

Research Workshop on Expert Judgment, Human Error, and Intelligent Systems

Barry G. Silverman

■ This workshop brought together 20 computer scientists, psychologists, and human-computer interaction (HCI) researchers to exchange results and views on human error and judgment bias. Human error is typically studied when operators undertake actions, but judgment bias is an issue in thinking rather than acting. Both topics are generally ignored by the HCI community, which is interested in designs that eliminate human error and bias tendencies. As a result, almost no one at the workshop had met before, and the discussion for most participants was novel and lively. Many areas of previously unexamined overlap were identified. An agenda of research needs was also developed.

On 3–4 August 1992, a 2-day workshop was held on how knowledge-based systems could reduce the mistakes of experts and other proficient task performers engaged in real-world tasks. Many have concluded that biased judgment (for example, repetitively using a mistaken schema or heuristic) and accidental human error (for example, a slip or lapse in implementing a correct schema) are the principal causes of major industrial catastrophes, transportation accidents, medical misdiagnoses, forecasting failures, and the like. The interesting questions examined at this workshop included, Why do errors occur? How

should we model error processes? What are the most promising routes to error detection and mitigation? What should be the role of knowledge-based systems?

The workshop brought together 20 participants from 8 European countries and the United States in the fields of psychology, judgment, and decision making, human-computer interaction (HCI), and AI to discuss the issues based on their own research. Most of the participants had not previously met or heard the subtleties of their respective views. Thus, the stage was set for useful discussion and bridge building.

The workshop was organized by Jens Rasmussen, Risoe Labs, Denmark; Paul Booth, Salford University, United Kingdom; Mike Donnell, George Washington University; Gerhard Fischer, University of Colorado at Boulder; John Fox, Imperial Cancer Research Fund, United Kingdom; Sture Hagglund, University of Linköping, Sweden; Erik Hollnagel, Computer Resources International, Denmark; Barry Silverman (chair), George Washington University; and Masoud Yazdani, Exeter University, United Kingdom. The workshop was sponsored by the European Coordinating Committee for AI and was held in conjunction with its biannual meeting in Vienna, Austria. The

workshop was cosponsored by the American Association for Artificial Intelligence.

What Should Be Replaced: The Human, the Technology, or Neither?

Any time researchers working in the error-mitigation field get together, there is generally some discussion of the three pathways to consider. As a first path, in some situations, human error can be eliminated through automation. A current example is receptionists and telephone operators being replaced by voice mail. One participant at the workshop kept suggesting that the automation route was the most reliable, but the bulk of the attendees did not subscribe to this view.

One reason that this view was discounted is that automation is costly. The domain of aircraft carriers in the United States Navy was discussed at some length and from several perspectives. On each sea tour, it seems that a few seamen lose their lives through carelessness, accident, and error (for example, falling off the side, getting sucked into engines, crashing into the ship). However, the navy does not have the funds—or even know-how—to automate all these jobs. Even in the far simpler voice mail example, how to automate doesn't seem obvious. One presentation showed that users can exhibit increased errors and reduced performance in telephone use as a result of voice mail. A related concern is that software itself can never be guaranteed to be error free (for example, witness recent near-nation-wide telecommunication system crashes). Finally, innumerable situations are too ill structured for automation. Humans must be kept in the task-performance loop. For these and related reasons, most participants were uninterested in researching the option of replacing humans with software.

Instead, a large minority of participants were acutely interested in the second error-mitigation pathway, which is to redesign organizations, procedures, technology, and so forth,

so that they reduce humans' propensity to err. The Apple MACINTOSH look and feel drove this point home to society, and other electronic-based industries (VCRs, televisions, cameras, and so on) are finally beginning to change their digital displays to a more intuitive, error-reducing approach as well. From an organizational design perspective, it seems that aircraft carriers are a superb model of error reduction because of the interface flexibility between different teams of sailors performing diverse but interrelated tasks.

