysis: do users prefer to actively participate in the solving process choosing the mixed-initiative approach, or do they prefer to entrust the system with the problem solving task thus choosing the automated approach? Do users of mixed-initiative systems rely on explanation during problem solving? Are there individual differences between expert and non-expert users? Is the difficulty of problems a relevant factor in the choice of the strategy or in accessing the explanation?

Given this set of general questions, the experimental methodology requires to carefully formulate hypotheses to be tested and the variables that should be monitored during the experiments.

Automated vs. mixed-initiative problem solving

Two main aspects have been considered as relevant variables in influencing users' choice in whether or not to employ the mixed-initiative approach, namely problem *difficulty* and the user's level of *expertise*. For the first variable, two levels have been considered: low and high difficulty, broadly corresponding to easy and hard problems. The two levels of this variable have been determined considering problem dimension in terms of number of activities to be scheduled, and alternative resources available for each activity. It is worth highlighting that a preliminary test proved the validity of our manipulation of the *difficulty* variable. Indeed, all participants correctly perceived the difference in the two levels of difficulty (easy and hard). As for the second variable, we considered the user's specific *expertise* in planning and scheduling problem solving. Two levels are considered for the variable: expert and non-expert.[1]

The first study aimed at investigating the influence of the two variables (*expertise* and *difficulty*) on the selection of mixed-initiative vs. automated strategy. In our experiment, the user is presented an alternative between a completely automated procedure and a mixed-initiative approach. By choosing the first alternative, the user will delegate all decisions to the artificial solver, thus maintaining no control over the problem solving process, whereas in the second case the system and the human solver will actively cooperate to produce a solution to the problem.

There are some interesting results that show how humans do not always accepts advice provided by artificial tools, rather ignoring them (Jones & Brown 2002). A possible explanation for this behavior is provided by previous research in human-computer interaction showing how humans tend to attribute a certain degree of anthropomorphism to computers and assign to them human traits and characteristics. In

---

[1] A future direction of this study will include the domain expert as additional level of the variable *expertise*.

(Langer 1992; Nass & Moon 2000), a series of experimental studies are reviewed reporting that individuals mindlessly apply social rules and expectations to computers. It is plausible to hypothesize that human problem solvers manifest the same tendency toward artificial solvers, and refuse to delegate the problem solving for many reasons. For instance, they could mistrust the effectiveness of automated problem solving, or they could enter in competition with the artificial agent. However, we have no data on possible differences in the behavior of users with different levels of expertise. Planning and Scheduling experts are people with some knowledge of the design of artificial solvers and they are aware of the limitations and merits of the system. We assume they would adopt a more pragmatic strategy, thus delegating the machine to solve the problem in order not to waste time. On the other hand they may be interested in understanding the procedure applied by the system. Hence, when facing difficult tasks, they might be motivated to test themselves and actively take part in the process. Conversely, non-experts do not know the mechanisms behind the automated algorithms and thus might have a different degree of trust. Nonetheless the greater the difficulty of the problems, the more likely the choice to commit the problem solving to the machine. For these reasons, we believe that some differences might exist between experts and non-experts while interacting with an artificial problem solver. In particular we formulate the following hypotheses:

Hypothesis 1. *Solving strategy selection (automated vs. mixed-initiative) depends upon user expertise. In particular it is expected that scheduling experts use the automated procedure more than non-experts. Conversely, non-expert users are expected to use the mixed-initiative approach more than experts.*

Hypothesis 1b. *In addition, it is expected that when solving easy problems, inexperienced users prefer the mixed-initiative approach, while expert users have a preference for the automated strategy. Conversely, for solving difficult problems, inexperienced users may prefer the automated strategy while expert users have a tendency to choose the mixed-initiative approach.*

The role of explanation in mixed-initiative systems

Among the numerous aspects involved in the development of mixed-initiative systems, one important requirement is the need to maintain continuous communication between the user and the automated problem solver. This continuity is usually lacking in current mixed-initiative systems. System failures that may be encountered in finding a solution typify this sort of deficiency. Typically, when a planning/scheduling system fails during the problem solving, or when the solution is found to be inconsistent due to the introduction of new world state information, the user is not properly supported and left alone to determine the reasons for the break (e.g., no solution exists, the particular algorithm did not find a solution, there was a bug in the solver, etc.). To cope with this lack of communication, the concept of *explanation* is brought into play. Indeed this concept has been of interest in many different research communities. Explanations, by virtue of making the performance of

a system transparent to its users, has been demonstrated influential for user acceptance of intelligent systems and for improving users' trust in the advice provided (Hayes-Roth & Jacobstein 1994).

