

A modular approach to the evaluation of Adaptive User Interfaces

A. Paramythis¹, A. Totter¹ and C. Stephanidis^{1,2}

¹ Institute of Computer Science, Foundation for Research and Technology – Hellas, Science and Technology Park of Crete, GR-71110 Heraklion, Crete, Greece

² Department of Computer Science, University of Crete, Greece
{alpar, totter, cs}@ics.forth.gr

Abstract. Adaptive User Interfaces (AUIs) are a continuously growing area of research, with numerous experimental and commercial systems reported in the literature, in diverse application areas. Although there exist today a number of established approaches and frameworks for the design and implementation of AUIs, their evaluation is yet to be addressed in a comprehensive way. Arguably, the main drawback of existing evaluation methods and techniques is that they fail to provide adequate feedback into the AUI design process, or to generate empirical findings that can be reused across application boundaries. This paper proposes a new, modular approach to the evaluation of AUIs, which is specifically intended to cater for the aforementioned problems. The adaptive Nautilus Web browser, on which the approach will be applied is also briefly presented.

1. Introduction

The application of adaptive methods and techniques in human-computer interaction (HCI) is gaining increasing attention in recent years. However, to date, there is limited knowledge as to which adaptation methods and techniques are appropriate for different users and for different interaction contexts. This is arguably due to the lack of reusable empirical findings coming from the evaluation of adaptive interactive systems, which, in turn, can be traced back to the way in which the evaluation of adaptive user interfaces is approached today.

To start with, adaptation is not sufficiently addressed by existing standardized evaluation frameworks (although, in some cases, it is a concern) (Stary & Totter, 1997). As a result, researchers have had to employ more “basic” evaluation tools to approach the assessment of adaptive systems. Among the best-known and widely used approaches in the field is the “with and without” adaptivity evaluation design, in which an adaptive instance of the system is compared with a non-adaptive one (see, e.g., (Kaplan, Fenwick, & Chen, 1993), (Meyer, 1994), (Boyle & Encarnacion, 1994), (Weber & Specht, 1997), (Brusilovsky & Pesin, 1998), (Brusilovsky & Eklund, 1998)). A major criticism of this evaluation approach has been that the non-adaptive instance cannot be “optimal” in any way, if adaptation is properly “designed into” the system (Höök, 2000). Another equally important problem is that, in this type of study,

the reasons behind the “success”, or “failure” of adaptation can only be traced back to the initial hypotheses of the adaptive system design. In other words, it is not possible to ascertain why, and under what conditions, a particular type of adaptation may be employed towards a specific goal. This situation is exemplified in the several studies that have addressed adaptive link annotation, often arriving at contradictory conclusions (Eklund & Brusilovsky, 1998).

A different perspective on the study of adaptive systems has been put forward by Oppermann (Oppermann, 1994), in the assessment of adaptation in Flexcel II (Krogsäter, Oppermann, & Thomas, 1994). Following this perspective, adaptation is treated as an integral part of the system and evaluation is not based on the presence of a non-adaptive counterpart. However, this approach is also limited with respect to assessing the degree to which the different factors influencing adaptation contribute to its “success” or “failure”.

In light of the above, there is an acknowledged need for a renewed look at the evaluation of adaptation, in which the employment of traditional HCI evaluation methods and techniques will be placed on a new basis, acknowledging the particular characteristics that differentiate adaptive systems from their “static” counterparts (Höök & Svensson, 1999), (Höök, 2000).

The main idea behind the approach put forward in this paper is that the evaluation of adaptive systems should not treat adaptation as a “monolithic” / singular process happening behind the scenes; rather, adaptation should be “broken down” into its constituents, and each of these constituents should be evaluated separately where necessary and feasible. The seeds of this idea can be traced back to (Totterdell & Boyle, 1990a), who propose that a number of adaptation metrics be related to different components of a logical model of adaptive user interfaces, to provide what amounts to adaptation-oriented design feedback. Furthermore, (Totterdell et al., 1990a) present two types of assessment performed to validate what is termed “success of the user model” (note that, in their case, the “user model” is also responsible for adaptation decision making): “... an assessment of the accuracy of the model's inferences about user difficulties; and an assessment of the effectiveness of the changes made at the interface.” (Totterdell et al., 1990a)

The contribution of this paper, along these lines, is the introduction of a modular approach, which offers a detailed view into the “decomposability” of adaptation, from the perspective of HCI-oriented evaluation. The main strength of this approach, which builds extensively on previous work in the field, lies with the potential it offers towards deriving detailed evaluation results that can be analyzed, extended and reused across user interfaces and application domains.

A related approach to the one presented in this paper can be found in (Weibelzahl & Lauer, 2001), where the authors introduce an evaluation framework for the assessment of interactive systems that employ Case-Based Reasoning techniques to support adaptation in their interaction with the user. Their framework bears many similarities to the approach postulated herein, especially in terms of how adaptation is “decomposed” and evaluated in a series of steps.

Although the proposed approach is presented in relation to a particular class of adaptive systems, namely adaptive user interfaces (AUIs), it should be noted that it is not exclusively relevant to AUIs; rather, the proposed evaluation approach is expected to be easily extensible to other classes / categories of adaptive systems.

