Improving decision-making performance through argumentation: An argument-based decision support system to compute with evidence

Joshua Introne, Luca Iandoli

Abstract

While research has shown that argument based systems (ABSs) can be used to improve aspects of individual thinking and learning, relatively few studies have shown that ABSs improve decision performance in real world tasks. In this article, we strive to improve the value-proposition of ABSs for decision makers by showing that individuals can, with minimal training, use a novel ABS called Pendo to improve their ability to predict housing market trends. Pendo helps to weight and aggregate evidence through a computational engine to support evidence-based reasoning, a well-documented deficiency in human decision-making. It also supports individuals in the creation of knowledge artifacts that can be used to solve similar problems in the same domain. An unexpected finding and one of the major contributions of this work is that individual unaided decision-making performance was not predictive of an individual’s performance with Pendo, even though the average performance of assisted individuals was higher. We infer that the skills activated when using the tool are substantially different than those enacted to solve the same problem without that tool. We discuss the implications this result has for the design and application of ABSs to decision-making, and possibly other decision support technologies.

1. Introduction

Numerous studies have characterized the ways in which people fail to make rational, consistent decisions in light of available evidence. Human decision-makers make generalizations on the basis of information that is more recent or familiar, ignore a priori probabilities, and overweight data that better fits stereotypes [48]. Cognitive limitations are often assigned blame for these difficulties; given our limited capacity to process information we tend not to carry out the proper search and assessment of evidence when making an inference and opt for easier, heuristic shortcuts [4,28].

Decision support technologies can address these difficulties by allowing people to keep track of larger amounts of information, and in some cases help people evaluate their information in a consistent and principled manner by applying mathematical decision theory. Yet these systems are not without their costs; they require training to learn [36] and, in the workplace, careful integration into existing organizational decision making routines [1,8,38]. Even once such tools are successfully integrated, they can be difficult for decision-makers to trust [22].

Argument based systems (ABSs) could address some of these problems. ABSs employ a visual formalism at the interface that embodies a handful of rhetorical primitives (e.g. claim, pro, con) that may be combined to create sophisticated decision structures. Because argumentation is a process that most people use in their daily lives, argument based interfaces may be easier to understand and more transparent than other kinds of decision support technology. There is also evidence suggesting that individuals benefit from being able to reflect upon the external representation of an argument structure [37] and experience has demonstrated that ABS can be a useful tool for storing and evaluating decision information in organizational contexts [7].

As reviews of argument system research [19,34] illustrate, most investigations have focused upon the representation of arguments and role of ABSs in the decision process rather than their impact on decision outcomes. Furthermore, in many cases where ABSs have been used to support decision-making, they have been applied to "wicked" problems that have no verifiably correct answer [8,13,26,37]. These research foci have made it difficult to determine if there are concrete payoffs to using argument-based systems to support decision-making.

In this article, we seek to address this gap by comparing the performance of individuals with and without an ABS in a complex, realistic decision task with an objective outcome—predicting housing market trends. Our findings demonstrate that the ABS employed in the study significantly improves individuals’ ability to make correct housing predictions when using the system compared to when using regular decision making routines.
market forecasts, and that the argument maps constructed by users to solve one set of problems outperform the users themselves on another set of problems from the same domain.

However, we also find that the decision performance of users without access to the tool is not predictive of their performance with it. ABS assisted individuals achieve performance scores that are normally distributed around a mean performance level that is significantly higher than the mean performance level for unassisted individuals, but uncorrelated with unassisted performance. Because the ABS under study has many similarities to other argument platforms and some other kinds of decision support tools, our results have interesting implications for the design of decision support tools.

In the following section we provide background on the benefits and limitations of ABS. We then motivate and describe our experimental tool. Finally, we describe a study in which subjects were required to solve a decision task with the support of the tool.

2. Argument-based systems and computational decision support

It is generally accepted that the practice of formal argumentation can improve the rationality of one's reasoning process. This has led researchers to develop argument-based systems (ABS) to support the construction of arguments using visual representations called argument maps. Early ABS platforms [43,51] were faithful to formal theories of argument, such as Toulmin's [47] theory of argumentation and Walton's presumptive reasoning [52]. However, many modern ABSs (e.g. DebateGraph,1 Rationale2 [12]) use somewhat more simplified representations, in part to address usability concerns [43].

2.1. Benefits of argument representation

Argumentation is a natural form of rhetoric that most people are already familiar with [25]. By the age of three, children can understand and apply the basic elements of argument, and children as young as seven use argumentation in their everyday life [40,41]. An interface based on these familiar rhetorical patterns should thus be easier to use than one based on more abstract mathematical formalisms like Bayesian belief networks and decision trees.

The practice of argumentation may be considered a form of quasi-logical reasoning that encourages context sensitive, grounded decision-making, and favors coherence [45], critical thinking [49], and evidential reasoning [5,30] over mathematical reasoning [47]. When embodied in a software tool, argument representations make certain types of relationships easier to see [44,46] and have been shown (with concomitant training) to improve critical thinking [12,29,46,49] and reflective assessment [29].

In decision-support contexts, ABSs may be particularly useful because they require the decision-maker to carefully articulate her rationale, and this process is considered to be an important step in many decision analytic techniques [2,38]. There is also evidence that practicing argumentation can help individuals make rigorous assessments of available information in the absence of objective statistical data [31].

Finally, argument maps can be used to build representations of domain knowledge that are easy to access and reuse [7,13,39]. Because argument maps follow a well-structured formalism, it may be possible to use them as a basis for semi-automated reasoning systems and to apply mathematical approaches to compute the relative support for the claims that are made [10,15,23]. We expand upon this point below.

