



# **The burden of computer advice**

## *Using expert systems as decision aids*

*Ronald T. Kessel*

**Defence R&D Canada**

Technical Report

DRDC Atlantic TR 2003-241

December 2003

This page intentionally left blank.

Copy No:

# **The burden of computer advice**

*Using expert systems as decision aids*

Ronald T. Kessel

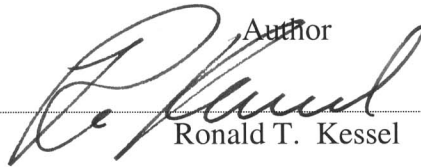
**Defence R&D Canada – Atlantic**

Technical Report

DRDC Atlantic TR 2003-241

December 2003

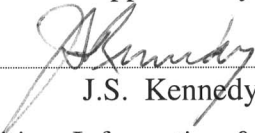
Author



---

Ronald T. Kessel

Approved by

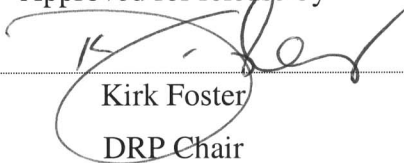


---

J.S. Kennedy

Head/ Maritime Information & Combat Systems

Approved for release by



---

Kirk Foster

DRP Chair

© Her Majesty the Queen as represented by the Minister of National Defence, 2003

© Sa majesté la reine, représentée par le ministre de la Défense nationale, 2003

## Abstract

---

Where expert systems are not trusted to replace human decision makers, they have instead been proposed as decision aids. By reporting its automatic decisions the expert system naturally helps the human arrive at better decisions, or so the reasoning goes. On further analysis, one quickly discovers that the usual vision for decision aids entails contradictions that are confusing if not disabling. Computer advice invokes unfamiliar deliberation and ambiguity about its interpretation within the context of a particular decision and about system performance, for which neither decision makers nor system developers are prepared. Here we examine the problematic nature of computer advice, examining in particular the mechanisms (or lack of them) by which a computer's decision becomes helpful to a human decision maker, identifying common fallacies in the decision-aid concept, setting minimum standards that decision aids must meet if they are to be useful, pointing to the crucial role of trust, and illustrating how system developers can integrate expert systems following rational principles of trust. Decision makers who have known the frustration of using expert-system decision aids should find confirmation of their experience and material for defence against system developers who do not appreciate the implications of the systems they propose, and developers should be redirected by this work toward more realistic and profitable principles of decision-aid design.

## Résumé

---

Lorsqu'on ne se fie pas suffisamment aux systèmes-experts pour remplacer les décideurs humains, on les propose comme aides à la décision. Lorsqu'il communique les décisions auxquelles il a conclu automatiquement, le système-expert aide l'humain à prendre de meilleures décisions, ou du moins c'est ce qu'on croit. Avec une analyse plus poussée, on découvre rapidement que la perception habituelle des aides à la décision est porteuse de contradictions confondantes ou même invalidantes. Les conseils produits par les ordinateurs entraînent des discussions relatives à des domaines non familiers et leur interprétation donne lieu à de l'ambiguïté dans le contexte de décisions particulières et en matière de performances des systèmes, pour lesquelles ni les preneurs de décisions, ni les développeurs ne sont prêts. Nous examinons dans le cadre de ce projet la nature problématique des conseils générés par les ordinateurs, en nous penchant plus particulièrement sur les mécanismes (ou les lacunes en la matière) faisant qu'une décision produite par un ordinateur sera utile pour un décideur, en identifiant les faussetés répandues au sujet du principe de l'aide à la décision, en établissant des normes minimales que les systèmes d'aide à la décision devraient respecter pour être utiles, en soulignant l'importance cruciale de la confiance et en illustrant la manière dont les développeurs de systèmes peuvent intégrer des systèmes experts en appliquant des principes

rationnels de confiance. Les décideurs qui ont connu la frustration d'utiliser des aides à la décision fondées sur des systèmes-experts doivent pouvoir trouver confirmation de leur expérience et acquérir des connaissances leur permettant de faire valoir leurs besoins auprès des développeurs de systèmes, qui ne perçoivent pas les incidences des systèmes qu'ils proposent. De plus, le présent projet devrait amener les développeurs à rediriger leur travail vers des principes de conception d'aides à la décision plus réalistes et plus profitables.

Kessel, R.T. 2003 Le fardeau des conseils formulés par les ordinateurs : Usage des systèmes-experts comme aides à la décision. RDDC Atlantique TR 2003-241

This page intentionally left blank.

## Executive summary

---

### Introduction

The amount of information collected by military operation centres is continually increasing, threatening to overwhelm the capacity of decision makers to take account of it all.

Automation is required. Decision aids are often proposed in which a computer (an expert system), though not trusted to replace the human decision maker, nevertheless reports its decision as a form of advice for the human. Experience with such systems has proved to be unsatisfactory. The computer advice is seen more as an annoyance than an aid, and decision makers are inclined to switch the aid off or simply ignore it. The design and training of the expert system usually bears the blame for this, but a more fundamental fault is diagnosed here.