The rest of this paper is structured as follows. The following section presents the proposed evaluation approach in two steps: in the first step, a high-level model for adaptation in AUIs is introduced, accompanied by a tentative classification of evaluation methods, intended to facilitate subsequent discussions; in a second step, the model is broken into (sometimes overlapping) modules and the evaluation of each module is discussed in detail. The subsequent and final section presents a brief overview of the Nautilus Web browser, which will serve as the platform for applying and further improving the proposed approach.

3. Modular evaluation of AUIs

The proposed approach is based on the premise that the evaluation of individual stages (referred to as “modules”) involved in the AUI adaptation cycles, enables the derivation of detailed findings, which, in turn, provide ample feedback back into the AUI design process. Specifically, the proposed approach:

- identifies “modules” of AUIs that can, and should, be evaluated both separately and in combination (i.e., the evaluation objects);
- presents the evaluation rationale underlying the decomposition of AUIs into modules and the subsequent assessment of these modules, based on specific criteria (i.e., the evaluation purpose);
- circumscribes the methods and techniques that can be employed for the evaluation of the different “modules”, in the different stages of the AUI development life-cycle (i.e., the evaluation process).

To that effect, the rest of this section will: (i) establish a basis for the discussion of evaluation methods / techniques; (ii) present a high-level model for adaptation in AUIs; identify the individual stages of adaptation that can be targeted as evaluation modules; and, (iii) propose specific evaluation methods and techniques that can be employed for each module.

3.1 A contextual perspective on evaluation

To facilitate a rather generalized treatment of user-based evaluation methods in the forthcoming sections, a tentative classification will be introduced. This classification is not intended to fully capture the characteristics of all existing evaluation methods, but rather to identify those dimensions of evaluation that are pertinent to the ongoing discussion. The classification scheme is based on two dimensions: (a) the types of evaluation measures that are supported by each method, and (b) the stage of the development life cycle that each method is best suited for.

The first dimension, i.e., evaluation measures (one could alternatively term this dimension “data collection methods”), is a simplification of the measures proposed by McGrath (McGrath, 1995), which are extensively used in the social and behavioral sciences (see Table 1, top row). Along this dimension, evaluation methods are separated into: *self reports* of participants (e.g. questionnaire responses, interview protocols, rating scales, etc.); *observations*; and, *trace measures*.

Regarding the second dimension of our classification, i.e., stage of the development life cycle that each method is best suited for, a broad categorization is employed, which distinguishes between methods that: (a) are best suited for the early (exploratory) stages of design, (b) require the existence of at least an interactive prototype, (c) are targeted towards complete (“finished”) products, and (d) can be used (in variations) at any stage of the design process.

Table 1 presents a classification of thirteen empirical evaluation methods commonly used for the investigation of usability in HCI, as identified in (Jordan, 1998), using the above classification scheme.

Table 1. Classification of empirical usability evaluation methods.

| <i>Empirical Usability Evaluation Methods</i> | <i>Types of Measures</i> | | | <i>Suitability for employment at different development stages</i> |
|---|--------------------------|--------------|----------------|---|
| | self reports | observations | trace measures | |
| Focus groups | ■ | | | at any stage of the design process |
| Interviews | ■ | | | |
| Questionnaires | ■ | | | |
| Private camera conversation | ■ | | | better suited for early design stages |
| Valuation methods | ■ | | | |
| User workshops | ■ | | | require at least an interactive prototype |
| Co-discovery | ■ | □* | | |
| Think aloud protocols | ■ | □* | | |
| Logging use | | | ■ | |
| Controlled experiments | | ■ | | better suited for “finished” products |
| Incident diaries | ■ | | | |
| Feature checklist | ■ | | | |
| Field observation | | ■ | | |

* When applied in HCI, Co-discovery and Think aloud protocols need to be combined with some form of observation in order to obtain a meaningful record of the interaction circumstances, so as to enable the contextual interpretation of the users' comments.

In addition to the above user-based evaluation methods, the proposed AUI evaluation approach will also consider expert-based ones (i.e., evaluation methods that require the participation of experts, but not of end users). Following the broad categorization of (Jordan, 1998), such methods can be classified as *expert appraisals* (which typically require the expert to judge the product against known principles, guidelines, rules, standards, etc.) and *cognitive walkthroughs* (which call upon the expert to approach the evaluation from the point of view of a typical user performing a particular task).

Finally, it should be mentioned that, following the norm in HCI, user testing in a “usability laboratory” (for hypothesis testing, or performance measurements such as error rate, task completion time, task frequency, etc.) is classified under controlled experiments (although controlled experiments are not restricted to this type of user testing).

3.2 A high-level model of adaptation in AUIs

In the context of the proposed approach, a base model for adaptation is required, which will reveal some important high-level architectural components of AUIs, as well as explicitly represent the fundamental stages involved in deciding upon and effecting adaptation in HCI. The model presented in Figure 1 is based on, and extends, the logical two-level architecture of adaptation in (Totterdell et al., 1990b). Our goals in deriving this model have been to: (a) make the individual stages of adaptation as concrete as possible, without, however, delving into technical issues, or implementation-oriented details, and (b) introduce details related to different approaches to AUI adaptation, which impact on the evaluation choices (affecting both the *objects* of evaluation, as well as the *process* for evaluating them). A number of points that should be noted regarding the model are: no assumptions are made as to the employed technologies and the targeted platforms; no assumptions are made as to the physical distribution of user interface components (e.g., over the network); although depicted separately at the conceptual level, some of the components may actually be combined in an implemented AUI.