Our work is aimed at studying the willingness of users' to rely on explanation. In previous research expectation of failures and perceived anomalies have been identified as an occasion for accessing explanations (Chandrasekaran & Mittal 1999; Gilbert 1989; Schank 1986). In accordance with these findings we formulate the following hypotheses related to the users' willingness to rely on explanation:

Hypothesis 2. *The access to explanation is more frequent in case of failure than in case of success.*

Hypothesis 2b. *The access to explanation is positively associated with the number of failures and negatively associated with the number of successes.*

We were also interested in understanding possible relationship between explanation recourse and solving strategy selection (mixed-initiative vs. automated). Previous studies investigated the role of explanations in cooperative problem solving (Gregor 2001) showing how participants make a greater use of explanations. Results presented in Gregor's paper are related to participants assigned by the experimenter to two different conditions (automated and collaborative problems solving). A difference in our study consists in the fact that the two conditions are chosen by the participants themselves. It is our intuition that users who choose the mixed-initiative approach possess a higher level of control in the problem solving, thus showing a lower need to access the explanation. For this reason we formulate the following hypothesis:

Hypothesis 3. *Access to explanation is related to the solving strategy selection. In particular, participants who choose the automated solving strategy rely more frequently on explanation than subjects who choose the mixed-initiative approach.*

The relationships between the user's level of expertise and recourse to explanation has been an additional interest of our study. Previous research proved that experts are more likely to use explanations for resolving anomalies (Mao & Benbasat 1996) or because of unexpected conclusions (Ye 1995).

In our case, we expect that non-experts will use explanation more frequently than experts, and thus we formulate the following hypothesis:

Hypothesis 4. *During problem solving, non-experts access to explanations more frequently than experts.*

It is finally plausible to hypothesize that the difficulty of problems will affect recourse to explanation. Specifically, we expect that the more difficult the problem, the more likely users will access explanation. In particular we hypothesize that:

Hypothesis 5. *Access to explanation is more frequent in case of difficult problems.*

The explanation messages used in our experiment describe and explain the reasoning behind the choices made by the problem solver. They are expressed in textual form and have a user-invoked provision mechanism.

Realizing the Experiments

This section first gives some additional information on the solver services that act as a basis for the experiments, then describes the various choices needed to create the settings for the experiments.

The reference solver

As mentioned, the experiments are designed on the basis of COMIREM's services. This planning and scheduling architecture[2] provides the user with the two options of either automatically generating a solution to the problem or iteratively building a solution. The automated resolution is based on an opportunistic constraint-posting scheduling procedure to allocate resources to activities over time, relying on a planning sub-procedure as necessary to determine appropriate resource reconfiguration actions. The system takes as input an initial *plan sketch* that specifies, at some level of abstraction, the actions needed to accomplish the goals for a given scenario. Starting from this initial plan, the scheduling procedure tries to feasibly allocate resources to input activities. If successful, the procedure returns a detailed plan, where each activity is assigned the resources it requires and is designated to execute in a specified finite time interval.

Due to its interactive nature, the system can exploit human-planner knowledge and decision making, and in fact promotes a mixed-initiative process. Through an Interaction Module it is possible to iteratively build a solution through a *step by step* procedure that interleaves human choices with system calculation of consequences. When an initial plan is loaded, COMIREM performs a temporal feasibility check, and creates new activities as necessary to carry out entailed supporting actions. A visual representation of the problem and its main features is provided to the user through a graphical spreadsheet-style model. For each unassigned activity in the plan, the system maintains the current set of feasible allocation options and presents them to the user through the Interaction Module. At any time and in any order, the user can manually specify resource assignments for particular activities. Whenever a user allocates a resource to a given activity, the impact of the user's choice is reflected in the plan and the system updates the set of possible options available for other pending decisions.