The HCI workshop participants placed the highest priority on this overall redesign pathway, and several presentations addressed it. Under the sponsorship of Bellcore, Javier Lerch and Brian Huguenard of Carnegie Mellon University presented a human error-modeling approach to the (re)design of voice-mail menus. The conventional wisdom in voice mail is to use short menus with many layers to avoid the user forgetting what is on a given layer of the menu. By simulating working memory (using a mixed production system-connectionist approach) and running verification experiments, Lerch and Huguenard showed that conventional menus promote *navigational errors*, errors in traversing the menu because of forgotten higher-level choices. Their results to date seem to indicate that broad but shallow menus might be better. Further research is warranted on the impacts of alternative menu designs, simulations of diverse human working-memory capacities, and so on. Also, the research thus far is *content free*; that is, it ignores ambiguity in the content of the menu selections.

A number of other participants offered equally excellent redesign presentations, particularly during the HCI panel session. The initial attempts to develop *ecological interfaces*, screens that display information at the level at which it is cognitively useful, are another example. Unecological interfaces load the screen with graphics and direct-manipulation objects that might be faithful domain metaphors but that don't directly improve the nature of

problem solving itself. Also, there was a short discussion about the nature of groups and how inappropriate corporate culture can foster error. It was pointed out that improved group culture might be one of the major, underresearched areas that could greatly reduce human error, particularly in safety-critical situations.

In the end, most workshop participants agreed with the HCI view that it is important to diagnose the errors that result from poor design and use after-the-fact redesign to minimize error reoccurrence wherever possible. It would be foolish, for example, to build an expert system to help people set the clock on their VCR when simple VCR design fixes could be made to make the process of how to set it intuitively obvious. Where it is not possible to eliminate all sources of human error through after-the-fact redesign or where designs are already locked in, then AI solutions might help to mitigate the consequences of human error. That is, in some situations, after-the-fact diagnosis and redesign might be too late. This point brings up the third and final pathway to human error mitigation—using knowledge-based systems to help humans detect and recover from their errors.

The third pathway assumes that complex human-computer systems might better be designed around the expectation of human errors. The philosophy here is that the problem is not the errors themselves but the consequences of leaving the errors unattended. This third path of having the machine help humans recognize and recover from their errors was a major concern for the bulk of the workshop participants. The next several sections explain this area more fully.

It became clear through discussion that there is a research gap in the HCI community. That is, there is such an overriding focus on the redesign of items so that they will be error free that the HCI community is currently paying almost no attention to error-recovery issues and research. However, organizational designs, procedures, technology, and so on, will

never entirely eliminate human error. This issue was not resolved at the workshop, but several attendees from the HCI and AI communities agreed to meet and discuss error recovery more fully in the future. Also, a draft set of research issues was prepared by the cochairs of the HCI panel and discussed by the panelists. This list is given later in A Shopping List of HCI Challenges.

Behavioral and Cognitive Psychology: A Role for Each

Workshop participants generally fell into one of two groups: those working on errors related to human action (*behavior*) and those working on errors related to human judgment (*cognition*). Judgments often are precursors to actions, but researchers tend to focus on the two as separate tasks. The action- and judgment-error research communities are similar in that they share three top-level goals: (1) both are trying to develop taxonomies of error types, (2) both are trying to overcome the general lack of models of human error (and error-recovery) processes, and (3) both are interested in getting machines to help detect and repair the errors.

There are two differences between the two communities. First, actions take place in time under dynamic circumstances, such as flying a plane, driving a car, or controlling a power plant. People who act are usually not experts in the expert system sense. Rather, they are proficient performers, competent practitioners, and professionals. Human action researchers generally assume that their subjects made the correct judgments about what schema to activate. It is the slips, lapses, and other real-time schema-execution errors that interest these researchers.

Second, judgments are usually studied in static environments without strict time limitations. Also, researchers who study judgments often are interested in heuristics and cognitive processes used to the point of deciding what course of action to pursue. They focus on systematic tendencies or reproducible cognitive

biases arising from these heuristics. These researchers ignore the action stage or assume the action schemas will be executed as planned.

These two communities also overlap somewhat. For example, chess players are experts, but what they do is not characterized as judgment but, rather, situation assessment, planning, and so on. They and certain other categories of professionals (for example, doctors and trial lawyers) execute schemas that might variously be studied by both communities. Conversely, action researchers increasingly are coming to see recurring patterns in some of the accidental schema-execution slips and lapses of process operators. Many apparently accidental execution errors, in fact, might be the result of systematic biases of heuristics used during action-based judgments.

