Towards a Reliable Evaluation of Mixed-Initiative Systems

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 are being applied in different real world domains and many systems have been developed to address specific problems. Though several successful examples of such tools encourage the use of this solving paradigm, it is worth highlighting that research in mixed-initiative interaction is still at an early stage and many important issues need to be addressed. In particular, while some work has been devoted to the design of working prototypes and to identify relevant features of the mixed-initiative interaction, little attention has been given to the problem of evaluating the approach as a whole and the diverse aspects involved. This work aims at highlighting the need for effective evaluation studies for this class of tools and provides a methodological contribution in this direction. In particular it uses an experimental methodology well known in psychology and human-computer interaction for the problem of understanding users' attitude with respect to mixed-initiative problem solving and investigates the importance of explanation services as a means to foster users' involvement in the problem solving.

Introduction

Several real world domains, such as manufacturing, space, logistics and transportation have demonstrated how the use of computer-based application to support users be useful and convenient. Automated techniques can relieve humans from solving hard computational problems saving their "cognitive energy" for higher level decision tasks.

Nonetheless, the introduction of intelligent systems for solving complex problems has been characterized by the raising consciousness that in most cases a completely automated approach is neither applicable nor suitable for a successful deployment of solving technologies. As a matter of fact, automated problem solving is difficult to integrate into human-centric activities, for both technical and psychological reasons.

Although there are certainly some exceptions, total automation of decision-making is not an appropriate goal in most practical 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 practical domains overwhelms the solving capabilities of automated planning and scheduling technologies, and some sort of problem decomposition and reduction is required to achieve solver 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.* 1999), is receiving increasing attention and interest.

This emerging paradigm fosters the human-computer cooperation during the resolution of complex problems. The approach is based on the idea that experienced users and automated technologies possess complementary abilities and the goal is to fruitfully integrate them to obtain a more efficient system. An integrated system ⟨*human, artificial solver*⟩ can create a powerful and enhanced problem solver applicable to the resolution of difficult real world problems.

It is worth saying that current proposals for mixed-initiative systems are very often presented as system description and developed on purpose for *ad hoc* scenario. 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 studying users' attitude 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, nonetheless, no empirical evidence is given to support this statement. This paper would like to contribute in this direction.

This work is driven by an additional observation that while 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 interaction 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 becomes crucial and the integration of advanced services such as explanation functionalities, what-if analysis, assumes an important role.

This paper is aimed at providing a methodological contribution to the synthesis of mixed-initiative system. In particular it 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. an automated approach; (b) users' willingness to rely on explanation as a mean to maintain the control on the machine; (c) possible individual differences between experienced an inexperienced users. In general we would like to stress the need of designing effective evaluation studies for evaluating this class of interactive systems and describe our particular experience on the subject.

Plan of the paper. In the following we first summarize the state-of-the-art in mixed initiative systems and highlights some research aspects that would deserve attention. Then we describe our work that inherits features from experimental research in psychology and human-computer interaction. We first set up an experimental apparatus, that design our experiments formulating hypothesis, gathering data, and then interpreting them. A section discussing implication for practice that this approach may have 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 the last 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 promote a problem solving style centered on the idea of a system as an intelligent black-board, where a user can posts her/his 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 compare different partial solutions. A similar opportunistic approach to plan development is used in TRIPS (Thinking, Reasoning, and Intelligent Problem Solving) (Ferguson & Allen 1998), an Integrated Intelligent Problem Solving Assistant which has been designed for use in transportation-related planning domains.

PASSAT (Plan Authoring System Based on Sketches, Advice and Template) (Myers *et al.* 2003) has recently introduced the concepts of plan sketches and supports a user and the system working collaboratively to refine plan sketches to a satisfactory solution.