2.2. Argumentation and decision-support

Despite their apparent potential, only a handful of studies have carefully examined ABS in such decision-making contexts [14,16,18,24], and there are few examples of ABS for decision support in the wild. We identify several reasons why argumentation research has focused on decision support, and argue that this is at least partially responsible for the dearth of commercial argument based decision support platforms.

From its inception, research on argumentation as a mediating artifact has focused on the structure and use of argument representations. Buckingham Shum [55] traces the idea of graphically depicting argument to John Henry Wigmore [53], who envisioned the design of a representation to capture the logic of legal argument. The embodiment of such representations was not realistic until the advent of the personal computers with graphical user interfaces. One of the earliest implemented argument platforms was the gIBIS system [6], which was created to help represent and evaluate design rationale. Such representations were novel at the time, and research consequently focused upon their design and use.

Subsequent research on argumentation has been carried out primarily within human–computer interaction and related fields and continues to focus on representational issues; the study of ABS is more commonly referred to as argument mapping, reflecting this bias. This has led to theories about the role of representations in guiding cognition [44,46], and to new approaches to training rational thought [11]. However, this work has also found that argument interfaces can be complex and difficult for users to understand, and that a significant amount of training is required if one is to reap their benefits [37]. For example, Van Gelder [12] attributes success in using argument maps to train critical thinking to a pedagogical strategy called LAMP (Lots of Argument Mapping Practice), and Conklin [7] has emphasized the importance of training, technical support, and enthusiastic advocacy for ABS within an organization.

In most of this work is that the intended effect of using an argument representation is a more rational thought process, rather than improved decision-making. Comparatively little work has sought to attach algorithms to argument structures to make decision support explicit, although there are several viable approaches (e.g. [9,15,18]), and this is a common design pattern among commercially available decision support tools (e.g. Bayesian Networks, Bayesian decision trees, the Analytical Hierarchy Process).1

Those studies that have sought to evaluate the efficacy of ABSs in decision-making contexts have used criteria other than decision performance. Hua and Kimbrough [14] demonstrated that a decision support system for helping people construct, evaluate, and communicate arguments can help people make correct inferential statements on abstract if-then logical statements, but never tied this improvement to decision making performance. Karacapilidis and Papadias [18] evaluated their HERMES platform for medical and engineering problem solving, but only reported that users enjoyed using the tool and that they thought it was helpful. Jarupathirun and Zahedi [16] showed how a dialectical argument system can help collaborative decision makers elicit underlying assumptions when reasoning with complex problems, but they do not extend their results to task performance.

Another reason few studies report directly on decision performance is the research community’s growing interest in “wicked” problems [32], which by their very nature do not present easily evaluated outcomes [8, 13,26,37]. Wicked problems are more ecologically valid than the toy decision tasks often used in controlled studies, but focusing only on problems with no objective solution makes assessment of decision support technology much harder.

---

1 http://debategraph.org/
2 http://rationale.austhink.com/
4 http://www.hugin.com/.
5 https://www.treeage.com/

Please cite this article as: J. Introne, L. Iandoli, Improving decision-making performance through argumentation: An argument-based decision support system to compute with evidence, Decision Support Systems (2014), http://dx.doi.org/10.1016/j.dss.2014.04.005
As it stands then, there are very few ABS decision-support platforms that offer explicit decision guidance. Research seems to indicate that ABSs can be used to train rational thinking if one is willing to invest significant resources in learning how to use them, but it is not clear what the payoff is in terms of decision-making performance. This presents the decision-maker or CIO with an unattractive value proposition: there is no clear evidence suggesting that ABS supports improved decision-making, but the costs are apparent and not clearly bounded. We believe it is possible to alter the value-proposition for ABS by focusing directly upon decision-performance. To do this, it is necessary to design a system that can provide unambiguous decision guidance, and identify realistic domains within which decision performance can be readily assessed.

3. Redesigning ABS for decision support

We designed a prototype system called Pendo (from the Latin verb pendere “to weigh” but also “to judge”, “to evaluate”) to help users distinguish between evidence and theory, and also to evaluate outcomes in light of different sets of evidence by computing the relative weight of competing claims under different evidence sets. Pendo uses a question/claim/pro-con structure that is derived from IBIS [32], a formalism that has been adopted by many ABSs [6,11,13,37,44]. Pendo includes the ability to attach evidence items to individual arguments, an interface feature that is also found in some argument mapping tools [5,11,29,44]. Unlike the majority of argument mapping tools, but similar to many computational decision-support tools, Pendo also includes a computational engine to evaluate user-generated arguments and calculate the weight for competing decision options. Thus, Pendo is a fairly typical argument-mapping tool, with some added functionality that makes it better suited for decision support.

In the following, we motivate the core design elements in Pendo, describe the computational engine embedded in Pendo, and describe the argument formalism and Pendo user interface.

3.1. Arguments and evidence

Research on ABSs has focused on supporting a representational process in order to mitigate less rational approaches to decision-making [17,28,48]. It is thought that ABSs may be able to do this by extending people’s capacity to store and reflect upon knowledge that is relevant to a decision and providing a structure that supports rational analysis. Yet, there remains little guidance about how best to design technology that helps people make decisions that are consistent with information that has been so encoded.

Empirical work by Deanna Kuhn [20,21] provides some guidance for developing this kind of support. Kuhn has carefully documented the ways in which people fail to practice “good argumentation” in evaluating their own everyday working theories. Her analysis yields a number of important insights, including:

1. People are typically unable to distinguish between the logic of an argumentative theory and evidence needed to support it.
2. People are not good at understanding how alternative sets of evidence might support different conclusions.