### Results

It is shown how computer advice imposes a burden of deliberation and ambiguity that decision makers would ordinarily never face in the absence of computer advice, deliberations about the relevance and meaning of the advice for the decision at hand, and about the reliability of the expert system. This cognitive burden more than offsets the intended utility of the aid. This diagnosis is reported here in non-technical terms that both decision makers and system developers should understand, while at the same time clarifying the paradigms, models, and mechanisms behind decision aids. It is concluded that decision-by-decision supervision by a human is an unproductive mode of integration for an expert system. Developers must discover more productive mechanisms of decision aiding, based on restricted and managed levels of trust and shared responsibility as described here, or they must finally produce expert systems that can be trusted implicitly, without reserve, to replace the decision maker.

### Significance

System operators who have known the frustration of using expert-system decision aids should find confirmation here of their own experience, as well as material for defence against system developers who do not appreciate the implications of the systems they propose. At the same time developers will be redirected by this analysis toward more realistic and profitable principles of decision-aids for use in military information and knowledge systems.

### Future plans

Solutions to the burden of computer advice are proposed. These will be considered by the Maritime Information & Knowledge Management (MIKM) Group at DRDC Atlantic for knowledge management in maritime intelligence, reconnaissance and surveillance (MISR).

Kessel, R.T. 2003. The burden of computer advice: Using expert systems as decision aids. DRDC Atlantic TR 2003-241.



## Sommaire

---

La quantité de renseignements recueillis par les centres opérationnels militaires croît continuellement, menaçant de submerger la capacité des décideurs à tenir compte de l'ensemble de l'information disponible. L'automatisation devient ainsi essentielle. On propose souvent le recours aux aides à la décision, des systèmes dans lesquels un ordinateur (système-expert), bien qu'on ne lui fasse pas suffisamment confiance pour remplacer le preneur de décision humain, formule toutefois des décisions sous forme de conseils à l'intention de personnes. L'expérience de ces systèmes s'est révélée insatisfaisante. On perçoit plutôt les conseils qu'ils communiquent comme une nuisance et les décideurs ont tendance à mettre ces aides hors service ou à simplement ne pas en tenir compte. La conception et la formation relative aux systèmes-experts sont habituellement blâmées pour cet état de fait, mais nous diagnostiquons une faille plus fondamentale.

### Résultats

Il est démontré que les conseils fournis par les ordinateurs imposent un fardeau de délibérations et d'ambiguïtés auquel les décideurs ne seraient normalement jamais confrontés en l'absence de conseils fournis par des ordinateurs : délibérations au sujet de la pertinence et de la signification des décisions fournies, ainsi qu'au sujet de la fiabilité du système-expert. Ce fardeau cognitif fait plus que neutraliser l'utilité projetée de l'aide. Notre diagnostic est ici formulé en termes non techniques que les décideurs et les développeurs de systèmes devraient comprendre et on en profite pour éclaircir les paradigmes, modèles et mécanismes sous-jacents aux aides à la décision. On en conclut que la supervision de chacune des étapes du processus de prise de décision constitue un modèle non productif d'intégration d'un système-expert. Les développeurs doivent découvrir des mécanismes de prise de décision plus productifs, basés sur les niveaux restreints et gérés de confiance et de partage de responsabilités décrits dans le présent rapport, ou ils doivent finalement produire des systèmes auxquels on peut faire confiance intégralement, sans réserve, pour remplacer le preneur de décision.

### Portée

Les décideurs qui ont connu la frustration d'utiliser des aides à la décision fondées sur des systèmes-experts doivent pouvoir trouver confirmation de leur expérience et acquérir des connaissances leur permettant de faire valoir leurs besoins auprès des développeurs de systèmes, qui ne perçoivent pas l'incidence des systèmes qu'ils proposent. De plus, le présent projet devrait amener les développeurs à rediriger leur travail vers des principes de conception d'aides à la décision plus réalistes et plus profitables, destinées aux systèmes d'information et de connaissances militaires.

### Plans futurs

Ce rapport propose des solutions au fardeau que représentent les conseils fournis par les ordinateurs. Elles seront discutées par le Groupe de gestion de l'information et du savoir maritimes (GISM) à RDDC Atlantique en vue de la gestion des connaissances en renseignement, reconnaissance et surveillance maritimes (MISR).

Kessel, R.T. 2003. The burden of computer advice: Using expert systems as decision aids. DRDC Atlantic TR 2003-241.

# Table of contents

---

Abstract.....	i
Executive summary.....	iv
Sommaire.....	v
Table of contents.....	vii
List of figures.....	ix
Acknowledgements.....	x
1. Introduction.....	1
1.1 Outline.....	3
2. Scope and background.....	5
2.1 Expert systems as decision aids.....	5
2.2 Background literature.....	6
2.3 Assumptions.....	6
2.4 Augmenting advice.....	8
3. System developer's perspective.....	9
3.1 The goal: better decisions.....	9
3.2 Minimum standards.....	10
3.3 Computer advice as human advice.....	11
3.4 Computer advice as decision fusion.....	12
3.5 The models predict failure.....	15
4. The decision maker's perspective.....	16
4.1 A thought experiment: in the decision maker's shoes.....	16
4.2 The burden of unfamiliar deliberations and new ambiguity.....	17
5. Dissecting computer advice.....	20
5.1 Automatic choice selection.....	20
5.2 Including a confidence indicator.....	21

5.3	When the confidence indicator is ostensibly a probability .....	22
5.3.1	The probability that the probability is right .....	23
5.3.2	Utility of probability .....	24
5.3.3	Non-classical decision making .....	26
6.	Realistic integration of expert systems .....	28
6.1	Trust management.....	28
6.2	Measure of last resort.....	29
6.3	Balancing responsibilities and abilities.....	30
7.	Summary and conclusions .....	32
8.	References.....	34
	Distribution list .....	40