A more specific system is the one developed at the Mitsubishi Electric Research Laboratory (Anderson *et al.* 2000), which proposes an effective and interactive schema called *human-guided simple search* devoted to the solution of a well-known and difficult combinatorial and optimization problem. The human-guided search paradigm allows users to explore many possible solutions in order to better understand the trade-offs between possible solutions and then choose a solution based on their understanding of the domain. Users can manually modify solutions, backtrack to previous solutions, and invoke a portfolio of search algorithms. More significantly, users can constrain 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 system and 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.) that promotes an interactive and incremental construction of a solution.

A somewhat missing issue which derives from this brief overview is the one related to the evaluation of such systems. 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. The diversity and complexity of the two involved entities, the human user with her/his unpredictable and sophisticated reasoning and the artificial machine, with its computational complexity and technicality, together with the uncontrollability and uncertainty of the environment, makes it difficult the design of precise and effective evaluation methodologies.

Which research topics for mixed-initiative systems?

Usually the synthesis of effective systems that exploit a certain methodology is instrumental for the establishment of any methodology as a 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 roadmaps for their solutions.

Although the concept of mixed-initiative systems has been recognized to be useful and many specialized meeting have been dedicated to it, the identification of open topics is still very subjective, with the consequent drawback of limiting the aggregation of different groups of researchers on the same topic. An attempt to identify 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) improvements of the underlining problem solving technology to better serve the mixed-initiative interaction; (b) empower the user with services that enhance their involvement and their active role. Example of the first type are the effectiveness of specific features of the underling constraint based technology (see the example of the features of the temporal network that represents the current solution, the role of preferences in connection with the same representation, etc.). As example of the second type, also underscored in the COMIREM paper, is the need for automated synthesis of explanations for the users. Indeed for systems that use a constraint based representation some works are appearing with initial interesting results (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 few works produce 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 (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 at contributing further on this specific issue.

Evaluating mixed-initiative systems

Once highlighted the need to direct research efforts toward a structured approach for the evaluation of mixed-initiative features, the problem remains of how to contribute with some specific effort. Somehow the lack of a principled design and of a robust evaluation methodology is expected being the theory of mixed-initiative a relatively recent effort. This paper, apart from further highlighting this open issues, proposes steps for systematic evaluation that rely upon a well-founded methodology to quantitatively analyze different features of these 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 system. This type of evaluation usually 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 capacitated-vehicle-routing-with-time-windows problem.
- *Involving users in the evaluation process*, that is evaluating different aspects related to the users' requirements and judgment on the system such as usability, level of trust of the system, clarity of presentation, etc. These aspects 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.

In our work we rely on an experimental methodology normally used in psychology and HCI and use it to evaluate various aspects of mixed-initiative. In particular we have set up a complete experimental methodology and used it to address two features.

A first general question we have addressed relates to the validity of the whole solving approach, that is the study of users' attitude toward the mixed-initiative approach in comparison with the use of a completely automated solution.

A second question is more specific. It is related to the emerging and interesting topic of the generation of automatic explanation. As already mentioned, mixed-initiative systems, implies a continuous communication between users and machines. Explaining 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 (Smith *et al.* 2005). We have been looking here to empirical evidence that proves the willingness of real users' to rely on explanation during interactive problem solving.

Setting up an empirical study

The design of the experimental methodology has focused on features of the COMIREM system (Smith, Hildum, & Crimm 2005). This is a web-based mixed-initiative problem solver devoted to the resolution of planning/scheduling problems. In accordance with the mixed-initiative theory, 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 automatic algorithms to efficiently solve problems and interactive facilities to keep the human solvers in the loop.

For this study we developed a simulated version of COMIREM, which is devoted to solve scheduling problem instances in a TV broadcasting station domain. The choice of developing a simulated version of the system is due to the willingness of extending the experimental evaluation to a large number of participants. This choice have forced us to simplify the system layout limiting to some extent the richness of information. However the generality of the proposed evaluation methodology does not rule out the possibility to use the methodology to test different features of the real system.

Once setting the experimental context we have formulated to motivating questions: 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 experts and non expert users? Is the difficulty of problem a relevant factor in the choice of the strategy or in accessing the explanation?