A web-based simulated version of COMIREM was developed for the purpose of our experiments. The specific application scenario was a TV broadcasting station domain. The simulation reproduced the two alternative solving procedures on the sample of problems which were proposed to the participants. By choosing the automated method, users could inspect the result provided by the system and possibly access explanation for problem solving choices. Choosing the interactive strategy entailed that users had to build a solution based on their own choices, and at each step they were given the opportunity to inspect the system's calculation of consequences, possibly accessing explanation. Indeed, the simulated version contains only information relevant to the

---

[2] For the sake of completeness, we insert here a partial description of COMIREM's functionalities. The interested reader should refer to the original work (Smith, Hildum, & Crimm 2005) for an exhaustive presentation.

purpose of the study. The design and implementation of the tool has been accomplished considering possible usability problems of the Interaction Module that would act as undesired biases for the subject. To this end an iterative usability test, based on the Thinking Aloud methodology (Nielsen 1993), has been performed on the user interfaces before the final experiments. The results of this usability test allowed to discover and solve interaction problems, and simplify some other aspects before the experimental study.

Experimental design

Once the reference apparatus is set up, the experimenter has to decide the *independent variables* (i.e., variables manipulated by the experimenter, and that cause changes in behavior) and the *dependent variables* or *measures* (i.e., variables that are observed, measured and recorded by the experimenter). As the name suggests, the latter depend on the behavior of the participant, which, in turn, depends on the independent variables.

A further decision an experimenter must make is how to assign subjects to the various levels of independent variables. The two main possibilities are to assign only some subjects to each level, or to assign each subject to every level. The first possibility is called a *between subjects* design, and the second a *within subjects*.

As already mentioned before the primary independent variables considered in our experiments are *expertise* and *problem difficulty*. Our general choice has been to consider *expertise* as a *between* factor with two levels, expert or non-expert, while the *problem difficulty* represents a *within* factor with two levels, low and high. A further independent variable is represented by *failure* during the problem solving. This last variable has two levels, present or absent.

As general measures, two main independent variables have been considered, namely the choice of the solving strategy and the frequency of access to explanation. In particular, with respect to the solving strategy, two general scores were computed (*n_auto* and *n_mixed*). They measure the overall frequency of choice of each strategy in the experiment.

As for access to explanation, the following indexes were calculated:

– *access_failure* which represents the frequency of access to explanation in case of failure during problem solving;

– *access_success* which measures the frequency of access to explanation in case of correct decision during problem solving;

– *access_easy* indicating the frequency of access to explanation in case of easy problems;

– *access_hard* indicating the frequency of access to explanation in case of difficult problems.

A web-based tool

The web-based apparatus, inspired by COMIREM, allowed us to extend the experiments to a large sample of participants. The experimental tool is accessible through a web browser and is organized as follows:

– *Presentation*: A general description of the study and the list of software requirements.

– *User data input form*: Data collected through this input form was registered in a data base implemented in MySQL. For each participant, the following data was recorded: identifier, profession, education, sex, age, language, expertise in planning & scheduling.

– *Instructions*: A list of instructions to be followed during the experiment.

– *Training session*: This session was implemented through a sequence of animated web pages showing the actions necessary to use the system. The layout of the screen was subdivided into two parts. On the left part the list of instructions was presented, which described the interface of the system and called upon the users to actively use the system. The right part of the screen was devoted to presenting the Problem Solver and its behavior consequently to user actions. The training session also allowed users to practice and gain experience with the system.

– *Session 1*: This session was implemented through a sequence of web pages showing an instance of a scheduling problem to be solved. A textual description of the problem was shown, followed by a graphical representation. Consequently to the user's actions, the system showed updated results.