## List of figures

---

- Figure 1.** a) In human relations, advice, if it is to be constructive, must be met with trust. It may then be followed by a transfer in a share of responsibility from the decision maker to the advisor. b) When the expert system is used as a decision aid, however, the exchange of trust and responsibility is deliberately blocked by the system developer's caution against trust, owing to ambiguity about the reliability of the expert system. Trust and responsibility rests solely with the human decision maker..... 12
- Figure 2.** a) Decision fusion is the combination of human and computer decisions into a single decision of better quality than either human or computer working independently. b) With computer advice, the decision fusion is carried out in the mind of the human decision maker, rather than by an independent agent..... 13
- Figure 3.** This Venn diagram depicts the domains of expertise of the decision maker (qualifying them to be responsible in the original decision space), and the developer of the expert system (qualifying them to propose and integrate an expert system as a decision aid), and then a third domain of expertise that the decision maker must have in order to meaningfully integrate the decisions of the expert system as advice of the decision at hand. .... 19

## **Acknowledgements**

---

Discussions with Dr. Nancy Allen at DRDC Atlantic were particularly helpful during this work. Her advice was clear and not burdensome.

# 1. Introduction

---

*Fools seek council from the ones they doubt.*  
—Heraclitus (500 b.c.)

*Whoever makes use of technology—and who does not?—entrusts himself to its functioning.*  
Hans-Georg Gadamer (1982)

There is a principle for decision aiding, often proposed, and sometimes put into operation, that nevertheless suffers from a flaw in its basic rationale, and therefore fails in practice. Its rationale is as follows. An expert system (computer) is designed to make decisions that a human currently makes, decisions that call for specialized expertise, entail uncertainty and difficulty, and are of significant consequence; decisions which are therefore prime candidates for computer assistance. The expert system is developed experimentally at first, to discover whether it could in fact *replace* the human decision maker, though most developers would admit from the outset that this is wishful thinking. In any case, the expert system shows promise during preliminary demonstrations, yet on the whole performance remains ambiguous. Either the demonstration and testing omits conditions that may realistically be encountered, or developers can see that the system falls short of desired performance for a subset of realistic conditions. The expert system cannot be trusted for autonomous operation, to replace a human decision maker, so it is offered instead as a decision aid. The adage “two heads are better than one” springs immediately to mind. That is, by reporting its decisions the expert system will naturally help the expert human arrive at better decisions. At least it cannot hurt matters, or so the reasoning goes, for in such critical decisions “every little bit helps”.

But beyond adages and a sense of urgency, system developers must follow sound principles of system design, which relate means to ends in a plausible chain of cause and effect. These cause-and-effect mechanisms, whether empirical, theoretical, or hypothetical, guide the developer through operational tradeoffs for best results, and justify the design of a system and its integration into the decision process. Yet it is precisely such mechanisms that are missing in the rationale for using an expert system as a decision aid. It has never been explained, for instance, how presenting an expert system’s automatic decisions to an expert human leads to better decisions, or at least produces the expectation of better decisions, particularly when the system is admittedly unreliable. The mechanisms behind computer advice, or the lack of

them, are the subject of this paper.

Human factors research has identified indirect mechanisms for decision improvement insofar as an automatic decision, when delivered under the right conditions and properly presented, can prompt the decision maker to be more alert or less biased by countering lapses of attention and drawing attention to possible oversights [Wickens et al. (2000); Kessel (2001); Johannsen et al. (1983)]. Thus a timely modicum of advice is automatically delivered by the expert system to produce effects in the stance of the decision maker that are in fact largely independent of the content of the advice. That is, a correlation between the content of the computer advice and the “best” decision is not required to produce the expectation of better decisions through these indirect mechanisms. Indeed, the delivery of the advice—its frequency, timing, and format—takes precedence over its content. In target detection, for instance, an occasional automatic alarm may be raised by an expert system to increase the vigilance of the human operator visually scanning signals for the signature of a target. Most of these alarms may be false, even obviously false to the human decision maker, but if delivered properly, they can nevertheless be effective. The veracity of the advice, though desirable of course, is in fact secondary to the goal of increased vigilance. As another example, the automatic suggestion of several decision options for consideration, in order to *de-bias* the decision maker, primarily reminds the decision maker that such options exist, while making no pretence of directing the decision maker’s deliberations toward one particular option or toward truthfulness.

But all parties concerned with decision aids—the developers, users, and purchasers—apparently expect that, beyond merely nudging the decision maker, the content of the expert system’s decision, properly considered, should enter into and sway the deliberations of the decision maker, simplifying or improving the decision in much the same way that expert advice from one human to another does. This is the primary role commonly envisioned for the expert system, much as in “two heads are better than one”. This advisory role is examined here, to expose its fallacies and propose solutions.

It is shown that, in practice, computer advice burdens the decision maker with unfamiliar deliberations and ambiguity falling outside their accepted domain of expertise—deliberations regarding the relevance and reliability of the automated advice within the context of the decision at hand, deliberations which system developers likewise cannot navigate as proven by their unconditional cautions against trusting the expert system during real-world operations. The unfamiliar deliberations and ambiguity constitute a significant cognitive burden that is immediately felt by decision makers, resulting in confusion regarding the role of the expert system and an aversion toward its advice. The source of the problem is difficult to identify because it is masked on the one hand by the deceptive clarity of the advice, and masked on the other hand by the prevailing fallacy that the decisions of an admittedly



unreliable expert system constitute advice. Thus, although dissatisfaction with operational computer advice has been widely recognized for some time, the diagnoses provided so far have missed its implicit source. Such is the thesis argued here.