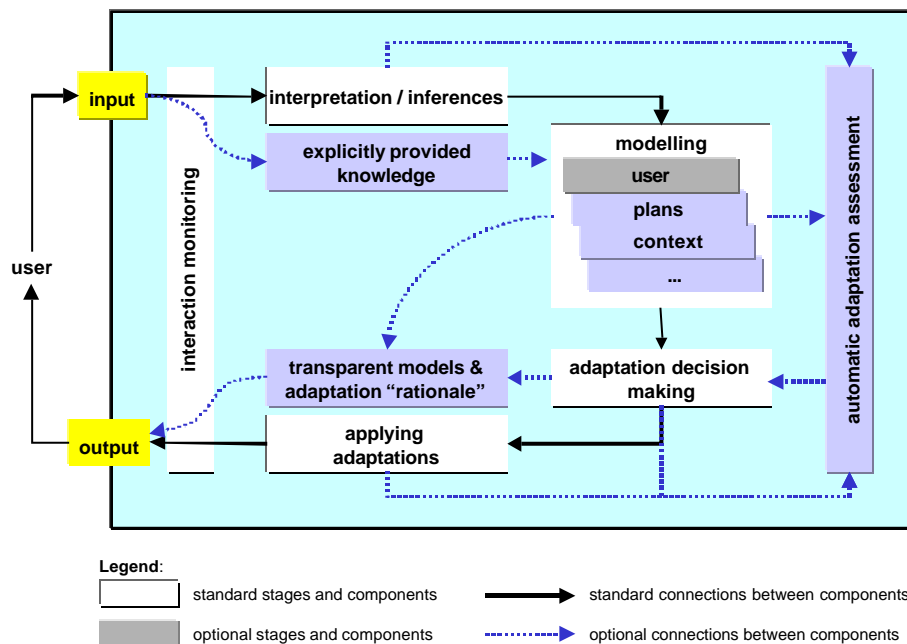


Fig. 1. High-level model of adaptation in AUIs.

- The model encompasses the following components / stages of adaptation:
- *Interaction monitoring*: Refers to facilities that are intended to capture the exchanges between the user and the user interface, at different levels of the interaction (i.e., physical, syntactic, semantic (Hoppe, Tauber, & Ziegler, 1986)).

- *Interpretation / inferences*: Refers to the part(s) of the AUI that is responsible for interpreting information made available through interaction monitoring, in order to update the models maintained by the system (e.g., user model).
- *Explicitly provided knowledge*: Refers to information about the users' characteristics, plans, tasks, context, etc., which is explicitly provided to the system (as opposed to indirectly inferred from interaction data), typically by users themselves.
- *Modeling*: Refers to explicit or implicit representations of the users (including, for example, their abilities, skills, requirements, preferences), their plans with respect to a particular (portion of an) interactive session, the tasks that can be performed with the system, etc. Of particular interest in the context of the present discussion are those models that are dynamically updated during interaction, based on knowledge acquired at run-time (the user model being a typical such case).
- *Adaptation decision making*: Refers to the part (or parts) of the AUI that is responsible for deciding upon the necessity of, as well as the required type of, adaptations, given a particular interaction state. Seen at an abstract level, decisions made at this stage match information found in the various models maintained by the AUI, with the alternative interactions designed to cater for variations therein.
- *Applying adaptations*: Refers to the actual introduction of adaptations in the user-system interaction, on the basis of the related decisions. Although typically subsumed by adaptation decision making in the literature, this adaptation component may be varied independently of the decision making process, e.g., to account for different adaptation strategies.
- *Transparent models & adaptation "rationale"*: Refers to the particular case of AUIs that enable users to review the models maintained by the AUI (at different levels of "transparency" – see (Höök et al., 1996) for a detailed discussion), or the rationale that underlies the adaptation decisions made by the system. In the case of transparent modeling, users may also be offered the capability to modify these models, so that the latter better reflect their individual or other characteristics.
- *Automatic adaptation assessment*: Refers to the run-time assessment of the effects of decided upon and effected adaptations, with the intent of evaluating their "success" (i.e., whether the goals underlying their introduction have been met). This stage is referred to as "second-level adaptation" in (Totterdell et al., 1990b) and may further involve the modification of aspects of the lower-level adaptation cycle (e.g., by enabling or disabling rules in rule-based adaptation, or by altering the "weight" of alternatives, in decision theory-based adaptation).

It should be noted that this high-level model is not intended to capture the characteristics of all AUIs reported in the literature. On the other hand, there do not exist to date AUIs that comprise all of the identified components. However, the modular nature of the proposed evaluation approach allows one to selectively apply it, or extend it to suit the particular needs of the AUI at hand.