– *Questionnaire 1*: an 11-item questionnaire was presented at the end of the first session. The questionnaire was subdivided into three sections:

1. the first section was devoted to the *manipulation check* of the variable *difficulty* (i.e., to check the validity of our classification of problems into easy and hard instances);

2. the second section was devoted to verifying how clear the two description modalities (textual and graphic) were;

3. the last section was aimed at investigating users' strategy selections and the reasons for their choices.

The first two sections included 6 items on a 5-step Likert type response scale (from "not at all" to "very much"). For the remaining items, related to reasons for the strategy selection, participants were asked to choose among different options. Participants were given the possibility to indicate possible suggestions or general comments.

– *Session 2*: This session was implemented through a sequence of web pages showing the instance of a scheduling problem to be solved.

– *Questionnaire 2*: The first three sections were the same as for Questionnaire 1. In addition, a fourth section was added with the aim of investigating the access to explanations during the whole experiment and the perceived utility of explanation services. Questions related to explanations were evaluated on a 5-step item Likert scale.

Questions related to the explanation recourse have been deliberately included solely in the last questionnaire so as to prevent users from being influenced in their access to explanation during the problem solving session.

Participants and experimental procedure

A group of 96 subjects participated in the study. The sample was balanced with respect to expertise in planning and scheduling (40 experts and 56 non experts) and with respect to gender, education, age and profession. All subjects participated in the experiment by connecting from their own computer to the experiment web site.

At the beginning of the experiment, the animated tutorial provided subjects with instructions on how to use the software, and showed which type of problems were to be solved. Then, it solved an example of scheduling problems by using both the automated and the mixed-initiative procedure. Participants could repeat the tutorial session until they felt confident with the use of the system. Then a problem was presented to the subjects and they were asked to choose between one of the two available solving strategies. During problem solving, participants could either access explanations through the *explanation* button or go to the next step. The user's interactions with the system were registered in the data base. At the end of the first session subjects were asked to fill in Questionnaire 1. The same procedure was followed for Session 2. In order to avoid effects due to the order of the presentation, the two sessions (which corresponded to different degrees of difficulty) were randomly presented to users.

Stimuli

As mentioned above, the stimuli presented to participants consisted in four scheduling optimization problems. The design choice of using a relatively small set of problems is motivated by the evident need of presenting an overall task which was not too tedious for the users so as to be sure that all participants could complete the whole experiment. Indeed, when an experimental study is conducted with real users, a trade-off exists between the complexity of the experimental design and the time needed to complete the whole experiment. Our choice has been that users could not be asked to spend more than one hour at a time on a single tasks, nor more than two hours in a single day, the duration including also the training to perform tasks. This temporal constraint motivated the choice of limiting the number of stimuli.

For each problem users had to provide a final schedule of activities assigned to resources so as to minimize the cost of the overall schedule. Two solvable problems (one easy and one hard) were presented during the first and the second session to all subjects, and two unsolvable problems (one easy and one hard) were presented only to subjects who chose the automated procedure. The reason for adding these further problems in case of automated selection is twofold:

– the mixed-initiative selection entailed more time to solve problems. In this way all subjects had a comparable workload in term of time spent in solving problems.

– the mixed-initiative selection entailed that almost all participants encountered some failures during the problem solving, thus introducing unsolvable instances (failure) which were also necessary to the automated procedure.

Results

This section reports the data gathered with the experimental apparatus subdivided according to the two main questions.

Automated strategy vs. mixed-initiative

A between subjects ANOVA was performed to test the influence of *expertise* on the solving strategy selection, separately for the two different strategies. We used as dependent variables the two indexes previously introduced: *n_auto* and *n_mixed*. Results show a significant effect of the variable *expertise* ($F_{(1,94)} = 20.62$ $p < .001$).[3] In particular we found that experts rely more often on the automated procedure, while non experts seem to prefer the mixed-initiative problem solving (see Table 1).

	expertise	N	Mean	Std. Dev.
	Non-experts	56	.6786	.7653
n_auto	Experts	40	1.3750	.7048
	Total	96	.9688	.8137
	Non-experts	56	1.3214	.7653
n_mixed	Experts	40	.6250	.7048
	Total	96	1.0313	.8137