Despite any grey areas, the differences in action versus judgment focus translate into real differences in how the two communities investigate the three goals they have in common. In terms of the first goal, several papers were devoted to error taxonomies. Action-oriented researchers tend to focus on the automaticity errors, such as the various types of slips and lapses that can occur when executing a proper schema. For example, Hollnagel gave an invited talk on developing a rigorous, consistent, and machine-implementable phenology of erroneous actions. *Phenotypes* are the manifestations of erroneous actions (behavioral view), as opposed to *genotypes*, which include the causes (cognitive processes). Hollnagel's work is largely based on timing (for example, premature start, omission, delayed ending) and sequence (for example, jump forward, jump backward, reversal) errors. Its value lies in the fact that the taxonomy only includes errors that can be observed (for example, loss of attention is a cause, not an observable phenotype); thus, the taxonomy can be machine implemented.

In contrast, I presented a taxonomy of genotypes (causes) of judgment errors. Although probably a million pages of published literature exist on judgment biases, almost

none of these bias models are machine implementable. I explained a feasible, yet time- and effort-intensive procedure that my colleagues and students are pursuing to develop judgment-related genotypes. Because of the large scope of this undertaking, rather than tackle specific classes of error (for example, time based), our approach is to examine all the types of error that arise in a given class of task (for example, information acquisition, forecasting, decision making under uncertainty). There is a similar lack of machine-implementable taxonomies in both communities; however, this obstacle is likely to be eliminated sooner for the action community.

The behavioral versus cognitive psychology distinctions carry into the second goal of how the two communities develop models of error and error-recovery processes. The behaviorists build situation, rather than mental, models. They find mental models too slow and inaccurate for error monitoring in real-time, safety-critical systems. Mental models and models of intentionality are too

erated conclusion that affordances missed are errors. Greenberg contends that affordances are too low level. Linguistic-semantic techniques such as those of Terry Winograd, Stanford University, are more useful. Instead of affordances missed, Greenberg favors tracking commitments not fulfilled. Also, Greenberg's research with error-monitoring systems shows that often, there is user resistance to suggestions of commitments missed or of active remediation. Simply allowing users the ability to authorize or unauthorize the machine-suggested action is insufficient. Among other things, more research is needed to show that machine-generated remedies are reliable and sometimes crucial. For example, as Booth pointed out, a Boeing 737-400 flight-management system demanded the complete attention of the copilot throughout the Kegworth air disaster. At no time, did the system offer suggestions to help avert or diminish the crisis.

Mental modeling, in turn, is proving useful to both the HCI redesign and the judgment-error communi-

All is not a bed of roses with situational models either...

coarse grained and lead to unacceptably high rates of false alarms. Moreover, the closer the system gets to a hazard, the less important the operator's mental state is, and the more important it becomes to focus on avoiding the hazard. Situational or engineering models are useful here because they can accurately infer whether observable action streams hold errors that might jeopardize the safety of the system.

All is not a bed of roses with situational models either. For example, Alan Greenberg from Search Technology Inc. pointed out that a lot of context information is required before inserting machine actions into the operator's action stream. Related to this idea are the Gibsonian theory of *affordances* (for example, a chair affords sitting) and the machine-gen-

ties. The presentation by Lerch and Huguenard illustrated the value of cognitive models to the redesign community. Paul Johnson from the University of Minnesota provided a lucid example for the judgment-error community. He studied why expert accountants at major accounting houses are so poor at detecting fraudulent bookkeeping at firms and banks that intend to deceive investors and auditors. By building cognitive models of their heuristics and biases, Paul Johnson was able to replicate and explain the fraud-detection (judgment) errors that accountants succumb to.

Given the importance of cognitive modeling to so many of the researchers, another welcome talk was by Fox, who discussed a specification language that he has been

developing to help cognitive modelers make their models more precise. Fox gave SOAR as an example; SOAR is a cognitive model that one can study by either reading a textbook theory that might or might not be faithful to the actual computational model or studying 30,000 lines of code. Using his tool, he reduced the actual code to a few pages of cognitive model specifications. These specifications are themselves rules. Researchers can run these rules to exercise the model and study the consequences of different model assumptions. The error community will more rapidly harness the cognitive research literature and results as more psychologists develop similar computational-level specifications for other cognitive models.

The third and last shared goal of the action- and judgment-error communities concerns the implementations of systems on the machine. The entire workshop was interested in real-world applications, as the discussion to this point indicates. The papers by Greenberg, Hollnagel, Lerch, Silverman, and others included descriptions of, and lessons learned from, systems being fielded.