Decision makers who have known the frustration of using expert-system decision aids should find confirmation of their own experience here, as well as material for defence against system developers who do not appreciate the implications of the systems they propose. More importantly, it is hoped that developers will be redirected through this discussion toward more realistic and profitable principles of decision-aid design.

For brevity, *developer* refers throughout to the developers of expert systems and/or decision aids, and *decision maker* refers the human decision maker, never the expert system.

## 1.1 Outline

Section 2 delimits the scope, assumptions, and context for the discussion.

Section 3 adopts a system developer's perspective of both computer advice and the decision maker. Good decisions and good decision aiding are first defined, and then used to set minimum standards by which decision aids in general can be evaluated. Mechanisms for the action of computer advice are drawn from complimentary paradigms that developers might naturally adopt to conceptualize and speak about the operation of computer advice: 1) a human-human model for the exchange of advice (anthropological paradigm), and 2) an instrumental model of automated advice as decision fusion (instrumental). It is shown that both predict the failure of computer advice.

Section 4 takes up on the decision maker's perspective, using a thought experiment to place the reader in the decision maker's shoes. It is shown that computer advice, if it is attended to at all, imposes a substantial cognitive burden of unfamiliar deliberations and ambiguity that system developers likewise cannot navigate.

Section 5 dissects computer advice to unveil the ambiguities beneath its deceptive simplicity, which pose the biggest problem when integrating expert systems as decision aids. The confidence indicator included in computer advice is given particular attention. Though intended to reduce ambiguity by qualifying the reliability of the advice, it imposes yet another burden of unfamiliar deliberation. This is true even when confidence is ostensibly expressed in terms of probabilities.

Section 6 proposes remedies to the problem of computer advice, arguing first of all for the

important role of trust in the expert system, and then showing by examples how trust can be managed so that an imperfect expert system may be turned to good use. The remedies are by no means complete or exhaustive, but they illustrate what purposeful integration of imperfect expert systems may look like in practice. Neither are the remedies especially new, but they have so far been viewed as secondary or specialized ways of integrating expert systems, whereas, in view of the present critique of computer advice, they are in fact representative of the ways in which an admittedly imperfect expert system must be integrated if it is to serve any purpose.

Section 7 summarises the discussion, and gives reasons why the fallacies of computer advice have not prevented introduction of decision aids into operational systems, despite the contrary evidence: the frustration of decision makers, the lack of success stories for operational systems, and the cautionary work of many researchers.

## 2. Scope and background

---

Advice, decision aids, and expert systems touch on many topics ranging from human psychology, to computer intelligence, to system design. The bounded scope, context, assumptions, and aim of the present discussion are outlined below.

### 2.1 Expert systems as decision aids

It must be emphasized that the present critique concerns the integration of expert systems as decision aids, when the expert system's decision is reported to a responsible human decision maker in order to improve the quality of the decision. It does not apply more generally to expert systems in their own right, when they function autonomously, without decision-by-decision supervision. The two roles for expert systems are very different. *A decision aid does not fail when the expert system makes an error in judgment*, for instance. Developers warned that such errors are likely to happen, which is why the expert system was subordinated to the human decision maker in the first place. Errors should be no cause for surprise. Rather, *the decision aid fails when the decision maker feels inclined to switch the expert system off, or when it biases the decision maker against good decisions rather than toward them*. In other words, when integrating an expert system as a decision aid, the challenge to developers is not to improve the expert system's performance to the point of replacing the human (for the replacement would then presumably be made and the decision aid be foregone), but the challenge is to improve the decision maker's performance using an expert system that is expected to make mistakes. This challenge is the motivation for the present work.

What is said here about decision aids therefore does not apply to expert systems when they are given autonomous control, as done to good effect in autopilots or process control for instance. The problem of decision aiding vanishes the instant one passes control to an expert system. If an expert human (pilot or engineer, say) supervises the autonomous expert system, then it is by monitoring the effects of its automatic decisions (the disposition of the aircraft, process alarms or indicators) rather than by monitoring individual automatic decisions, of which there may be thousands made at high speed in a real-time feedback loop (the continual adjustment of trim, rudder, and so forth). To intervene, moreover, does not mean overruling a particular decision, but taking control back from the expert system to set matters right, at which point control may be transferred back again to automatic. This human-machine interaction of control transfer [Chu & Rouse (1979)] is quite different than the advisory role of the decision aid considered here, where the expert system reports its automatic decisions as an advisor—without control or responsibility—to the decision maker for consideration.

## 2.2 Background literature

The shortcomings of expert systems as decision aids have been treated by others, touching on many matters such as poor expert system performance, poor human-machine interface design, lack of understanding by developers of the decision maker's task, and inflated expectations [Chalmers et al (2000); Hopple (1986); Keyes (1989); Norman (1990); Pomranky et al. (2001); Skitka et al. (1999); Sloane (1991); Suh & Suh (1993); Waern et al. (1995); Woods (1985, 1986, 1988); Wooldridge & Jennings (1998)]. But the common element passing without challenge through virtually all critiques is the nature of the computer advice delivered to the decision maker. The advice typically consists, as we have said, of the expert system's decision, including a choice selection and a confidence indicator. At first glance the computer advice is clear and concise, with little room for discussion, which would be true so long as expert systems are treated as if they will one day replace human decision makers. But it is when the rules of the game are changed, when expert systems are integrated as a subordinate decision aid, not as a human substitute, that confusion suddenly arises. Human elements of advice, trust, and responsibility, once safely omitted, now become paramount.