Table 1: Influence of expertise on solving strategy selection (statistics)

A χ^2 test was performed to test Hypothesis 1b, separately for easy and hard problems. A significant effect was found in the first case (χ^2=9.80, df=1, $p < .01$). In particular, the analysis of standardized residual shows that when solving easy problems, experts prefer the automated strategy, while non-experts prefer the mixed-initiative approach (see Fig. 1).
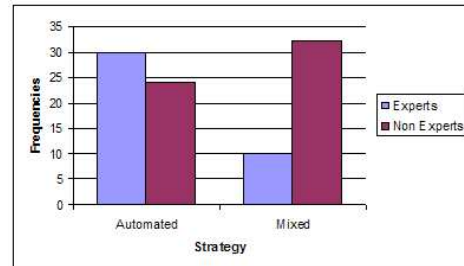

Figure 1: Strategy selection preferences: easy problems

No significant significant difference has been found between the two groups in case of difficult problems (χ^2=3.6, df=1, n.s.) (see Fig. 2).

We also analyzed answers to questionnaires related to the reasons for choosing one strategy or the other. The statistical analyses are not presented here for the sake of space.

[3] The F-ratio (F of Fisher) represents the ratio between the *Between* variance and the *Within* variance. The greater the F value, the greater the difference between the means. Associated to the F ratio there is the p value which represents the probability of making a mistake refuting the null hypothesis (i.e., the means are the same). As a consequence, the value of p indicates the level of confidence with which we can assert the validity of the alternative hypothesis (the difference in the means). Usually, a conventional threshold for p is decided, (e.g., 0.05), and if p if less than this value, the null hypothesis is neglected.
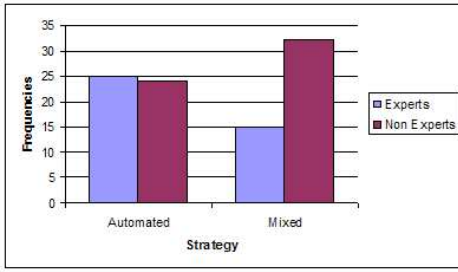
Figure 2: Strategy selection preferences: difficult problems

	N	Mean	Std. Deviation
automated	54	.8769	.3373
mixed-initiative	42	.2802	.3202
total	96	.6158	.4430

Table 3: Index of access to explanation: easy problems

	N	Mean	Std. Deviation
automated	49	.6297	.2959
mixed-initiative	47	.2790	.2709
total	96	.4580	.3329

Table 4: Index of access to explanation: difficult problems

Hoverer, results show how the reasons for choosing the automated approach as opposed to mixed-initiative solving is generally the same both for experts and non-experts. In particular, in case of easy problems, both experts and non-experts choose the automated procedure because they trust the system, while they rely on the mixed-initiative approach to maintain control over the problem solving. The reason for choosing the mixed-initiative approach remains the same in case of difficult problems, while a significant difference has been found for the automated choice: while experts choose this approach because they trust the system, non experts rely on the automated procedure in order to not waste time.

Access to explanation

To assess the relationship between failures and access to explanation, a repeated-measures ANOVA was performed using as dependent variables the indexes *access_failure* and *access_success*, previously defined. Results show a significant effect of failure on the access to explanation, $F_{(1,89)} = 85.37$, $p < .001$. In particular, users access explanation more frequently in case of failure than in case of success (see Table 2).

Additionally, a correlation analysis between the number of failures (and successes) and the number of accesses to explanation was performed in order to test Hypothesis 2b. Results show a significant correlation between failures and number of accesses to explanation (r = .86 $p < .001$). No significant correlation between number of correct choices and number of accesses to explanation has been found (r=.035, n.s.).