3.3 Modular evaluation

In this section we will identify adaptation "modules" (comprising one or more of the adaptation stages / components in the previous section), which can be evaluated individually and in combinations. Before proceeding to the presentation of the modules

and their evaluation, we would like to make the following clarifications, which hold true throughout the presentation of the approach:

- In some cases, evaluation methods that do not involve the users directly assume that the evaluator / expert takes into account the characteristics (abilities, skills, knowledge, etc.) of the “typical” user of a system. Since the concept of a “typical” user is contrary to the very notion of AUIs, this assumption cannot be applied in AUI evaluation. Thus, an explicit requirement that permeates the proposed approach is that, in all cases where users are not directly involved in the evaluation, each and every individual evaluation task takes into account a particular user (conveyed through relevant characteristics, which are encoded in some type of user profile), in a particular context of use (conveyed in a way analogous to the user).
- Expert-based evaluations in HCI are, in general, assumed to be conducted by usability experts. In the description of the proposed evaluation approach we will occasionally refer to expert-based evaluation tasks which are foreseen to be undertaken by individuals that possess expertise relevant to the application domain, the target user group(s), etc., but do not necessarily have a background in usability evaluation.

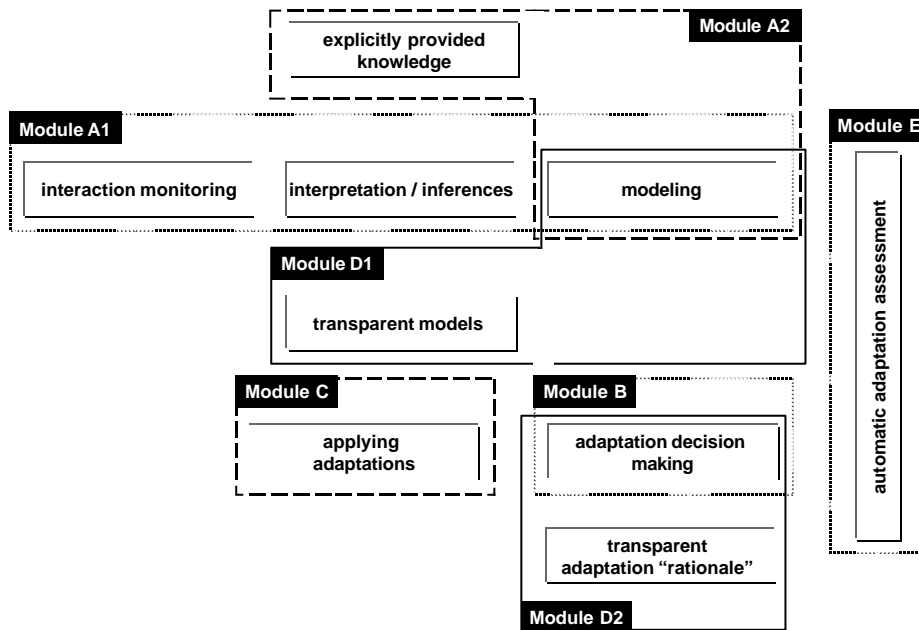


Fig. 2. The correspondence between evaluation modules and AUI model components.

Let us now move on to the presentation of the modules, wherein, for each identified module, the following information is provided: (i) components comprising the module; (ii) evaluation goal(s) and potential evaluation criteria; and, (iii) proposed evaluation methods, and prerequisites for supporting these methods.

3.3.1 Module A1

Comprises: *interaction monitoring*, *interpretations / inferences*, and *modeling*. The goal of evaluation in this module is to ensure that the models derived by the system through dynamic interaction assessment are “optimal”. Optimality in this context may be related to the following evaluation criteria¹: *correctness* of the interpretations / inferences (i.e., do the inferences / interpretations reflect that actual state of the entity being modeled?); *comprehensiveness* of the model (i.e., can the model represent in its entirety the inferred / interpreted information about the entity being modeled?); *redundancy* of the model (i.e., does the model contain “attributes” of the entity being modeled, which cannot be inferred from interaction?); *precision* of the model (i.e., how accurately does the model reflect the entity being modeled?); *sensitivity* of the modeling process (i.e., how fast does the modeling process converge to a comprehensive and accurate representation of the entity being modeled?); etc.

In the case of models that directly or indirectly involve the user (e.g., user modeling, plan recognition), one would need to employ a combination of evaluation methods to assess the degree to which the above criteria are met. Specifically, due to the fact that both observations and trace measures can only be used on overt behavior, not on thoughts or feelings or expectations (McGrath, 1995), methods in the *self report* category have to be used. Additionally, methods which allow the users to offer feedback *during* interaction are to be favored (to avoid remembering effects), although care should be taken that these methods are not too obtrusive with respect to the interaction itself.

Eliciting user feedback regarding the modeling process requires that at least a prototype of the system exists, with functional *interaction monitoring* and *inferencing / interpretation* components (the *modeling* component could be simulated). Furthermore, users should have some representation of the modeling process itself, which, in this case, can be constrained to the results of the process (i.e., the resulting model or models). If the AUI under evaluation also comprises a functional version of a *transparent models* component, then the latter can be used to that effect (although this might also necessitate a working *modeling* component). If such a component is not foreseen in the AUI (or not available at the time of the evaluation), then an alternative ad-hoc approach to the representation of the model should be sought (e.g., with an observer simulating the model, in a “wizard of oz” type of study).

Expert-based evaluation might also be of use in the early design and evaluation stages for Module A1. In particular, experts may be able to contribute towards the evaluation of *correctness* of inferencing / interpretations, and *comprehensiveness* and *redundancy* of the model. Such involvement of experts could be beneficial if part of the user model is related to the application’s domain model (e.g., in student models); if the inferencing / modeling process seeks to capture some special user characteristics (e.g., user’s ability to interact through a particular input device); etc.