Developers have been hesitant to address these human dimensions, though researchers in human factors have noted their importance for some time [Chu & Rouse (1979); Lee & Morray (1992); Muir (1987); Muir & Morray (1996); Waern et al. (1995)]. Closest to the present work is that of Woods (1996) and Parasuraman (1997) who consider automation in all forms, not just computers. They point out (as shown here for computer advice) that automation often creates new, unforeseen burdens and complexities that counter the potential benefit of automation, possibly overwhelming its benefit entirely. Hence they advise more care for the human dimensions when integrating automation operationally. A similar theme is taken up here. Where they looked at automation very broadly, the focus here is on mechanisms of computer advice in particular.

## 2.3 Assumptions

For our purposes it does not matter what type of expert system generates the advice, whether neural networks, pattern recognition, Bayesian networks, a support vector machine, etc.. It is the nature of the advice that is important. The expert system is therefore treated here as a black box, much as the decision maker would typically view it in practice. Indeed, it is usually too much to ask for a decision maker to wear two hats at once: to be expert in their primary domain of responsibility, and then again in the developer's domain specializing the inner workings of an expert system.

There are of course many factors, cognitive and operational, that affect a person's ability to make decisions, and hence affect the role of computer advice. A novice attends to advice more closely than an expert, for instance, much as an expert will attend to advice more in unfamiliar conditions than in familiar [Silverman (1992)]. The timing of advice, too, whether delivered before, alongside, or following other critical decision information, may affect its role. Where the importance of such factors has been recognized, developers have accommodated them by making provisions for on-site, fine-tuning adjustments of the decision-aid system. These may be sensitivity adjustments whose net affect is to change the system's error rate while the nature of the advice remains unchanged. It is therefore reasonable to simplify matters here, if only because this is the way expert systems are used in practice, by likewise treating so many decision-affecting factors as if their net effect were a matter of degree only, changing the inclination of the decision maker to attend to the computer advice without changing the essential nature of the advice or the mechanisms by which it aids the decision process. This is a necessary first order approximation to a very complex human dynamic.

Finally, it need hardly be said that it is difficult to advise an expert in matters about which they are certain. The advice is perceived as duplication or error. Advice from an admittedly unreliable subordinate is furthermore tolerated as a test of skill, in order to pass judgment on the subordinate more so than to improve one's own deliberations. To accept advice, rather, the decision maker must genuinely question his or her ability in some way, or suspect that the subordinate advisor is privy to hidden information. One or both is required for advice to play any role at all.

In summary, then, there are five operating principles that characterize the broad class of computer advice considered here:

- 1) the expert system, for whatever reason, is not trusted to replace the expert human;
- 2) the expert system is nevertheless invited into the decision-making process as a decision aid, with the expert human ultimately holding responsibility for the decision;
- 3) this aid amounts to a report of the expert system's automatic decision, consisting of the expert system's choice drawn from a list of allowable options plus an indication of confidence;
- 4) the expert human attends to this advice out of a sense of genuine need;
- 5) the goal of a decision aid is to increase the expectation of good decisions or reduce cognitive workload.

In the course of this discussion it will quickly become evident that item (1), the lack of trust owing to the cautionary stance advised by developers, is the primary source of the ambiguity and confusion surrounding the operation of computer advice, and that it must therefore be

qualified or eliminated in remedial measures. The remedies proposed here (Section 6) restrict or modify operating principle (1) in one way or another.

## **2.4 Augmenting advice**

It may be possible to augment computer advice beyond the mere choice selection and confidence indicator assumed here. Human advisors typically give an account of their reasons for the advice they give, telling of prior information, experience, or plausible story lines, in this way giving shape and direction to the deliberations of the responsible decision maker. The same is possible to a some extent using expert systems like Bayesian networks, belief networks, and decision trees, in which the expert system can be probed for its reasons [Pollack et al. (1982); Waern et al. (1995)]. The grammar-checking algorithm in the word processor used to prepare this paper is a good example of this augmented advice. It elaborates suspected grammatical violations by stating the rule at issue, and gives examples of fixes. Unfortunately, a similarly fluid dialogue with an expert system regarding course-of-action or command-style decisions calls for an esoteric language that is peculiar to the type of expert system, which, if it does not force deliberations into unfamiliar modes of rationality, demands an expertise and proficiency with the expert system that decision makers would not ordinarily have, and which to acquire and exercise would burden the decision maker considerably. Many criticisms made here about computer advice also apply to elements in the dialogue with the expert system, thus the potential for improved decision aids through augmented advice brings with it a counter potential for added confusion and ambiguity. Only one-way, two-component advice without augmentation (as in item (3) above) is considered in this paper.

### 3. System developer's perspective

---

It is remarkable that developers have apparently never related means to ends in a plausible chain of cause and effect for the decision aids they have been developing. How is it that the content of an admittedly unreliable expert system's decision results in better human decisions? Mechanisms of operation are drawn here from two complimentary paradigms that developers naturally adopt for computer advice when visualizing and speaking about its operation—one anthropomorphic (Section 3.3), and the other instrumental (Section 3.4). The paradigms are not new, but they are pressed now to demonstrate the plausibility (or implausibility) of computer advice given that one paradigm or the other applies. To this end it is necessary to first give definitions and criteria by which decision-aid mechanisms can be judged as operational successes or failures.