	N	Mean	Std. Deviation
access_failure	90	.8111	.3716
access_success	90	.3702	.3354

Table 2: Access to explanation (statistics)

To test Hypothesis 3, aimed at investigating the relationship between the strategy and the recourse to explanation, an ANOVA for independent groups was performed separately for the two levels of difficulty. The indexes *access_easy* and *access_hard* previously defined were used as dependent variables. Results show a significant effect of the strategy selection on the recourse to explanation. In particular, both for easy ($F_{(1,94)} = 77.26$ $p < .001$), see Table 3 and hard problems ($F_{(1,94)} = 36.60$ $p < .05$), see Table 4, access to explanation is higher when the automated strategy is selected.

Finally, to test our last two hypotheses, a mixed-design ANOVA was performed choosing *expertise* as a between-subjects factor and *difficulty* as a within-subjects factor. We used the indexes *access_easy*, and *access_hard* as dependent variables. A significant effect of expertise on recourse to explanation has been found ($F_{(1,94)} = 7.34$, $p < 0.01$). Experts were shown to access explanation significantly more than non-experts. An effect of problem difficulty on the recourse to explanation was also found ($F_{(1,94)} = 12.54$, $p < .01$). Access to explanation was shown to be significantly higher when an easy problem is to be solved. No significant interaction effect was found ($F_{(1,94)} = .002$, n.s.) (see Table 5).

	expertise	N	Mean	Std. Dev.
	Non-experts	56	.5423	.4760
access_easy	Experts	40	.7187	.3740
	Total	96	.6158	.4430
	Non-experts	56	.3829	.3177
access_hard	Experts	40	.5632	.3289
	Total	96	.4580	.3329

Table 5: Index of access to explanation: effect of expertise and problem difficulty

Discussion

The overall results of the present research are consistent with the expectation that non-expert users prefer the mixed-initiative approach rather than the automated strategy, while experts rely more frequently on the automated strategy. Moreover, explanation is frequently used and the frequency of access is higher in case of failure than in case of success.

More specifically, non-expert users show a tendency to actively solve problems keeping control over the problem solving process. This result can be considered in accordance with the idea that non-experts tend to be skeptical toward the use of an automated system, probably because they do not completely trust the solver capabilities. Conversely, expert users show a higher trust toward the automated solver. Expert users are usually system designers and are used to implement algorithms, thus knowing how effective machines can be in solving problems.

Results also confirmed previous studies (Gilbert 1989; Schank 1986), according to which access to explanation is more frequent in case of failure. These findings are consistent with some intuitions in the field of mixed-initiative sys-

tems, to consider system failures in achieving some goals as a specific occasion for providing explanation (see (Bresina *et al.* 2005)). Furthermore the main reason for accessing explanation seems to be the willingness to "understand" the artificial solver. Interestingly we found that, as expected, the more the failures the more the accesses to explanation; on the other hand no relationship was found between successful solving and access to explanation. As a consequence it is possible to assert that success is not predictive of any specific behavior with respect to access to explanation.

Hypothesis 3, which asserts a greater use of explanation in case of automated solving strategy selection, has been confirmed. In both sessions of our experiment it was found that participants who chose the automated strategy, access explanation more frequently than subjects who chose the mixed-initiative approach. It is possible to speculate that by selecting the mixed-initiative approach, subjects actively participate in the problem solving and keep a higher control on the solving process. As a consequence, the need for explanation might decrease. Conversely, participants who chose the automated strategy delegate the artificial solver but at the same time they need to understand solvers's choices and decisions. A somewhat surprising finding of the study was that experts access explanation more frequently than non-experts; in addition, the access to explanation is more frequent when facing an easy problem than in case of a difficult problem.

Implications for practice

This paper has described an experimental approach to evaluate some key features of mixed-initiative problem solvers. Our long term goal is to create a path toward establishing a methodology to compare features of such systems, and, in a future perspective, to compare different systems or specific solutions to the same task.

At present we have inherited the experience from disciplines that study the behavior of human beings (e.g., psychology and human-computer interaction), and slightly adapted them to the specific case. The same approach can be followed to broaden testing on interactive features. It is worth mentioning that to obtain experimental validity, a consistent amount of work must be put behind the logical design of the experiments. For this reason, a mix of competencies is required.