Kuhn suggests that for this kind of reasoning to occur, people require “a mental representation of the theory that can be acted on and evaluated” [20:267], and furthermore, that this representation should be distinct from the evidence that is used to support a decision. We conclude from Kuhn’s work that practical support for decision-making should: 1) help people distinguish between the argumentative theory and evidence used to support a claim; 2) provide a computational mechanism that allows users to express confidence in the relative strength of the available pieces of evidence supporting alternative claims; 3) help users to easily evaluate different sets of evidence to discover differential support for competing claims.

We designed our prototype to meet these design goals. In particular, our tool embeds a computational engine to satisfy requirements 2 and 3.

3.2. Pendo computational engine

The computational core of the system is based upon a technique described by Introne [15] that combines an argument map with a Dempster-Shafer belief aggregation [35] engine to assess the relative strength of competing claims in an argument. In the following, we describe this computational engine and the user interface of the Pendo system. Pendo’s includes:

1. A question that frames the decision
2. A set of competing claims or hypotheses
3. A set of pro or con arguments that respond to claims or other arguments.

In addition, arguments are constrained to have only one target, so that the formalism generates well-formed trees. A fully specified set of claims and arguments is often referred to as an argument map.

To provide support for assessing an argument map, we apply a belief aggregation procedure based on the Dempster-Shafer (D-S) theory. Like the Bayesian theory, the D-S theory is built on the axioms of probability. However, the unit of analysis in D-S theory is not a probability value but rather a belief function, which describes how a unit of belief is distributed over the power-set of mutually exclusive decision options (called the frame of discernment). The D-S theory offers several features that make it useful for users who do not have a mathematical background.

1. The D-S theory affords a lack of precision by allowing belief values to be assigned to sets of states without defining the apportionment of belief between these states.
2. The D-S theory allows some portion of belief to remain uncommitted (ignorance).
3. The Dempster rule of combination allows conflicting sets of evidence from multiple sources to be readily combined.

In order to apply the D-S theory to the argument formalism described above, Introne [15] associated each argument with a frame of discernment over “true” and “false” states, and the set of competing claims as states within a single frame of discernment (see Fig. 1). Belief

![Fig. 1. Combining the argument formalism with the D-S theory; heavy line = visual element in the interface, dashed line = frame of discernment, shaded box = non-visual element inserted by the algorithm.](image-url)
proportion occurs from the terminal nodes in an argument map to the claims at the root of the map. The topology of the connections between frames of discernment dictates how the support for individual arguments and claims is computed.

There are four ways in which frames might be connected:

1. An argument supports another (Fig. 1; Pro Arg A)—In this case, belief mass calculated for the true state of the supporting argument is propagated to the true state of the supported argument.
2. An argument refutes another (Fig. 1; Con Arg C)—The belief mass calculated for the true state of the refuting argument is propagated to the false state of the refuted argument.
3. An argument supports a claim (Fig. 1; Pro Claim 1)—The belief mass calculated for the true state of the supporting argument is propagated to the supported claim.
4. An argument refutes a claim (Fig. 1; Con Claim 1)—The belief mass calculated for the true state of the supported argument is propagated to the focal element consisting of all claims not including the refuted claim.

As a user encounters evidence for or against individual arguments, he may “post” his beliefs at these nodes, and these are propagated over connections in the map to determine the allocation of beliefs among the claims. Because Dempster’s rule of combination specifies how different belief functions may be combined, multiple pieces of evidence may be placed at each node.

In summary, the D–S based argument system used in Pendo can be thought of as a decision engine that is similar in function to a Bayesian belief net, but differs in a number of important ways:

- It requires less precise information about the specific probabilistic relationships between nodes, and no information about background (prior) probabilities.
- It is more conveniently thought of as a rhetorical representation of a chain of arguments that may be true or false with some certainty, rather than a map of causal relationships between random variables.
- It can handle competing evidence from multiple sources.

3.3. Argument formalism and user interface

Pendo is a prototype system that provides users with a typical argument interface on top of the D–S based aggregation engine described above. It provides affordances that help the user distinguish between the argumentative theory, and the evidence for individual arguments in the map. The user can save and recall different sets of evidence to evaluate different “what-if” scenarios. The system runs in a web browser, and all calculations are handled on the server. Many of the features that would be necessary in a true production system (e.g., secure logins, user management of previously stored maps) are absent in the prototype, but the core functionality is sufficient for experimental purposes.

As discussed in the previous section, three basic elements are used in constructing an argument map in Pendo: a question to be resolved, a set of evidence sessions, and a horizontal bar at the top of each node indicating the current belief assigned to that node. The valence of the connection between nodes is reflected in the color of the link (green for pro, red for con), and it is possible to set the strength of the connection by clicking the number on the link and moving a slider (varying from 1 to 10).

In order for the user to determine the relative support for competing claims, she must attach evidence to arguments. The prototype allows users to post three types of evidence (in order of decreasing weight): a “Fact”, an “Opinion”, or a “Guess.” Once evidence is attached, support for each claim is displayed using qualitative terms that are associated with intervals between zero and one. In the prototype version, the qualitative terms used were Unlikely (0−25), Possible (25−50), Likely (50−75), and Very likely (75−100).

A final piece of functionality in Pendo is that every piece of evidence is associated with a “session”, and users may create any number of such sessions. The list of evidence sessions is shown along the left hand side of the interface in Fig. 2; clicking on any of these will recall evidence from that session and update the interface accordingly.