3.3.2 Module A2

Comprises: *explicitly provided knowledge*, and *modeling*. This module is very similar to the preceding one, with the following exceptions:

¹ The proposed criteria sometimes refer to what is subjectively perceived by the user, while at other times they refer to what could be “objectively” measured or proven.

- Since there is no automatic assessment of the interaction, nor any attempt to elicit / infer information based on such assessment, any related evaluation criteria (including, for example *correctness*) are not relevant.
- Additional criteria that may be considered include: the *transparency* of the process (i.e., whether, and to what extent, the users can understand and / or predict how the information they provide affects the models maintained by the AUI); the *overhead* that may be imposed on the main interaction tasks by the explicit provision of knowledge; etc.
- The involvement of experts in the evaluation of this module might not yield as valuable results as in the case of Module A1. This is due to the fact that the direct “manipulation” of the model(s) is tightly coupled to the users’ mental model of what is being modeled and how, which may be quite hard to simulate or predict.

3.3.3 Module B

Comprises: *adaptation decision making*. The goal of evaluation in this module is to ensure that the adaptation decisions made by the respective component are “correct”. Correctness in this context may be related to the following evaluation criteria: *necessity* of adaptation (i.e., is an adaptation indeed required in the current interaction context?); *appropriateness* of adaptation (i.e., is the adaptation decided upon one that can cater for the requirements posed by the current interaction context?); *acceptance* of adaptation (i.e., does *the user* think that the adaptation is both necessary and appropriate?); etc.

A fundamental difference between this module and previous ones is that it does not (initially) require that any parts of the adaptation infrastructure have been implemented (although it does require that the alternative interaction artifacts have been designed). This is due to the fact that the adaptation logic relates interaction states (as these are depicted in the maintained models) to specific adaptations; thus, if such states can be reproduced or even simulated, it is possible to evaluate the related decisions “in context”. The decomposability of adaptation logic is of course constrained by the degree to which adaptation decisions affect each other (e.g., two decisions may be mutually exclusive, if they affect the same facets of interaction in different ways).

In practical terms, in a typical adaptation design cycle a theory, a set of hypotheses, or past empirical findings, will serve as input to the initial corpus of adaptation logic. This corpus can then be validated, in a first stage, using mostly formative evaluation methods to assess the *necessity* and *appropriateness* of adaptations.

Contrary to the above, it may be difficult (or even impossible) to extrapolate the overall *acceptance* of an adaptation decision in the same manner. This, combined with the requirement to further explore the other two criteria, when the entire corpus of adaptation logic is “active”, points to the necessity of a second stage of evaluation in this module, in which users will experience adaptation decisions in “real time”.

In either stage, to enable the participation of users in the evaluation, there needs to exist an explicit representation of the decisions made. This is a non-trivial requirement, especially in the case that the components that undertake decision making and adaptation application are separate (because, then, users would have to attain an understanding of a decision, without detailed knowledge of how it would be applied in practice). If the AUI comprises a *transparent adaptation* rationale component, then this could be utilized to offer the users the required representation. Otherwise, like in

the case of Module A1, a different (probably ad-hoc) approach to the representation of decision triggering (presumably on the basis of dynamic modifications in the maintained models) and decision making would be required. If none of the preceding were feasible, an alternative approach would be to treat modules B and C jointly, in terms of evaluation, as described in the section entitled “Evaluating across modules”.

Expert-based evaluation can also play a fundamental role in this module. This is true especially for the first of the two stages described above, and for the first two criteria proposed (i.e., *necessity* and *appropriateness*). The involvement of experts could lift the requirement for functional prototypes (or simulations) of the involved adaptation modules, if a structured approach was followed, within which adequate documentation and instruments were provided to the experts in order for them to be able to: (a) fully associate the interaction context (including user characteristics) that triggers a decision, with all the facets of interaction that the decision affects (and, of course, the ways in which it affects them), (b) assess the interplay of decisions on interaction, i.e., assess the possible / potential *combined* effects of sets of decisions triggered in the same (or similar) interaction contexts. This approach was followed quite successfully in the evaluation of the adaptation rule base of the adaptive user interface of the AVANTI Web browser (Stephanidis, Paramythis & Sfyarakis, 1999).

Finally, it is interesting to note that the evaluation of the current module as well as of Module C (described next), is both the most challenging and most interesting part of the evaluation of AUIs, as these two modules “contain”, in effect, the theory underlying adaptation in the user interface. This theory (typically expressed in the form of higher-level hypotheses) is what researchers have actually sought to evaluate in previous work. However, evaluating the AUI as a “black box” is what has restricted the scope and validity of efforts in the past, since not “separating” the module from the rest of the AUI, results in the concurrent assessment of possibly numerous influencing factors.

3.3.4 Module C

Comprises: *applying adaptations*. The goal of evaluation in this module is complementary to the one for Module B above, and can be expressed through criteria such as: *timeliness* of adaptation (i.e., is the decided upon adaptation applied in a timely manner - e.g., not too late?); *obtrusiveness* of the adaptation (i.e., how obtrusive, or obstructive is the application of an adaptation, with respect to the users' main interaction tasks); *user control* over the adaptation (i.e., can the user disallow, retract, or even disregard an adaptation?); etc. Furthermore, these criteria can be thought of as directly contributing towards the criterion of *acceptance* in Module B.