Quite interesting are the implications of the current findings for future practice. In particular, we paid attention to basic user attitude concerning the choice of automated rather than interactive strategies, as well as the bias toward the use of explanation. As a result, we have empirically proved that the mixed initiative approach responds to the willingness of end users to maintain control over automated systems. Conversely, expert users prefer to entrust the system with the task of problem solving. The existing difference between individuals with different levels of expertise highlights the need for different styles of interaction in the development of intelligent problem solving systems.

Our work also demonstrates the utility of explanation during problem solving, and the achievement of a *failure* state has been identified as a main prompt to increase the frequency of explanation access. One aspect related to explana-

tion that is worth reminding is the increased use by experts, who can also often actually contribute to the problem solving cycle with their expertise. This strengthens one open issue in the research agenda for mixed-initiative problem solving, namely explanation synthesis. This aspect is currently under-addressed, and deserves further investigation. Notice that our investigation has focused on the generic use of explanation. The answer to the more specific question "what is good explanation" is one we have left open for future studies.

Conclusions

Previous work in the area of mixed-initiative problem solving has been mainly focused on designing models, developing *ad hoc* systems and conceiving interactive features to foster human-system collaborative problem solving. Most of these studies are based on the assumption that this new solving paradigm is useful and appreciated by users. There is a widespread tendency of presenting users' needs for maintaining control over the machine as a proof of utility of mixed-initiative interaction. However, no empirical evidence to support this statement has been provided. The work presented in this paper shows empirically that the mixed-initiative approach responds to a specific need of real users. In particular, non-expert users have been shown to prefer an incremental and interactive procedure to build solutions rather than a completely automated approach. Indeed, the main reasons for this preference relies on the desire to be personally involved in the solution process.

Besides this important result, our study contributes with an investigation of different aspects of the mixed-initiative paradigm, identifying several issues that should be taken into account for designing future systems. It also highlights the importance of this kind of evaluation effort for mixed-initiative planning and scheduling technology. Indeed, the general desiderata of system designers consists in obtaining tools that can be fruitfully adopted in real world contexts. To this end, it is necessary to take into consideration potential needs of end users, and to gain a deeper understanding of how users can interact with decision support tools.

Our study has availed itself of a rigorous methodology, which is inherited from experimental psychology research. It is worth noticing that the effort to apply established methodologies such as the above is rather time consuming but extremely precise and useful. To this end, some work still remains to be done in order to better understand how to speed up and facilitate the application of this kind of methodology in the specific context of mixed-initiative system evaluation, and to understand further the generality of the outcomes.

Several points remain open for future investigation. We would like to use the same experimental apparatus to evaluate different types and depths of explanation, as well as the influence of access to explanation or solving strategy selection on problem solving performance. It would be also interesting to apply the same experimental methodology to study other aspects of mixed-initiative interaction considering also different domains, in order to validate the generality of the results presented here.

Acknowledgements

References

Ai-Chang, M.; Bresina, J.; Charest, L.; Chase, A.; Hsu, J.; Jonsson, A.; Kanefsky, B.; Morris, P.; Rajan, K.; Yglesias, J.; Chafin, B.; Dias, W.; and Maldague, P. 2004. MAPGEN: Mixed-Initiative Planning and Scheduling for the Mars Exploration Rover Mission. *IEEE Intelligent Systems* 19:8–12.

Anderson, D.; Anderson, E.; Lesh, N.; Marks, J.; Mirtich, B.; Ratajczack, D.; and Ryall, K. 2000. Human-guided simple search. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI 2000), Austin, Texas, USA. AAAI Press*, 209–216.

Bresina, J. L.; Jónsson, A. K.; Morris, P. H.; and Rajan, K. 2005. Mixed-Initiative Planning in MAPGEN: Capabilities and Shortcomings. In *Proceedings of the ICAPS-05 Workshop on Mixed-initiative Planning and Scheduling, Monterey, CA*.

Burstein, M., and McDermott, D. 1996. Issues in the development of human-computer mixed-initiative planning. In Gorayska, B., and Mey, J., eds., *Cognitive Technology*. Elsevier. 285–303.