The existence of sessions helps to address Kuhn’s recommendations in a couple of ways. First, it allows the user to switch rapidly between different sets of evidence to examine how different claims are supported under different conditions. Moreover, it affords a workflow whereby a user can incrementally refine an argument map over the course of several decision problem instances (e.g., encountering a new kind of evidence and updating the argument map accordingly). In so doing, the user is encouraged to become aware of the difference between the argumentative theory, which might be considered to be a general construct, and evidence, which may vary from instance to instance.

4. User study

4.1. Study’s objectives and hypotheses

As discussed above, we hope to improve the value proposition for ABS by demonstrating clear performance improvements in a realistic decision task, without requiring users to undergo extensive training. We developed a task in which participants were asked to forecast the short-term trend of the housing market, using real data drawn from the U.S. housing market. The task is non-trivial, realistic, has clear value, and it also has an objectively correct solution against which to compare performance, here defined as decision maker’s ability to produce an accurate prediction. This is a departure from many empirical studies on argumentation in which participants solve ill-structured problems and argumentation performance is measured with respect to platform usability or aspects of the argumentation process itself.

The primary aim of our study was to determine whether or not the use of Pendo would help subjects to make better use of their available information and improve their performance in the prediction task. We measured prediction performance as a subject’s ability to correctly answer the question “What will be the price trend at the end of the year?” using historical data as ground truth (discussed in more detail in the following section).

We anticipated that Pendo would improve our subjects’ forecasting accuracy because it was designed to address well-known human limitations in information processing (reasoning with evidence). Thus, we state our primary research hypothesis as:

\[ H_1 \] subjects will make more accurate predictions when supported by Pendo.
We were also interested to know if the argument maps developed by our users could be considered to be general argumentative theories about housing market trends. If this were the case, we would expect the argument maps to make accurate predictions on held out data. We state this research hypothesis as:

H2. the argument maps created by individual users will produce good predictions also when applied on a different data set for the same task.

Finally, we were interested to know whether the use of Pendo would result in users being better able to solve domain problems without the system. As we have discussed, much of the ABS research has been aimed at helping people adopt more rational thought processes, with an implicit belief that this would result in better performance in the domain task. However, this same research indicates that substantial training is required to achieve these outcomes, and we chose to eschew extensive training for our study. On the other hand, the type of decision support provided by Pendo extends beyond the representational guidance typically offered by other ABS because Pendo is designed to help people distinguish between an argumentative domain theory and the evidence presented in each problem instance, a deficiency documented by Kuhn [20]. We expected that the lack or training combined with enhanced representational guidance might lead to some limited learning that would appear as improved domain performance following exposure to the platform, which leads to our third hypothesis:

H3. after using Pendo subjects will produce somewhat better predictions in the same task even when the tool is not available to them.

4.2. Domain task and measures

All subjects performed a housing market prediction task. Housing market forecasts are very relevant for economic analysis because of their impact on the overall economy and on the market forecasts are very relevant for economic analysis because of

The housing market is more complex than some other markets (e.g. consumer goods) for a variety of reasons: houses are both investments and consumer goods; transactions are characterized by information asymmetry and high costs, and this favors the presence of a network of intermediaries and facilitators (real estate agencies, banks, insurance companies, etc.); housing markets are subject to speculative maneuvering and are quite sensitive to the variations of macroeconomic forces such as interest rates and inflation. Because of this complexity, housing prices are characterized by a high degree of uncertainty and market experts are required to interpret forecasts produced by quantitative models.

We developed a set of twenty prediction problems about the US housing market in the years from 1986 to 2006. Each of these problems presented subjects with the question: “How do you think the housing prices index will change in year t + 1 given the following data for year t?” and a set of four or five indicators, such as:

- Number of households is increased 2.5% this year
- Unemployment has decreased 1%
- Interest rate on a ten year fixed rate mortgage is on average about 10% and is slowly increasing
- The stock market has been growing at a moderate pace

Subjects were asked to respond with one of three options drawn from the following answers: i) the index will be stable (+5%); ii) the index will increase (>5%); and iii) the index will decrease (<−5%).

The 5% threshold was selected based on the average rate of appreciation of residential real estate in the US market in a period of regular but sustained growth (1984–1998, 4.5%). A 5% increase usually compensates for inflation-adjusted expected wage growth and in a normal economic scenario any rise beyond 5% could be safely considered as a significant increment.

We developed indicators for the problems as follows. First, we collected data on about 20 indicators using data drawn from a number of official sources including the US Census Bureau, Federal Reserve, Dow Jones and the Case Shiller indices as well as websites and blogs run by independent experts, consultancy companies and professional associations [27,33,42,54]. These data were classified into higher order variables that are typically deemed to influence market equilibrium in the housing market, such as demographics, economic growth, access to credit, consumer preferences and tastes, substitute goods, construction costs, existing residential stock and sellers & customer expectations. Finally higher order variables were used to generate the indicators provided with each problem. The complete list of indicators and higher order variables are listed in Appendix A.
An individual’s prediction performance was measured in terms of his or her ability to make the right prediction on the price trend in the short term (one year). Given some data on the current year, subjects were required to select one of the three options as stated above. The procedure was replicated on five different years chosen at random in the considered time window. For each year, we compared subjects’ guesses with the actual price trend as derived from historical data; we computed the percentage of correct answers over the five years and used this value as the dependent variable.