#### 3.1 The goal: better decisions

Following Langer (1994) and others [Pierce (1878); Dewey (1938); Levi (1997)], a decision may be defined as a cognitive commitment to act or respond. By cognitive commitment is meant that information gathering and rational deliberation about options cease (or find their resting place) in a single option. The commitment may be the satisfying result of lengthy, careful deliberation, or the unsatisfying result under pressure of time and uncertainty. In any case the commitment marks the transition from thought to action<sup>1</sup>. The stability of the commitment depends on quality of the information, subsequent contradictory information, and the psychological stability of the decision maker. An *expert* decision, moreover, is a decision whose information gathering and deliberations require specialized knowledge, experience, or discipline, which to have entitles one to bear authority and responsibility for making such decisions.

Regarding good decisions, it must be said that, in the moment it is reached, every decision naturally appears to be good to the person making it, otherwise they would have made another. The decision maker may not feel entirely satisfied, and may qualify their decision with many cautions, but to their mind it is the best decision under the circumstances, at least until information to the contrary is received. The decision maker therefore has an internal subjective standard of goodness. A more objective definition of what constitutes a good

---

<sup>1</sup> Poincare (1952; Chapt.XI) points out that decisions made on the basis of doubtful information call for a degree of willfulness that is analogous to faith or trust: "We must therefore make up our minds without knowing. This must be done whatever may happen, and we must follow the risks, although we may have little confidence in them. What I know is, not that such a thing is true, but that the best course is for me to act as if it were true."

decision is desirable, but very difficult to construct. Indeed, Langer (1994) opts instead to define the *wrong* decision rather than a good decision, but in a way that leads again to subjectivity: “A wrong decision may be when one is untrue to one’s own cognitive commitments.” Although subjective, this is a crucial minimum standard that all decision aids must meet. That is to say, a cause-and-effect mechanism of utility for decision aids cannot require decision makers to act against their better judgment, or force them to consider matters falling outside their proper domain of expertise, for to do so would be to force decision makers to act in a way that appears irresponsible to them.

This may sound obvious, but it bears mentioning because this is precisely what would happen, for instance, if the expert system had access to information (as developers are often quick to point out), from another sensor, historic data base, or mathematical computation that was hidden from the decision maker. For in cases of disagreement between human and computer, it would be irresponsible or irrational for the decision maker to change his or her mind in favour of the expert system, simply on the faith that it has exploited hidden information, particularly since developers have cautioned against this precisely this kind of blind faith. Thus what ought to be the strength of the expert system—seeing what the human decision maker cannot see—is of no avail because the mechanism behind the operation of the decision aid violates the minimum standard: it requires the decision maker to act irresponsibly, trusting the system when developers have expressly cautioned against trust, on the purely speculative grounds that the expert system “knows” things that remain hidden from the decision maker. This dilemma sets the stage for the present analysis, but it is not the full story as we shall see.

An objective definition of a good decision is simpler with expert systems than with real-world decisions generally because the expert systems’ allowable options are limited, being specified in advance rather than generated fresh for each decision, making it is possible then to speak narrowly of good in the classical sense using the calculus of expected costs [von Neumann & Morgenstern (1944)]. That is, a good decision maximizes benefits (minimizes costs) where benefits (costs) are given a numerical value but may broadly include advantage (disadvantage) of any kind. The calculus of minimum expected costs is in fact the basis for the design and training for most expert systems [Hastie et al. (2001); Ripley (1996); Duda et al. (2001)]. For our purposes, then, improving the quality of decisions therefore means increasing the benefits (reducing the costs) of operation through the action of the decision aid on the decision maker.

## **3.2 Minimum standards**

Turning now to decision aids, the objective and subjective standards of goodness given above imply that a decision aid should produce the objective expectation of greater benefit (lower costs and risks) of operation, without compromising a decision’s legitimacy in the mind of the



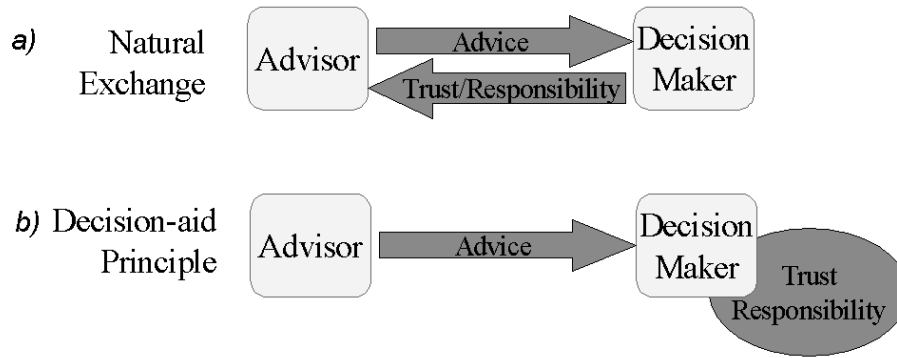
decision maker. A decision aid that fails in either respect will be an operational failure, being either of no utility or jeopardizing the mission. These are the standards used to critique computer advice below.