The evaluation of this module should be treated very carefully, and any related evaluation activity should be designed so as to measure only the criteria relevant to this module. The difficulty in doing so arises from the fact that the users “experience” the grand total of the system’s adaptive behavior through the adaptations that are effected (and of which they are aware). In other words, users, in all but the most trivial cases, might not be able to (and need not be asked to) distinguish between modeling, decision making and adaptation application. It is the task of the evaluator to factor out any “interference” from the preceding adaptation stages. One feasible approach to achieving this would be to evaluate this module only after Modules A_x and B have

been evaluated, and any necessary modifications to the respective AUI components have been made.

The evaluation of adaptation application with the involvement of end users requires that the AUI “feels” like a complete interactive system, i.e., functional prototypes (or simulations) of all AUI components should be present. This requirement stems from the need to enable users to situate themselves in the actual interaction context in which adaptations would take place. If this requirement is not met, then the only criterion that may be assessable is the degree of control that users feel they have over adaptations. *Timeliness* and *obtrusiveness* cannot be evaluated by end users unless they are actually “immersed” in realistic tasks or interaction situations.

The assessment of this module by end users also points to the direction of summative evaluation methods, and especially ones where factors that are external to the user-system interaction and may influence the interaction situation are reduced or controlled (such as, for example, controlled experiments). Expert-based evaluation is not likely to render any significant results in this module, with the exception of initial design cycles (targeted at identifying and treating any major flaws in how the system applies adaptations).

3.3.5 Module D1

Comprises: *modeling*, and *transparent models*. The goal of evaluation in this module is to ensure that the users’ perception of the maintained models matches the actual state of the models. This translates into evaluation criteria such as: *completeness* of the presentation (i.e., does the user have a full –perhaps abstracted– view of what is modeled and the current contents of the model?); *coherence* of the presentation (i.e., how well can the user understand the attributes of the model); *rationality* of the presentation (i.e., does the user understand why the model is in its current state?); etc.

The evaluation of this module can follow a two-staged approach. In the first stage, end users and experts can be involved in the assessment of the actual representation of the model, addressing issues such as the level of detail to be employed, the level of transparency that is necessary to provide the user with adequate information, but without exposing internal modeling details, etc. This stage, which may not even necessarily require the presence of interactive prototypes, can thus explicitly target the *completeness* and *coherence* criteria.

The second stage of the evaluation would have to address how users experience and perceive the model(s) *during* interaction. This stage is complementary to the previous one and is mainly intended to address the *rationality* criterion. To assess the latter, a user would actually need to have a full understanding of the interactions that have led to a particular situation, in order to be able to judge whether the presented model “makes sense”. During this stage, it would be preferable to elicit user feedback during interaction, so as to avoid any rationalization effects that may interfere with post-interaction evaluation. A possible compromise might be to structure interaction into small tasks and request users to provide their feedback between tasks (e.g., by answering short, targeted questionnaires).

3.3.6 Module D2

Comprises: *adaptation decision making*, and *transparent adaptation rationale*. This module is similar to the preceding one, with the main difference being that what is presented to the user is not a model, but rather the rationale underlying a particular adaptation (which could also have the form of a recommendation made by the system). Thus, evaluation criteria that may be relevant include: *coherence* of the adaptation rationale (i.e., how well can the user understand what the rationale refers to – e.g., what is / will be adapted and in what way?); *causality* of the rationale (i.e., does the user understand what triggered a particular adaptation?); etc.

Although there exist subtle differences between this module and Module D1, the evaluation can follow, in general, a similar approach. A notable difference is that the second stage of the evaluation (as described for Module D1) would, in this case, require that the AUI is functional (or simulated) almost in its entirety.

3.3.7 Module E

Comprises: *automatic adaptation assessment*. The goal in this module is to ensure that the system shares the same views as the users with regards to the “success”, or “failure” of adaptations. This goal differs significantly from the ones expressed in the previous modules, in that the users’ feedback regarding specific adaptations and their effects on interaction needs to be captured and subsequently compared to the system’s view of the same adaptations. From a different perspective, if this AUI component assesses and modifies the lower-level adaptation “strategies”, what needs to be evaluated is whether any such modifications are optimal from the perspective of the user.

Although, from an engineering perspective, the AUI component(s) involved in “adapting the adapter” operate at a meta-level with respect to the rest of the AUI components, this distinction may not be relevant from the perspective of evaluation. Specifically, it may be possible to treat these “meta-adaptations” as just another type of adaptations taking place in the interface. This would mean that meta-level adaptations are amenable to the same treatment as first-level adaptations, and can thus be included in the modular evaluation as this has been described so far. The feasibility of this approach will be investigated in the context of the evaluation of the adaptive Nautilus Web browser (see section entitled “Summary and future work” below).

3.3.8 Evaluating across modules

Evaluating each of the above modules in isolation may deliver significant results, but there remain a number of questions that cannot be answered unless modules are evaluated in combination. Additionally, some of the components may be tightly coupled in an implemented AUI, which might render it impossible to evaluate some of the modules in isolation. The rest of this section briefly presents some tentative module combinations and their significance in terms of AUI evaluation.