4.3. Experiment design

We designed an experiment to evaluate our three hypotheses. The independent variable in the experiment is whether or not the subjects had access to the Pendo platform, and the dependent variable was the subjects’ ability to forecast market trends, as described above. We used a crossover experimental design to examine the performance of subjects with and without access to Pendo. A crossover experiment is a longitudinal study in which subjects receive a sequence of different treatments in different orders [3]. Crossover experiments are very useful with smaller numbers of subjects because repeated measurements of the same subject are expected to yield more precise results than comparisons across subjects.

In our study, each subject was exposed to two treatments in different time periods, as shown in Table 1. All subjects solved the same problems, but in the Pendo condition were required to use Pendo to represent their arguments and evidence they felt was relevant, while in the NoPendo case they solved the problem using only paper and pencil. Subjects in the Pendo condition were allowed to indicate the answer they thought was correct regardless of the answer recommended by Pendo.

We tested our primary hypothesis (H1), that using Pendo would help people perform better in the domain task, by comparing Pendo vs. NoPendo prediction performance. We tested hypothesis (H2), that decision structures developed using the argument Pendo could be used to solve other domain problems, by evaluating the performance of maps developed in the Pendo condition upon the problems in the NoPendo condition.

In addition to offering increased statistical power, the crossover research design also allowed us to investigate our hypothesis (H3) that some learning might occur as a result of platform use. If some learning occurred, we would expect the magnitude of the performance change with group A (exposed to the Pendo first) to be smaller than that with group B (Table 1). Within the language of crossover studies this is called a carryover effect, which occurs when a treatment condition has an impact on task performance beyond the end of exposure to the treatment.

Note that the carryover effect can be difficult in cases to distinguish from the time effect, which occurs when external time-varying environmental factors influence task performance (e.g. learning due to external exposure or over the course of the experimental session). To help reduce the possibility of time effects, each group worked through both conditions in the same lab session, thus eliminating the possibility of external exposure to the domain. We also included a training period with sample problems drawn from our dataset, reducing the likelihood that performance would generally improve as people became accustomed to the domain during data collection.

5. Participants

Participants were recruited from among undergraduate students attending a course in Economics in the second year of a program in Industrial Engineering. 76 students participated in the experiment (34 in group A and 42 in group B; the groups differed in size because some students did not attend the experiment session). Participants were evenly split by gender (54% male), and were randomly assigned to the two groups. By the time of the study, students taking the course had acquired knowledge about the theory of market equilibrium and received one lecture about the housing market. Participation in the experiment and the preliminary training activities was voluntary and rewarded with extra credit.

Because participants were neither from the US nor were experts in real estate market analysis, and the data used to construct experimental problems preceded the mortgage crisis by several years, we did not expect that students would have prior knowledge of our data set. As a further check, the instructor probed the students on this topic during the class activities and confirmed that students were ignorant about the price trend during this period.

5.1. Procedure

Prior to the experiment, participants received a general introduction to argument mapping and some hands-on experience with Pendo. The total time spent in these training activities was 2 h. The general introduction included information on the theory of argument and the structure and uses of argument maps. For hands-on training, they were given a basic tour through the interface. Some time was spent to illustrate how the belief aggregation function worked without going into technical details about the algorithm. Finally, participants were given a question similar to the one used in the experiment about a different market, and given an opportunity to play with Pendo on their own. Training concluded one week before the experiment began.

The experiment itself took place in a single two-hour lab session. The activities were organized in three steps as reported in Table 2. In the warm up phase, participants solved five sample problems drawn from the same data set to familiarize them with the rubric. Participants did not have access to Pendo during this phase, and were given 3 min to solve each problem. Performance on the warm-up problems is not included in the following analysis.

Ten problems were randomly drawn from the fifteen remaining problems, and all participants solved these problems in the same order. Problems were presented one at a time, and participants were given 7 min to solve each problem. Because the conditions were presented to the groups in different orders (see Table 2), each problem was solved in both conditions, allowing us to screen for the possibility that one group of five problems was harder than the other.

In the NoPendo condition, participants were allowed to solve the problem however they saw fit, using a pen and paper for notes as necessary. In the Pendo condition, they were asked to read the problem first, then to update an argument map using Pendo containing their forecasting rationale and post evidence in a new session to determine the most likely outcome. Participants were explicitly told to use whatever knowledge was available to them, even if it was not an indicator provided as part of the problem. All participants recorded their final answer on a sheet of paper, and participants in the Pendo condition were free to use the prediction made by the system or not.

Table 1
Schema for a two-period, two-treatment crossover design.

<table>
<thead>
<tr>
<th>Step</th>
<th>Group A</th>
<th>Group B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Warm-up</td>
<td>5 problems, no software</td>
<td>5 problems, no software</td>
</tr>
<tr>
<td>I</td>
<td>Problems 1–5 (Pendo)</td>
<td>Problems 1–5 (NoPendo)</td>
</tr>
<tr>
<td>II</td>
<td>Problems 6–10 (Pendo)</td>
<td>Problems 6–10 (NoPendo)</td>
</tr>
</tbody>
</table>

Table 2
Experimental procedure design.