Needed: A Theory of Context

One issue that kept arising in session after session was context. Expertise in one context might be an error in another and vice versa. For example, Vassilis Moustakis from the Technical University of Crete, Greece, presented an example of a medical domain where the theoretically correct expertise was being contradicted in practice. It turned out that patients from remote villages, who were unlikely to continue textbook-prescribed medication and office visits, were being recommended for immediate, outpatient surgical correction of their problems. The local doctors were seeing to their patients' needs in the context of the delivery system rather than as recommended by a theoretically correct approach.

Most participants agreed with Ken Ford's (University of West Florida) framework for handling such exper-

tise and error-context issues. He suggested three types of experts: Type-1 experts are socially selected experts, those who have some sort of societal recognition or perception of expertise by their blind-faith followers. Examples include con artists and quacks and sincere-but-misguided individuals. Type-2 experts are socially selected, but they also possess personally constructed expertise that comes from functioning and practicing with their degree, title, job, and so on. These experts, through (possibly fallible) experience, have constructed a rational domain model that allows them to make functional decisions that their local constituents value and find difficult to make themselves. The doctors in Moustakis's case study fall into this category, as do other professionals, such as lawyers, accountants, coaches, and teachers, who also must keep passing some domain-relevant performance tests. Type-3 experts, reality-relevant experts, are able to pass all these tests plus the scrutiny of scientific society. Type-3 experts include major theoreticians of a given discipline. Most experts exist at the middle level, and only a few reach the highest level. At all three levels, however, expertise is located in the expert in context; that is, experts exist in social niches dependent on social validation (Moustakis's doctors' rules are wrong for a metropolis). Also, expertise is subject to a relatively short half-life. The duration of the half-life grows as one progresses up the three rungs of the expertise ladder. Conversely, the fallibility frequency of the expert decreases as the ladder is ascended. However, at each level, expertise remains at the mercy of a variety of metaselectors.

Virtually all participants agreed on the need for AI to have a better handle on the context problem, but they disagreed on what was needed. There were two basic camps: the engineers and the scientists. The engineers felt it was best to work toward a simple theory of context, one they could immediately implement. This group offered various proposals. For example, Laurent Siklossy from University of Savoy, France, suggested all expert

systems should contain a metaframe that is used to describe the limits of their knowledge and the extent of their ignorance. Siklossy coined the term *ignorance representation* and indicated that such a frame could help an expert system with such problems as advising its users when it was being used out of context. The metaframe is an admittedly weak implementation that would be unable to handle unexpected situations and would leave a number of context issues unresolved.

The scientists felt that only a full-blown theory of context would allow knowledge-based systems and their designers to handle the problem properly. Such a theory would allow a knowledge-based system to be socially aware. It would help the knowledge-based system to know what knowledge and tools to use, when to use them, and for whom. Further, this theory would permit the knowledge-based system to handle communication difficulties and other types of trouble. Although a laudable long-term research goal, developing a full-context theory involves solving the complete AI problem as well as having an accepted theory of human cognition and group behavior. Although research is needed to overcome the overall context obstacle, pragmatic solutions are also needed. The latter could provide significant benefits in the near term.

Bug Theory Might Work After All

In intelligent tutoring systems, the errors that students make in a learning task are the *bugs*, and an enumeration of these errors is the *bug theory* of this domain. Many researchers have concluded that bug theory has limited value because in even seemingly simple domains, there is an explosion of possible bugs. For example, in the domain of learning subtraction, school children's bug catalogs are vast and innumerable.

Despite its problems in novice training, bug theory might pose a successful paradigm for professional domains. Paul Johnson presented results that show that in sophisticated professional domains (for example, accounting,

medicine), the bugs are limited to a few repeated procedural errors. A half dozen bugs often account for 60 to 90 percent of all professionals' errors because unlike open domains, such as learning, professions can be highly constrained to a few heuristics that everyone follows. This finding is consistent with results concerning cognitive bias in judgment.

What is interesting here is that Paul Johnson's findings (1) provide a bridge between the two sets of concepts (bugs and biases) and (2) open the door to a new (non-educational) use for bug theory. The general significance of Johnson's results prompted discussion among several participants. It was pointed out that the use of the term *bugs*, rather than biases, might help computer scientists better understand the task of designing expert-critiquing systems to help humans repair their buggy procedures in decision support applications.