Modules Ax and D1: These modules capture the entire process of constructing models (automatically, or through explicitly provided information) and presenting these models to the user. Evaluating these modules in combination may offer a more global perspective on how users perceive modeling in the AUI, and allows one to investigate other relevant aspects, such as whether users are comfortable with whatever

private information is included in the models, whether they “trust” the system with such information, etc.

Modules B and C: These modules capture the process of deciding upon and applying adaptations in the AUI. Evaluating them in tandem may be inevitable if the AUI does not distinguish between the respective components in a way that allows for treating them separately. On the other hand, even if the AUI does distinguish between the components, it is possible to treat them jointly by: (a) enumerating all the possible methods in which an adaptation decision can be applied, and (b) treating each decision-method pair as a distinct decision.

Modules C and D2: By evaluating these modules in combination, one could, for example, address the questions of how “predictable” and how “controllable” users perceive the AUI to be. Whereas the perception of predictability might result from the user’s ability to understand the circumstances under which adaptation decisions are made, the perception of controllability might result from the users’ ability to control both the circumstances that lead to an adaptation decision, and the application of the decision as such.

Modules A, B and C: The combination of these modules captures the entire “traditional” adaptation cycle in an AUI, and can thus be thought of as evaluating the AUI as a “whole”. Although it is argued that the evaluation of this combination should not commence until the modules have been addressed individually, there are questions regarding the adaptive theory employed in the AUI, which can only be posed at this level (such as, for example, “does adaptive task guidance improve the ability of novice users to complete complex tasks in the user interface?”)

4. Summary and future work

This paper has proposed a new approach to the evaluation of AUIs, based on an approach that can facilitate and guide the modular assessment of various components / stages of the adaptation cycle. The paper has presented the rationale underlying the derivation of the approach, as well as the anticipated implications of its introduction in the evaluation of AUIs.

Portions of the proposed evaluation approach have been applied successfully in the evaluation of the adaptive user interface of the AVANTI browser (see Acknowledgements). The AVANTI browser is described in detail in (Stephanidis et al., 1998) and (Stephanidis, Paramythis, Sfyraakis, & Savidis, 2001), whereas details of the related evaluation activities can be found in (Stephanidis, Paramythis, & Sfyraakis, 1999) and (Stephanidis et al., 2001). The approach will be applied in its entirety in the evaluation of the Nautilus Web browser, a follow-up development of AVANTI. The Nautilus browser is currently under development in the context of the Nautilus project². Adaptation in the Nautilus browser is intended to facilitate the accessibility and increase the usability of the browser, for all its potential users. The browser will inherently support interaction with able-bodied users, users with vision problems (includ-

² Project EPET 98AMEA28 Nautilus, funded by the Hellenic Ministry of Development.

ing blindness), and users with various degrees of motor-impairment. Furthermore, the browser will support synchronous interaction between sighted and blind individuals.

The evaluation of Nautilus is scheduled to commence in August 2001, and will be based on the proposed modular evaluation approach. A multi-disciplinary group of experts will participate in the evaluation activities, with a background in usability evaluation, user modeling, AUI engineering, and assistive technology. The user population sample will include both able-bodied and disabled users, with varying degrees of experience in using computers and the Web. In a first round of evaluation, all individual modules will be addressed, except Module E (comprising automatic adaptation assessment). In a second round of evaluation, Module E and all module combinations except B-C will be addressed. The overall goal of the evaluation is to assess the degree to which automatic interface adaptation can increase the accessibility and usability of a general-purpose user interface (such as that of a Web browser) and to identify the particular types of adaptation that are of most benefit to the users in this direction.

Finally, we expect that the full application of the proposed approach in the case of Nautilus will provide feedback both towards the improvement of the approach, and towards a better understanding of what existing (or new) evaluation methods and techniques have to offer in the direction of AUI evaluation.

Acknowledgements

Part of the R&D work reported in this paper has been carried out in the context of the ACTS AC042 AVANTI project "Adaptive and Adaptable Interactions to Multimedia Telecommunications Applications", partially funded by the European Commission (DG XIII). The AVANTI consortium comprises: ALCATEL Siette (Italy) - Prime contractor; CNR-IROE (Italy); ICS-FORTH (Greece); GMD (Germany); University of Sienna (Italy); MA Systems (UK); MATHEMA (Italy); VTT (Finland); ECG (Italy); University of Linz (Austria); TELECOM ITALIA (Italy); EUROGICIEL (France); TECO (Italy); Studio ADR (Italy).

References

- Boyle, C. & Encarnacion, A. O. (1994). Metadoc: An Adaptive Hypertext Reading System. *User Modeling and User-Adapted Interaction*, 4, 1-19.
- Brusilovsky, P. & Eklund, J. (1998). A Study of User Model Based Link Annotation in Educational Hypermedia. *Journal of Universal Computer Science*, 4, 429-448.
- Brusilovsky, P. & Pesin, L. (1998). Adaptive navigation support in educational hypermedia: An evaluation of the ISIS-Tutor. *Journal of Computing and Information Technology*, 6, 27-38.
- Eklund, J. & Brusilovsky, P. (1998). The value of adaptivity in hypermedia learning environments: A short review of empirical evidence. In P. Brusilovsky & P. De Bra (Eds.), *Proceedings of Second Adaptive Hypertext and Hypermedia Workshop (at the Ninth ACM International Hypertext Conference - Hypertext'98) (Computing Science Reports, Report No. 98/12) (pp. 13-19)*. Eindhoven: Eindhoven University of Technology.