<table>
<thead>
<tr>
<th>Step</th>
<th>Group A</th>
<th>Group B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Warm-up</td>
<td>5 problems, no software</td>
<td>5 problems, no software</td>
</tr>
<tr>
<td>I</td>
<td>Problems 1–5 (Pendo)</td>
<td>Problems 1–5 (NoPendo)</td>
</tr>
<tr>
<td>II</td>
<td>Problems 6–10 (Pendo)</td>
<td>Problems 6–10 (Pendo)</td>
</tr>
</tbody>
</table>

Please cite this article as: J. Introne, L. Iandoli, Improving decision-making performance through argumentation: An argument-based decision support system to compute with evidence, Decision Support Systems (2014), http://dx.doi.org/10.1016/j.dss.2014.04.005
Because one objective was to determine if the maps generated during the study were general representations of the subjects’ domain knowledge, we sought to have participants create maps that were sufficiently general representations of the domain. Consequently, we adopted the workflow described at the end of Section 3—participants in the Pendo condition were initially presented with a map that included just the three possible forecasts, and then asked to incrementally modify the same map as they were exposed in each problem. Although participants were not explicitly told to create “general” maps, we expected that their maps would become more general as subjects progressed through their series of problems.

### 6. Results

Of the 76 participants, fifteen were unable to complete the study or encountered technical errors that prevented data from being collected; these data were omitted from the analysis. The remaining 61 participants each generated an argument map, and data for these participants is reported. Descriptive statistics for results from the Pendo condition are provided in Table 3.

On average, participants posted three or four pieces of evidence per session, and usually recorded an answer that matched the system’s prediction. Note that in the context of the performance differences reported below, the agreement between the system’s output and user response serves as a manipulation check because it indicates that performance differences were most likely due to the subject’s use of the system’s output. No significant differences for the measurements reported in Table 3 were found between the two groups.

Table 4 provides performance statistics for the two groups. Note that here and throughout the remainder of this discussion, scores are reported as the ratio of correct answers to the number possible correct answers (n = 5). To develop support for research hypotheses H1 (the system’s impact on performance) and H3 (whether or not learning occurred), we screened the data to determine whether any time or carryover effects could be detected, which would appear as either a main effect for period or a significant interaction between period and treatment. A two-way ANOVA revealed that there was no detectable effect either for period alone, (F(1,118) = .048), or for the interaction between period and treatment, (F(1,118) = .004), leading us to rule out effects due to time or learning. However, there was a highly significant main effect due to treatment (F(1,118) = 10.60, p = .001). Thus, the use of Pendo led to a significant mean improvement in performance on the domain task (H3 is confirmed), but we detected no discernable learning as a result of tool use (H3 is rejected).

Scores were in general quite low across the population. A two-tailed t-test demonstrates that mean participant performance in the NoPendo condition was not significantly better than random at this sample size (r(304) = 1.46; p = .15). However, with access to Pendo, participants chose the correct response about half of the time, and the difference between this level of performance and that of a random process is highly significant (r(304) = 5.4143; p < .001).

Despite the improvement in the population average, we observed a striking lack of correlation between the performance of individuals between conditions (r(61) = −.06, n.s.), which suggested that the impact of Pendo was not a simple amplification of individual ability. To gain insight into this phenomenon, we examined how changes in score were distributed across the population. Fig. 3 presents a histogram of the change in score from the Pendo to the NoPendo condition (a positive number indicates that an individual did better with access to Pendo). It is apparent from this analysis that just about half of the participants experienced no improvement or actually performed worse in the Pendo condition. We then compared each participant’s score in the NoPendo condition to the change in their score when moving to the map condition (Pendo score − NoPendo score). We found a highly significant inverse correlation (R(61) = −.68, p < .0001) between individual scores and performance change, and a linear model predicts 47% of the variance in the data (see Fig. 4). One interpretation of this result is that Pendo “pushed” the performance of each individual participant towards a mean independent of his or her performance without the tool.

As noted previously, participants were not obligated to follow the recommendations of the map, even though the majority did. Although we found that there were no significant differences between the mean accuracy of predictions generated by the system in the Pendo condition and those reported by the participants, restricting our analysis to just those users whose answers deviated from the system’s predictions sheds some additional light on user performance. Any participant choosing not to follow the system’s recommendation could, for any given problem, choose a correct or incorrect answer. Using this distinction, we classified subjects into three groups: (a) those who always followed the system’s prediction; (b) those who correctly ignored the system at least once; and (c), those who incorrectly ignored the system at least once; note that (b) and (c) are non-exclusive. 23 subjects ignored the system recommendation at least once, and three of these people fit into both categories (b) and (c). Thus, the majority of participants who chose to ignore the Pendo’s output at some point either did so correctly or incorrectly, but not both. Table 5 presents data comparing the change in mean performance of these three groups across conditions. Those users who were able to correctly ignore the map at certain points were able to improve their scores in the Pendo condition much more than the other users; each of these users did as well or better in the Pendo condition than they did in the NoPendo condition. It is notable that these subjects were also among the best performers in the NoPendo condition, though the difference is not significant with this small a sample.

Users who always followed the system’s guidance also had higher average scores than they did in the NoPendo condition, but the change was not as great. Finally, the average score of users who incorrectly ignored the system did not increase at all in the Pendo condition. Although the sizes of the populations at this level of analysis are too small to draw definitive conclusions, this data suggests that Pendo can improve scores for low-performing individuals, but an individual’s ability to reflect upon the system’s guidance and use it wisely is also important.

The preceding results demonstrate that the argument maps created by subjects using Pendo performed better than humans on average. This is sufficient to demonstrate that the system functioned effectively as a decision support tool, but it does not indicate if the knowledge encoded by the system was sufficiently general to solve unseen problems (hypothesis H2).

To examine this question, we used the maps generated by participants in the Pendo condition to process the evidence provided in the NoPendo condition. In order to do this, it was first necessary to...
determine how this evidence should be applied to the arguments in each map. This was feasible within our dataset because all of the evidence was developed on the basis of a finite set of indicators, and subjects almost always defined argument nodes in terms of these indicators.