### **3.3 Computer advice as human advice**

The human-human exchange of trust and advice is the most widely used paradigm for thinking about the operation of computer advice. The “advice” in its name “computer advice” is proof of this. In human relations, advice is rejected wherever it is deemed untrustworthy. Trust is therefore a precondition for advice to be constructive. An inclination to consider the advice of a colleague, for instance, is proof of one’s trust in his or her judgment. This human exchange of trust and advice is shown schematically in Fig. (1a). Responsibility for a decision is also shared to some extent with the advisor, though perhaps morally more so than legally.

With computer advice, however, the natural exchange of trust, responsibility and advice among humans faces a deliberate partial blockage. Advice is given by the expert system, but developers caution the decision maker against placing trust in it, while insisting that responsibility rests solely with the human decision maker, much as shown in Fig. (1b). Perhaps advice can work this way, but it would be an unnatural, if not inhuman mode of operation. Developers therefore cannot simply draw an analogy in their mind between computer and human advice without further explaining how the substantial difference in the exchange of trust is inconsequential in practice.

The need for such explanations has not been recognized, nor have any been forthcoming. But until one is discovered, the success of computer advice as it is implemented in practice (paradigm (b) in the figure) remains no more than wishful thinking. On the other hand, if developers intend to capitalize on the strength of a truly human model of advice (paradigm (a) in the figure), then computer advice fails in contradiction: trust being a precondition for constructive advice, and yet an attitude that developers cannot in good conscience recommend. Like most operational contradictions, it results in confusion: decision makers do not know what to make of the “advice”, whether to trust it or ignore it, the latter being the more rational and responsible way out. Whether following paradigm (a) or (b), then, computer advice fails the minimum standards set above, neither producing an expectation for better decisions, nor making sense to the rational decision maker.



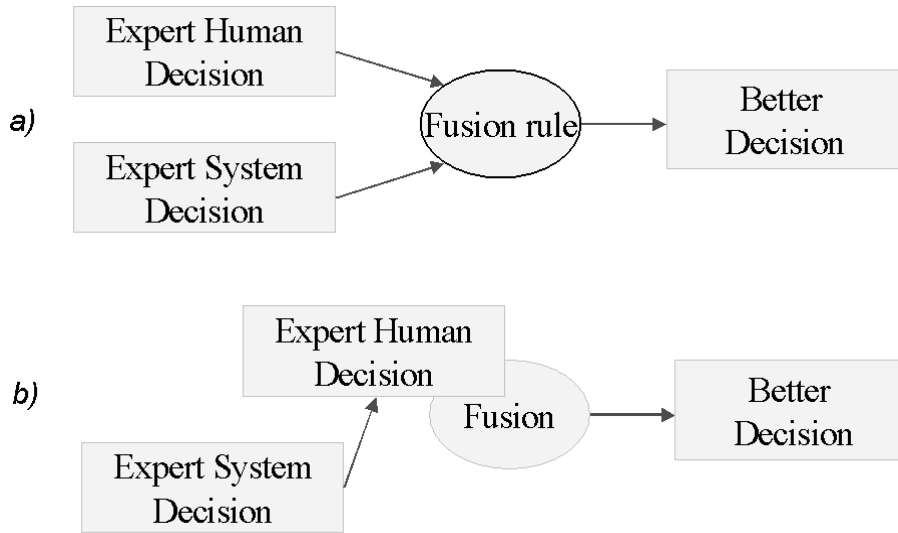
**Figure 1. a)** In human relations, advice, if it is to be constructive, must be met with trust. It may then be followed by a transfer in a share of responsibility from the decision maker to the advisor. **b)** When the expert system is used as a decision aid, however, the exchange of trust and responsibility is deliberately blocked by the system developer's caution against trust, owing to ambiguity about the reliability of the expert system. Trust and responsibility rests solely with the human decision maker.

### 3.4 Computer advice as decision fusion

Some developers prefer to set the notions of trust, responsibility, and advice aside, claiming that the expert system merely provides its decision as information for the decision maker to consider, in much the same way as any sensor provides information for consideration. The expert system functions then as an *information transducer* of sorts, transforming large amounts of diverse information into a single nugget, the automatic decision, serving as one information cue among many that are relevant to the decision, rather than as advice proper. It is still possible to speak of trust in this case, but it is a different kind of trust that admits gradations and doubt, like the trust placed in sensory perception, vision for instance, rather than the trust placed in advice or an advisor. Thus we trust our vision, but we do not trust it entirely; we know that it can be tricked by speed, illusion, and lighting, yet make good use of vision all the same. The expert system may likewise be trusted (or doubted), as a sensor merely, rather than as an advisor.

This *instrumental* view of the expert system as information transducer amounts to viewing computer advice as an instance of *decision fusion*. Decision fusion is the combination of the human's decision and the expert system's for a result better than each operating independently, as shown schematically in Fig. (2a). The connection between computer advice and decision fusion is made in the following way. In principle, a decision aid brings a computer's decision and a human's decision together for combination into a single decision. However this may be accomplished, we insist that it be done in a way that, as a rule, raises the quality of the combined decision above the quality of either agent operating independently (otherwise one would do just that, use only the better agent). Thus, in principle at least, the

goal of computer advice is precisely the same as decision fusion: to bring multiple decisions together to produce a better decision. Since decision fusion entails well-defined mathematical techniques for optimizing the quality of decisions, moreover, it sets the upper bounds on the best performance that we can hope to achieve for any given class of fusion rules, and it characterizes the properties of improved decision performance generally requires.