Can We Improve Critiquing Systems?

Expert-critiquing systems already are a viable, commercially successful technology. However, they often fail to say the right thing at the right time, intrude when users don't want them, and appear to be situation-insensitive automatons. Four speakers explained how they were addressing these challenges. From these discussions, it seems that a need exists for research on alternative designs of both the differential analyzer and the text generator of the critiquing algorithm. Some of the research presented by these four speakers sheds further light on these needs.

Consider the challenge of trying to get a critic to say the right thing at the right time and be more situation sensitive. In judgment situations, this task might require the critic to evaluate the human's choice of mental model for a task, determine if this model is suboptimal, and only then interrupt and attempt to shift the human to a more normative mental model. I presented an actual working example for a forecasting task for the United States Army. Here, the critics test for several commonly occurring

bugs or biases related to nonregressive models. If these biases are present, the critic sends the user through a step-by-step procedure that causes him or her to realize the value of a regression approach to the forecast. No one actually performs a regression, but the result is that the user adopts a more normative solution to the problem. Paul Johnson raised the concern that this mental-modeling approach relies on bias theory to help detect errors. However, bias theory has a history of nonrelevance to real-world tasks. I countered that an increasing amount of recent empirical evidence, such as the army forecaster and accounting fraud examples, indicates that professionals might regularly succumb to common judgment biases. The use of published bias theory helps critic designers get started (it is a generative bug theory), but they must adapt and extend these theories for the particular application.

By contrast, this same challenge (situational sensitivity and saying the right thing) leads to an entirely different architecture and implementation concept for action than for judgment situations. Critics of operators in time-critical activities probably don't have the time or ability to infer mental models, as mentioned earlier. It is more important for the critic to monitor the hazard consequences of a human's actions than it is for the critic to infer the cause of the human's actions. For action critics to work in real time, it appears necessary to replace the differential analysis module used in judgment critics, where the critic compares the user's solution, intentions, mental state, and so on, to a normative ideal. Instead, action critics must be streamlined to react only when system hazards appear imminent. These ideas were advanced by Greenberg along with the use of assessment nets to constantly monitor the distance of the system from a hazard state. Assessment nets are organized in order of situation severity and are constantly monitored to anticipate the most hazardous consequences of a human operator's actions. If a hazard is approached, an assessment net

triggers the appropriate remedy suggestion. An assessment net can incorporate factors such as the human operator's beliefs, cognitive work load, and level of fatigue, but research to date has omitted these factors.

Another related challenge—the critic saying the right thing in a way that avoids the impression of being a mechanical automaton—was also discussed. Rather than worrying about how to design the differential analyzer, this challenge brings up issues of architectures for eliminating canned text and generally improving the quality of a critic's textual dialog. Hagglund gave an overview of several critic research efforts that his group has been doing to improve the quality of a critic's textual dialog.

Afterward, his student, Jonni Harrius, presented some research on dynamic text generation for a criticism using rhetorical structure theory (RST) and other linguistic methods. In particular, once the differential analyzer identifies an error, the critique builder selects a schema from a library of aggregate schemas of argument structures. The aggregate schemas are descriptions of how to construct the argumentation. For example, an aggregate schema for *disagree* presents a negative statement and motivates it with support. RST indicates the nucleus and satellite pieces of text for any given element of an aggregate schema. With the help of a user model (what the user knows already), Toulmin microargument forms (for example, data, warrant, claim) and RST are used to instantiate the aggregate schema with actual text relevant to the current situation. At the end of a given criticism, the user model is updated.

Research to date has progressed along the lines of studying human-to-human criticism in a real-world domain to isolate the relevant rhetorical structures, relations, and schemas. The goal of the research is to develop high-quality machine texts. The goal does not include two-sided human-computer discussions or other natural language issues.