Chandrasekaran, B., and Mittal, S. 1999. Deep versus compiled knowledge approaches to diagnostic problem-solving. *Int. J. Hum.-Comput. Stud.* 51(2):357–368.

Cohen, R.; Allaby, C.; Cumbaa, C.; Fitzgerald, M.; Ho, K.; Hui, B.; Latulipe, C.; Lu, F.; Moussa, N.; Pooley, D.; Qian, A.; and Siddiqi, S. 1998. What is initiative? *User Modeling and User-Adapted Interaction* 8(3-4):171 – 214.

Ferguson, G., and Allen, J. F. 1998. TRIPS: An integrated intelligent problem-solving assistant. In *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI 1998), Madison, Wisconsin, USA. AAAI Press*, 567–572.

Gilbert, N. 1989. Explanation and dialogue. *Knowledge Engineering Review* 4(3):205–231.

Goodwin, C. J. 2005. *Research in Psychology: Methods and Design. 4th ed.* Hoboken, NJ: John Wiley & Sons.

Gregor, S. 2001. Explanations from knowledge-based systems and cooperative problem solving: an empirical study. *International Journal of Human-Computer Studies* 54:81–105.

Hayes, C. C.; Larson, A. D.; and Ravinder, U. 2005. Weasel: A MIPAS System to Assist in Military Planning. In *Proceedings of the ICAPS-05 Workshop on Mixed-initiative Planning and Scheduling, Monterey, CA*.

Hayes-Roth, F., and Jacobstein, N. 1994. The state of knowledge-based systems. *Communications of the ACM* 37:27–39.

Jones, D. R., and Brown, D. 2002. The division of labor between human and computer in the presence of decision support system advice. *Decision Support Systems* 33:375–388.

Jussien, N., and Ouis, S. 2001. User-friendly explanations for constraint programming. In *Proceedings of the ICLP'01 11th Workshop on Logic Programming Environments, Paphos, Cyprus*.

Kirkpatrick, A.; Dilkina, B.; and Havens, W. 2005. A Framework for Designing and Evaluating Mixed-Initiative Optimization Systems. In *Proceedings of the ICAPS-05 Workshop on Mixed-initiative Planning and Scheduling, Monterey, CA*.

Langer, E. J. 1992. Matters of mind: Mindfulness/mindlessness in perspective. *Consciousness and Cognition* 1:289–305.

Mao, J., and Benbasat, I. 1996. Exploring the use of explanations in knowledge-based systems: a process tracing analysis. Working Paper 96-MIS-002. Faculty of Commerce, University of British Columbia, Canada.

Myers, L. K.; Jarvis, P. A.; Tyson, W. M.; and Wolverton, M. J. 2003. A mixed-initiative framework for robust plan sketching. In *Proceedings of the 2003 International Conference on Automated Planning and Scheduling (ICAPS-03) Trento, Italy*.

Nass, C., and Moon, Y. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues* 56:81–103.

Nielsen, J. 1993. *Usability Engineering*. San Diego, CA: Academic Press.

Schank, R. C. 1986. Explanation: A first pass. In Kolodner, J. L., and Riesbeck, C. K., eds., *Experience, Memory and Reasoning*. Erlbaum Associates, Hillsdale, NJ. 139–165.

Smith, S.; Cortellessa, G.; Hildum, D.; and Ohler, C. 2005. Using a scheduling domain ontology to compute user-oriented explanations. In Castillo, L.; Borrajo, D.; Salido, M.; and Oddi, A., eds., *Planning, Scheduling, and Constraint Satisfaction: From Theory to Practice*. IOS Press.

Smith, S. F.; Hildum, D. W.; and Crimm, D. R. 2005. Comirem: An Intelligent Form for Resource Management. *IEEE Intelligent Systems* 20:16–24.

Wallace, R., and Freuder, E. 2001. Explanation for Whom? In *Proceedings of the CP-01 Workshop on User-Interaction in Constraint Satisfaction, Paphos, Cyprus*.

Ye, L. R. 1995. The Value of Explanation in Expert Systems for Auditing: An experimental Investigation. *Expert Systems with Applications* 9(4):543–556.