To establish a mapping, we coded each of the nodes in the collection of argument maps using the indicators that were originally developed to create the domain problems. To establish inter-rater reliability, the first author and one independent coder coded five randomly sampled maps; there was a 100% agreement between the coders. We found this level of agreement to be sufficient to establish that the coding process was indeed straightforward, and the first author coded the remaining maps. Once all of the maps were coded, evidence from each of the problems in the NoPendo condition was posted to the participants’ maps, answers recorded, and performance compared to the subjects.

The mean performance scores for the participants’ maps on the NoPendo problems are reported in Table 6. Participant scores are included for comparison. For both groups, the maps out-performed their authors on the problems in the NoPendo condition. We ran a two-way ANOVA with Group (A or B) as one factor, and source of the performance data (human or Pendo) as the other factor, and found a significant main effect for the source of the performance data ($F(1,118) = 6.12, p = .015$). No significant effects were found for either the other factor or the interaction between the factors. Combining the data from the two groups, we find that, when using maps on held out data, Pendo performs significantly better than random ($t(304) = 5.4461; p < .001$).

These results suggest that the system can function effectively as a means for extracting general domain knowledge from its user, and that the maps people develop can in cases apply this knowledge to perform better than their human authors (H2 is confirmed).

The performance of the maps on problems from the NoPendo condition also helps elaborate our understanding of how the technology improves performance. As with data collected in the Pendo and NoPendo conditions, there was no correlation between the performance of individuals and the performance of their maps on the same data ($R(61) = .05$), but there was a strong inverse correlation between the difference in scores (Pendo score − Human score) and the human score in the NoPendo condition (see Fig. 5; $R(59) = −.71; p < .001$). Once again, a linear model explains about 40% of the variance in the data ($R^2 = .42$).

Fig. 6 suggests that Pendo’s performance is normally distributed around a roughly consistent mean across each level of the human author’s performance. Together with the analysis of individual performance described above, the data suggests that independent processes govern the performance of the unassisted users and the performance of their maps. We conclude that people’s ability to perform well in the domain task is not correlated with their ability to create argument maps that perform well.

6.1. Summary

The preceding results show that the use of Pendo leads to a 36% improvement in the average performance of the population studied in the housing market forecasting task. This is a substantial and significant improvement, and could translate into monetary gains if applied to housing market speculation. If confirmed on a larger scale, these results would present a clear and attractive value proposition for anyone considering the adoption of similar technology as a decision support aid.

Our analysis also shows that Pendo does not simply amplify the decision-making ability of individuals, though. Instead, the system appears to replace a human decision process with a technologically embodied one that performs at a higher level on average, but there is no simple correlation between the two processes. Effective use of the

Fig. 3. Count of participants at various levels of improvement in the map condition.

Fig. 4. Plot of mean subject scores in the NoPendo condition (x-axis) vs. mean score improvement in the Pendo condition (y-axis).

Table 5

Summary of participant scores grouped by their choices to follow Pendo’s guidance (* indicates $p < .05$, and ** indicates $p < .01$).

<table>
<thead>
<tr>
<th># of participants</th>
<th>NoPendo mean score</th>
<th>Pendo mean score</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always followed</td>
<td>38</td>
<td>.36</td>
<td>.48</td>
</tr>
<tr>
<td>Incorrectly ignored at least once</td>
<td>15</td>
<td>.4</td>
<td>.4</td>
</tr>
<tr>
<td>Correctly ignored at least once</td>
<td>11</td>
<td>.42</td>
<td>.7</td>
</tr>
</tbody>
</table>

Please cite this article as: J. Introne, L. Iandoli, Improving decision-making performance through argumentation: An argument-based decision support system to compute with evidence, Decision Support Systems (2014), http://dx.doi.org/10.1016/j.dss.2014.04.005
system is not correlated with unassisted individuals’ performance in the domain task—the system appears to push both high-performing and low-performing unassisted individuals towards a common mean when they use the system. Finally, although there were too few subjects to draw concrete conclusions, the best performers seem to be those who are able to reflect on Pendo’s output, and ignore it when it is wrong. Our data does suggest that this ability may be correlated with an individual’s unaided domain performance, but further investigation is required to confirm this finding.

In addition to serving as a decision aid, Pendo may also be an effective, low-cost tool for the storage and reuse of domain knowledge. With very little training and no explicit instructions regarding the generality of the maps they constructed, undergraduate students were able to build maps that performed as well on unseen problems as they did on a training set. Thus, Pendo can function as an expert system that is trained in the course of being applied in a decision support context.

7. Discussion

One goal of this article has been to make a stronger case for ABS in decision support. We interpret empirical work by Kuhn [20] and others to suggest that there is a value in a tool that people can use to encode and reflect upon their rationale for a decision, and then use to evaluate the support for competing claims. Many existing approaches to decision support seek to do this (e.g. Bayesian Networks, Decision Trees, the Analytical Hierarchy Process), but their interface representations are based on abstract mathematical concepts that end-users may have trouble using fluidly.

Argument-based systems, on the other hand, are inspired by human approaches to communicating about the rationale for competing claims, so people are accustomed to organizing information in this manner. Unfortunately, most research on argument-based systems has focused on the proper use of formal argument representations and the cognitive benefits of ABSs, rather than on actual decision-making performance. We have shown that an argument platform that computes support for different claims and helps users distinguish between evidence and representation improves decision-making in a task that is complex and realistic. We intentionally bypassed considerations about the use of the formalism or the user’s decision processes and found that, with minimal training, people were able to use Pendo to improve their housing market forecasting ability. In the process of being used to solve a series of problems from the same domain, the system accumulated enough general domain knowledge to perform equally well on novel problems. Together, these outcomes satisfy our initial goal of providing a concrete case and some systematic evidence that make ABSs more attractive in decision support contexts.