A Shopping List of HCI Challenges

An international HCI panel consisting of Peter Johnson (Queen Mary College, London, United Kingdom) and Sotiris Papantonopoulos (George Washington University), cochairs, and Booth, Frances Brazier (University of Amsterdam, The Netherlands), Mark Maybury (Mitre-Bedford), and Pat Patterson and John Rosbottom (Portsmouth Polytechnic, United Kingdom) presented and discussed the following list of issues in HCI and multimedia:

First, before he retired, Rasmussen, along with Vincent, published a provocative proposal for *ecological interfaces* that reduce human error by making the invisible visible in safety-critical situations (Vincent and Rasmussen 1988). This proposal suggests that in safety-critical situations, most interfaces fail to present information in units that are cognitively natural to problem solving and error mitigation. Errors arise in part because human users must cognitively map screen information into the units needed to solve the problems.

Second, explanation does not need to be in text form. From an error-reduction viewpoint, questions arise about what the optimum mix is of media, audio, video, and so on.

Third, one of the features of direct manipulation is the availability of immediate, fine-grained semantic feedback about the consequences of the user's actions. In real-time systems, the time delays are so large that this style of interaction might be inappropriate. What should be done about the lag between the command and the effect on the world in terms of feedback to the user?

Fourth, the theory of minimalism (for example, John Carroll's [1990] Nurnberg Funnel) suggests that computers (1) be concise with the ability to elaborate (for example, through hypertext), (2) support guided exploration, and (3) offer error-recognition and recovery abilities. Little has been learned to date about how to do the last of these items. That is, it is often taken for granted that error recognition and recovery will be built into

any good interface; so, few guidelines have emerged, and many systems overlook this feature. What can be done to reverse this situation, and what have we learned already?

Fifth, much trouble and many misunderstandings could be mitigated if the computer could just recognize and react to the human's intentions and goals. However, interfaces that adapt to individual differences and personal objectives have been disappointing. More effective user models and adaptive interfaces appear unlikely in the near term.

Sixth, there seem to be no good cognitive models of core activities, such as browsing, searching, controlling, and fault finding. Although these activities underlie much of the interaction required between people and complex systems, little advice is available about the way the user interface should support these core activities. What are the prospects for better cognitive models of core activities that would lead to prescriptive guidance for interface designs?

Seventh, it seems that part of the error problem is the lack of any understanding about how to predict the way a system is going to behave when a person is interacting with it. There is a need for case histories of known design errors and discussion of how the error(s) could have been predicted.

Conclusions and Next Steps

In summary, it seems that the field of knowledge-based mitigation of proficient practitioners' errors is still immature. The years since several predecessor workshops have witnessed a blossoming of many new lines of investigation, but they have yet to bear much fruit in practice. Mid-1980s researchers were merely proposing new lines of investigation, but in 1993, many researchers are able to discuss real results from working knowledge-based systems and actual field experiments. We have a better idea of where the true obstacles lie and what hasn't proved to be helpful. There are a number of laboratory prototypes and about-to-be-

fielded systems that promise to yield still further insights in the near term.

It will be vital for this community to maintain interdisciplinary links and to regularly reconvene, particularly because major breakthroughs might be just around the corner. It would also be useful for certain research communities to expand the range of what they typically do. As examples, (1) psychologists should attempt to produce computer-implementable cognitive model specifications; (2) HCI researchers should consider ways to add active error mitigation into their designs; (3) critic builders should more aggressively pursue both taxonomies and mental models of professional judgment bugs and linguistic and rhetorical approaches to improving criticism texts; and (4) for surface credibility purposes, the automaticity error (slips, lapses) field needs to attain and document several application successes, particularly in real-time hazard avoidance.

These and other extra community research efforts would lead to a test of an integrated set of the latest psychological, AI, and HCI suggestions and advance the error-mitigation field significantly.

References

Carroll, J. M. 1990. *The Nurnberg Funnel: Designing Minimalist Instruction for Practical Computer Skill*. Cambridge, Mass.: The MIT Press.

Vincent, K. J., and Rasmussen, J. A. 1988. *Theoretical Framework for Ecological Interface Design*, Technical Report M-2736. Risoe, Roskilde, Denmark.

Barry G. Silverman is a professor of AI and human factors at George Washington University. Most of his research for the past half-decade has been on expert-critiquing systems, some of which led to an Innovative Application of AI label in 1991. Silverman is a member of the American Association for Artificial Intelligence, a fellow of the Institute of Electrical and Electronics Engineers and the Washington Academy of Science, and a holder of the Schubert award for teaching and an honorary mention in the Edelman Prize for Management Science Achievement (1992).