Once given the general questions the experimental methodology requires to carefully formulate hypothesis to be tested and the variables that should be monitored during experiments. Before giving these details a comment is worth doing. Although we are performing experiments relying on features of a specific system we are here interested to questions that are system independent, hence the validity of the current findings extends to analogous features in other systems.

Automated vs. mixed-initiative problem solving

In studying users' attitude toward using or not a mixed-initiative approach two main aspects have been considered as relevant factors in influencing users' choice, the problem *difficulty* and users' level of *expertise*. The first study aimed at investigating the influence of both these factors on the selection of mixed-initiative vs. automated strategy. In our research, 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 each action to the artificial solver, thus keeping 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 is some evidence that humans do not always adopt an optimal strategy in getting help from artificial tools, ignoring advices or solutions proposed by the system (Jones & Brown 2002). A possible explanation for this behavior is provided by some research in human-computer interaction area, reporting that humans tend to attribute a certain degree of anthropomorphism to computers, assigning to them human traits and characteristics. In (Langer 1992; Nass & Moon 2000) a series of experimental studies are reviewed, showing that individuals mindlessly apply social rules and expectations to computers. It is plausible to hypothesize that human problem solvers show the same tendency toward artificial solvers, and refuse to delegate the solution of the problem, for many reasons. For instance, they could mistrust the automated agent's ability to solve the problem or they could enter in competition with it. However, we have no data on possible differences in the behavior of users with different levels of expertise. 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 expert users 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 interactive 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 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 aims at studying the willingness of users' to rely on explanation. In previous research (Chandrasekaran & Mittal 1999) expectation of failures and perceived anomalies have been identified as an occasion for accessing explanations (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.*

In the context of knowledge-based systems, the role of explanations in cooperative problem solving has been investigated (Gregor 2001) and results show that participants in cooperative problem solving conditions, made a greater use of explanations. In accordance with the Mixed-Initiative Theory we hypothesize that the human solver, actively participating in the problem solving, possesses a higher level of control in the problem solving, thus showing a lower need to access the explanation. In particular 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 access more frequently the explanation than subjects who use the mixed-initiative approach.*

Based on the taxonomy of explanation types provided in (Chandrasekaran & Mittal 1999) and on a preliminary work on explanation generation within COMIREM (Smith *et al.* 2005), the explanation we will use in our experiment aims at explaining problem solvers' choices, it is expressed in textual form and has a user-invoked provision mechanism.

Evaluation Method

The general experimental design of this research aims at investigating the influence of the variables *expertise* and *problem difficulty* on the solving strategy selection (automated vs. mixed-initiative) and access to explanation. The variable *expertise* is a *between* factor with two levels, expert or non-expert, while the *problem difficulty* represents a *within* factor with two levels, low and high[1]. A further independent variable is represented by *failure* during the problem solving. This last variable has two levels, present or absent. As general measures, the choice of the solving strategy and the frequency of access to explanation have been considered. With respect to the solving strategy, two general scores were computed (*choice_auto* and *choice_mixed*). They measure the overall frequency of choice of each strategy in the experiment.

As for the 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_low_difficulty* indicating the frequency of access to explanation in case of problems of low difficulty;
- *access_high_difficulty* indicating the frequency of access to explanation in case of problems of high difficulty.

Tools

A web-based software has been developed, inspired by the software COMIREM. The simulated system allows users to solve instances of scheduling problems by means of two alternative procedures, automated and mixed-initiative. The system 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 were registered in a data base implemented in MySQL Language. For each participant the following data were registered: identifier, profession, education, sex, age, language, expertise in planning & scheduling and participant's problem solving pattern.
- *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 has been 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 use and practice the system.
- *Session 1*: It 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 presentation. 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*;
  2. the second section was devoted to verifying how clear the two description modalities (textual and graphic) were;
  3. the last section 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*: It 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 session was added designed for investigating the access to explanations during the experiment and their perceived utilities. Questions related to explanations were evaluated on a 5-step item Likert scale.