- Hoppe, H., Tauber, M., & Ziegler, J. (1986). A Survey of Models and Formal Description Methods in HCI with Example Applications. ESPRIT Project 385 HUFIT. Report B.3.2.a. 1986.
- Höök, K. (2000). Steps to take before intelligent user interfaces become real. *Interacting with Computers*, 12, 409-426.
- Höök, K. & Svensson, M. (1999). Evaluating Adaptive Navigation Support. In M. Maybury (Ed.), *Proceedings of the 1999 International Conference on Intelligent User Interfaces - IUI'99* (pp. 187). Redondo Beach, CA: ACM Press.
- Höök, K., Karlgren, J., Waern, A., Dahlbäck, N., Jansson, C. G., Karlgren, K., & Lemaire, B. (1996). A Glass Box Approach to Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*, 6, 157-184.
- Jordan, P. W. (1998). Methods for Usability Evaluation. In *An Introduction to Usability* (pp. 51-80). London - Bristol: Taylor & Francis.
- Kaplan, C., Fenwick, J., & Chen, J. (1993). Adaptive hypertext navigation based on user goals and context. *User Modeling and User-Adapted Interaction*, 3, 193-220.
- Krogsäter, M., Oppermann, R., & Thomas, C. G. (1994). A User Interface Integrating Adaptability and Adaptivity. In R. Oppermann (Ed.), *Adaptive User Support: Ergonomic Design of Manually and Automatically Adaptable Software* (pp. 97-125). Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- McGrath, J. E. (1995). Methodology Matters: Doing Research in the Behavioral and Social Sciences. In R.M. Baecker, J. Grudin, W. A. S. Buxton, & S. Greenberg (Eds.), *Readings in Human Computer Interaction Toward the Year 2000* (2nd ed., pp. 152-169). San Francisco, California: Morgan Kaufmann Publishers.
- Meyer, B. (1994). Adaptive Performance Support: User Acceptance of a Self-Adapting System. In B. Goodman, A. Kobsa, & D. Litman (Eds.), *Fourth International Conference on User Modeling (UM94)* (pp. 65-70). Hyannis, MA.
- Oppermann, R. (1994). Adaptively supported adaptability. *International Journal of Human Computer Studies*, 40, 455-472.
- Sary, C. & Totter, A. (1997). How to Integrate Concepts for the Design and the Evaluation of Adaptable and Adaptive User Interfaces. In C. Stephanidis & N. Carbonell (Eds.), *3rd ERCIM Workshop on "User Interfaces for All"* (pp. 7.1-7.15).
- Stephanidis, C., Paramythis, A., Sfyarakis, M., & Savidis, A. (2001). A Case Study in Unified User Interface Development: The AVANTI Web Browser. In C. Stephanidis (Ed.), *User Interfaces for All - Concepts, Methods, and Tools* (pp. 525-568). Mahwah, NJ: Lawrence Erlbaum Associates.
- Stephanidis, C., Paramythis, A., & Sfyarakis, M. (1999). Evaluating Adaptable and Adaptive User Interfaces: Lessons Learned from the Development of the AVANTI Web Browser. In A. Kobsa & C. Stephanidis (Eds.), *5th ERCIM Workshop on "User Interfaces for All"* (pp. 22.1-22.6).
- Stephanidis, C., Paramythis, A., Sfyarakis, M., Stergiou, A., Maou, N., Leventis, A., Paparoulis, G., & Karagiannidis, C. (1998). Adaptable and Adaptive User Interfaces for Disabled Users in AVANTI Project. In *Proceedings of the 5th International Conference on Intelligence in Services and Networks (IS&N '98, "Technology for Ubiquitous Telecommunications Services")*.
- Totterdell, P. (1990). Introduction. In D. Browne, P. Totterdell, & M. Norman (Eds.), *Adaptive User Interfaces* (pp. 1-14). London: Academic Press.
- Totterdell, P. & Boyle, E. (1990a). The Evaluation of Adaptive Systems. In D. Browne, P. Totterdell, & M. Norman (Eds.), *Adaptive User Interfaces* (pp. 161-194). London: Academic Press.
- Totterdell, P. & Rautenbach, P. (1990b). Adaptation as a Problem of Design. In D. Browne, P. Totterdell, & M. Norman (Eds.), *Adaptive User Interfaces* (pp. 61-84). London: Academic Press.

- Weber, G. & Specht, M. (1997). User modeling and adaptive navigation support in WWW-based tutoring systems. In C. Paris & C. Tasso (Eds.), *Proceedings of the Sixth International Conference on User Modeling (User Modeling '97)* (pp. 289-300).
- Weibelzahl, S. & Lauer, C. U. (2001). Framework for the Evaluation of Adaptive CBR-Systems. In U. Reimer, S. Schmitt, & I. Vollrath (Eds.), *Proceedings of the 9th German Workshop on Case-Based Reasoning (GWCBR01)* (pp. 254-263). Aachen: Shaker.