Our findings also reveal a more nuanced story about the application of ABSs for decision support. Pendo appeared to replace the user’s decision procedure instead of improving it. We hypothesize that the unaided individual uses a decision process that is substantially different from the one embodied by the system, and because of this distance, subjects encode decision information for the platform to process and delegate analysis and decision-making. This has two direct consequences. First, the performance of the user with access to the system depends on an individual’s ability to convert relevant information into the system’s representation, rather than to make good decisions. For the particular constraints of our experiment, it appeared that this ability was uncorrelated with unaided decision-making ability.

This would help explain why we detected no learning in our study and why so much training with argumentation platforms is required if the goal is to help people think more rationally. Though it appears to be easy enough to use, the system’s representation (and perhaps argumentation in general) seems to be quite unlike natural cognitive decision structures. We might conclude that as a cognitive scaffold, an argument representation is outside of the typical individual’s zone of proximal development [50], and so it is not possible for the user to internalize the representation without significant instruction.

A further consequence is that we would expect that both high performing and low performing unaaided decision makers would be brought to the mean performance level of the system itself. Therefore, a decision maker who generally outperforms the algorithmic engine would be more likely to experience a performance impairment when using the system. Similarly, underperforming decision makers would be more likely to experience an improvement. This expectation is borne out in our findings.

This does not reduce the impact of our contributions; millions of non-experts make decisions daily to invest in the purchase of a new home, and software such as that presented here could be used to help. Yet our findings indicate that the deployment of decision support technologies should be done thoughtfully and with consideration for the decision-making expertise of the user. This is an important caution that should be examined in other domains, and will be an important variable to consider in future studies.

8. Conclusion

Our results illustrate how ABS might be used to improve decision-making, with very little upfront cost and further potential benefits as the system begins to accumulate domain knowledge. ABS can provide an interface through which people are readily able to encode reusable decision-making knowledge that can be processed by a rational decision engine. Additional research is required to understand how these results generalize, but they present an easy to understand and attractive value-proposition for ABS as a decision-support technology.

A limitation of our study is that we were unable to determine which of Pendo’s features are responsible for the performance improvement. This is an important issue to consider if we seek to further understand the platform’s role in mediating a user’s decision process. Our findings point to several possible research questions about the design and use of the interface. For example, how much does the separation between
Appendix A. List of indicators provided to subjects as problem data

<table>
<thead>
<tr>
<th></th>
<th>PENDO 0</th>
<th>PENDO 0.2</th>
<th>PENDO 0.4</th>
<th>PENDO 0.6</th>
<th>PENDO 0.8</th>
<th>Total</th>
<th>Mean Map Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0.5</td>
<td>1.1</td>
<td>1.6</td>
<td>2</td>
<td>2.4</td>
<td>.44</td>
</tr>
<tr>
<td>New houses</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>61</td>
<td>.4</td>
</tr>
<tr>
<td>Price to rent</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>foreclosure</td>
<td>0.8</td>
<td>0.4</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.4</td>
</tr>
<tr>
<td>Mean Human Score</td>
<td>.4</td>
<td>.33</td>
<td>.37</td>
<td>.39</td>
<td>.35</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 6. Performance of subjects and the maps they create on the problems in the NoPendo condition. Cells indicate the number of participants at each level of performance.

References


Evidence and representation in the interface impact performance? Do untrained users use the argument formalism properly, and how does proper use correlate with outcome? We hope that future studies will focus more directly upon such questions.

Within the context of ABSs research, our results suggest that novice users can easily use an argument representation to encode information that can drive an automated belief aggregation algorithm, but that this does not lead them to change the way they think about a decision problem. As a consequence, decision performance (in problems where there is a correct outcome) depends upon the ability of a user to use the tool and the performance of the belief aggregation algorithm.

Within the larger context of decision support, we submit that our findings raise questions about other tools. Providing a decision maker with a representation intended to scaffold a more rational thought process does not necessarily amplify that individual’s existing decision-making ability. If the system embeds a high-performing algorithm, this might not be a problem for a majority of users. However, there may still be cases in which an expert human outperforms an algorithm, and this raises several research questions: Can we determine, either in advance or at runtime, if a system will replace or if it will amplify a user’s decision process? What sorts of design features lead to these two kinds of interaction? Which of these strategies for mediating human decision processes are best suited to different situations? We believe that these are exciting and fruitful avenues for future inquiry.

Appendix A. List of indicators provided to subjects as problem data

All data refer to the time window 1986–2006.

- Number of households (millions)
- Net immigration rate (%)
- Income (thousands $)
- Savings (as % of income)
- Unemployment
- Inflation
- GDP growth
- Mortgage interest rate (national average of 30 years, fixed rate mortgage)
- Released asset-backed securities (as % of total mortgages)
- Subprime origination (billions of $)
- Mortgage equity withdrawal (% of disposable income)
- Interest only and negative amortization rate approved mortgages (as % of total mortgages)
- National ownership rate (% of households)
- Dow Jones
- Price to rent ration (price/monthly rental)
- Construction costs
- Foreclosures (as % of total sales)
- New houses (units completed in thousands)
- Builder confidence survey

Please cite this article as: J. Introne, L. Iandoli, Improving decision-making performance through argumentation: An argument-based decision support system to compute with evidence, Decision Support Systems (2014), http://dx.doi.org/10.1016/j.dss.2014.04.005