[1] The two levels of this variable have been determined considering the problems dimension in terms of number of activities to be scheduled and alternative resources available for each activity.

Participants and procedure

A group of 46 subjects was contacted, aged from 23 to 58 years (Mean 33,3). The sample was balanced with respect to expertise in planning and scheduling (23 experts and 23 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[2].

At the beginning of the experiment, an 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 the problem solving, participants could either access explanations through the *explanation* button or go to the next step. 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 the users.

Stimuli

Four scheduling problems were defined in the field of a broadcast TV station resources management. Two solvable problems (1 low difficulty and 1 high difficulty) were presented during the first and the second session to all subjects, and two unsolvable problems (1 low difficulty and 1 high difficulty) 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

Strategy Selection: automated vs. mixed-initiative

The first analysis investigated the influence of *expertise* on the solving strategy selection. A between subjects ANOVA was performed to test Hypothesis 1, separately for the two different strategies. The dependent variables used in this analysis were the indexes *choice_auto* and *choice_mixed* respectively. With respect to the strategy selection no significant difference was found (F(1,22)=1.94, n. s.).

To test Hypothesis 1b a χ^2 test was performed, separately for low and high level of difficulty. In case of low difficulty problems, a significant effect was found (χ^2=5.58, df=1, $p < .05$) . 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).

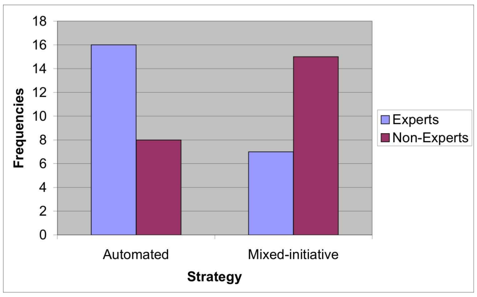

Figure 1: Strategy selection preferences: easy problems

The analysis shows no significant difference between the two groups in case of difficult problems (χ^2=0.11, df=1, n.s.) (see Fig. 2).

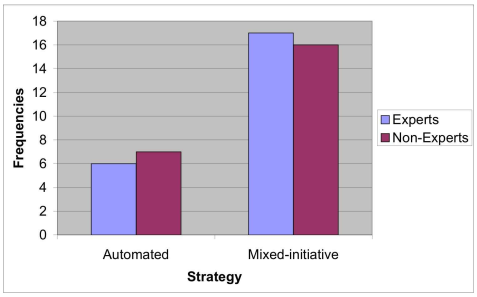

Figure 2: Strategy selection preferences: difficult problems

Access to Explanation

To test Hypothesis 2 which aimed at investigating the relationship between failures and access to explanation, a repeated-measures ANOVA was performed using as dependent variables the indexes *access_failure* and *access_correct*, previously defined. Results show a significant effect of failure on the access to explanation (F(1,41)=24.32, $p < .001$). In particular users rely on explanation more frequently in case of failure than in case of success (see Table 1).

Moreover, a correlation analysis between number of failures (and successes) and 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=0.623, $p < .001$). Conversely there is no significant correlation between number of correct choices and number of accesses to explanation (r=.01, n.s.).

Table 1: Access to explanation (statistics)

	N	Mean	Std. Deviation
access_failure	42	.6794	.4352
access_success	42	.3141	.3174

To test the relationship between the selected strategy and the access to explanation (Hypothesis 3), an ANOVA for inde-

[2] http://pst2.istc.cnr.it/experiment/

pendent groups was performed separately for the two levels of difficulty. The indexes *access_low_difficulty* and *access_high_difficulty* were used as dependent variables. A significant effect of the strategy selection on the recourse to explanation was found . In particular the access to explanation is higher when the automated strategy is chosen both in case of easy problems (F(1,43)=67.22, $p < .001$), see Table2, and in case of difficult problems (F(1,44)=10.97, $p < .05$), see Table 3.

Table 2: Index of access to explanation: easy problems

	N	**Mean**	**Std. Deviation**
automated	23	.8043	.2915
mixed-initiative	22	.1667	.2242
totale	45	.4926	.4128

Table 3: Index of access to explanation: difficult problems

	N	**Mean**	**Std. Deviation**
automated	13	.5769	.3444
mixed-initiative	33	.2625	.2666
total	45	.3513	.3204

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, the explanation is frequently used and the frequency of access is higher in case of failure than in case of success. More specifically, the study showed that non experts prefer the mixed-initiative procedure independently from the problem level of difficulty. Conversely, experts prefer the automated strategy when solving easy problems, while tend to move to the mixed-initiative approach while solving difficult problems.

As expected 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. One possible explanation, consistent with (Nass & Moon 2000), refers to the tendency to anthropomorphize machines and to believe that they can make mistakes just like human beings.

Conversely, expert users show a higher trust toward the automated solver. Nonetheless, they find it stimulating to actively participate in the problem solving process when a difficult task is given. The relevance of the problem difficulty emerged to be a key variable in their choice. Expert users are usually system designers and are used to implementing automatic algorithms, thus knowing how effective machines can be in solving problems. When an easy task is to be solved they are likely to consider the mixed-initiative approach as a time-wasting choice. On the other hand the idea of facing a puzzling problem can drive them to conceive alternative methods to generate solutions.

Results 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 systems, 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 will to understand the artificial solver (see (Cortellessa 2005) for more details on the motivations of this choice). 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 success 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, asserting a greater use of explanation in case of automated solving strategy selection, was 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 key features of mixed-initiative problem solvers. Our long term goal is to pave the way for stable methodologies 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 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 the testing on interactive features. It is worth mentioning that to obtain experimental validity a consistent amount of work stay behind the logical design of the experiments. For this reason a mix of competencies has been needed.

Quite interesting are the implications of the current findings for future practice. In particular we paid attention to basic users attitude concerning the choice of automated rather than interactive strategies and 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 keep 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.

It was also demonstrated 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 explanation that is worth reminding is the increased use by expert people that more often are those who may actually contribute most to the problem solving cycle with their expertise. This strengthen one open issue in the research agenda for mixed-initiative, current proposals with respect to synthesis of explanation are very initial but deserve further fostering. Notice that we have just investigated the generic use of explanation without entering in a more specific question on “what is a good explanation” a question that we have left open for future studies.

Acknowledgements. The acquaintance with COMIREM has been possible thanks to a visit of the first author to the Intelligent Coordination and Logistics Laboratory at Carnegie Mellon University. Special thanks to Steve Smith for supporting the visit and the whole COMIREM team for assistance. We would like to thank Maria Vittoria Giuliani and Massimiliano Scopelliti for their collaboration and support in setting up the experimental study. Authors’ work is partially supported by MIUR (Italian Ministry for Education, University and Research) under project RoboCare.

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 National Conference on Artificial Intelligence (AAAI 2000), Austin, Texas. 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*, 54–61.

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. 1999. What is initiative? In Haller, S.; McRoy, S.; and Kobsa, A., eds., *Computational Models of Mixed-Initiative Interaction*. Kluwer Academic Publishers. 171–212.

Cortellessa, G. 2005. *Problem Solving with Automated and Interactive Tools. Users’ choices and the role of automated explanation*. Ph.D. Dissertation, Cognitive Psychology Program, University of Rome ”La Sapienza”. (in Italian).

Ferguson, G., and Allen, J. F. 1998. TRIPS: An integrated intelligent problem-solving assistant. In *AAAI/IAAI*, 567–572.

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

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*, 18–27.

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 *ICLP’01 11th Workshop on Logic Programming Environments*.

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*, 5–11.

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

Myers, L. K.; Jarvis, P. A.; Tyson, W. M.; and Wolverton, M. J. 2003. A mixed-initiative framework for robust plan sketching. In *Procedings of the 2003 International Conference on Automated Planning and Scheduling (ICAPSi03) Trento, Italy*.

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

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 *CP01 Workshop on User-Interaction in Constraint Satisfaction*.