**Figure 2.** a) Decision fusion is the combination of human and computer decisions into a single decision of better quality than either human or computer working independently. b) With computer advice, the decision fusion is carried out in the mind of the human decision maker, rather than by an independent agent.

There are established techniques for fusing decisions from multiple agents according to optimized, mechanistic (mathematical) rules [Dasarathy (1994); (1998); (2003)]. These may be Boolean (AND/OR) operations applied to the candidate decisions, or weighted-voting schemes specially derived from knowledge (expressed as probabilities) of the decision performance of decision-contributing agents, in each case rendering a single decision from several. These techniques show that, in addition to well-designed fusion rules, improved performance requires that the agents providing candidate decisions adjust their decision-making inclinations (sensitivity settings or decision thresholds) in a narrowly prescribed fashion that depends on the expected performance of the other agents offering candidate decisions. In other words, across each agent's full range of independent inclinations, the enhancement of decision fusion only occurs within a narrow window of inclinations that is unlikely to be hit by chance, if agents operate without purposeful orchestration. Hence fusing decisions in an ad hoc or biased manner is likely to result in worse overall decision performance rather than better.

So long as we do not specify the class of fusion rules for decision aiding, the connection between the decision aid and decision fusion remains a connection in principle only, with decision fusion serving as a conceptual paradigm for envisioning its operation, and, most importantly, in this way qualifying our expectations for the operation of computer advice. On the other hand, if we are prepared to prescribe the use of a particular class of fusion rules, then the paradigm can be pressed much further, to the point of designing an optimal decision aid if only the performance of both human and expert system are known and stated probabilistically. Much the same kind of connection with decision fusion has been used elsewhere for human-machine interaction, in the area of psychophysics, for the analysis and evaluation of human-monitored alarm systems<sup>2</sup> [Sorkin and Woods (1985); Green and Swets (1989); Kessel (2001)].

In either case, the decision-fusion paradigm enlightens the operation of computer advice considerably. The developers' caution against trust, for instance, implies a makeshift fusion rule of sorts—not to trust the expert system—which is operationally equivalent in operation to a logical identity: the fused decision equals the human decision. This is clearly an unproductive fusion rule, leaving the quality of the fused decision unchanged from that of the decision maker operating without the expert system, yet it precisely captures the developer's caution against trust.

Under the decision-fusion paradigm, then, the first hurdle that developers face when integrating an expert system as a decision aid is to develop a productive fusion rule. We can expect that any ad hoc approach to fusion, if strictly enforced, is more likely to reduce decision quality than to help it. For improved performance, both the computer and human must regulate their decision-making inclinations in a narrowly prescribed fashion that depends on each other's decision performance, and this is impossible to do so long as the performance of human and expert system remain unknown. Note that if optimal fusion rules could be derived, then they risk being perceived as giving too much responsibility to the expert system, or as an affront to the decision maker's expertise—the mechanistic rules for settling human-machine disagreements appearing artificial, bureaucratic, or as a limited mode of rationality amid the rich context of real-world decisions, in violation of our standard that the decisions reached must look good to the decision maker. The rule must furthermore be simple if it is to be implemented in the mind of the decision maker without burdening them.

---

<sup>2</sup> Human-monitored alarm systems are often analyzed using a more general form of *feature fusion* or equivalent signal processing, of which decision fusion is a special case. In feature fusion a confidence indicator (or an equivalent decision variable) from each agent are fused rather than the decisions of each agent. The benefit of feature fusion versus decision fusion is a current topic of research. The distinction between the two types is not important here. The point here is that signal-processing styled fusion of one kind or another can be applied to the analysis of human-machine systems like decision aids, with the complexities of each agents' internal workings modeled probabilistically (feasible for humans) rather than deterministically (infeasible).

Finally, it is clear that the schematic in Fig. (2a) should be changed slightly, though in an important way, as shown in Fig. (2b), to reflect the fact that the decision fusion is carried out in the mind of the human decision maker, rather than by an independent third party as ordinarily assumed in decision fusion. This does not break with the decision-fusion paradigm because decisions are still being combined, though combined in more complicated ways than previously imagined. Indeed, if the computer advice influences the outcome of decisions at all, then it will be through the bias of agreement or disagreement with the human's decision, rather than through optimal fusion rules. The bias is perfectly natural and acceptable if the human holds all of the responsibility and the computer holds none. Being ad hoc from a decision fusion perspective, however, the bias makes the prospect of better decisions unlikely. If we allow that the decision maker might be instructed against bias, then the developer faces the difficulty once again of deriving and instituting a purposeful fusion rule.

Using the decision-fusion paradigm, then, we have several mechanisms accounting for the difficulty of computer advice, if not its failure as well, but no positive mechanism for its success.

### **3.5 The models predict failure**

We have seen that computer advice, modeled as advice between humans, naturally requires trust as the precondition for constructive advice in contradiction to the developer's unconditional caution against trust, resulting in confusion about the purpose and use of the decision aid. On the other hand, modeled instrumentally, computer advice becomes ad hoc decision fusion with little expectation of better decisions. If other models of computer advice can be found, then they must ultimately take account of the two paradigms given here, either overcoming the problems advice identified in each, in the terms of each, or else functioning in ways that these two models do not apply—in ways, that is to say, that cannot be construed as human advice or decision fusion.

## 4. The decision maker's perspective

---

The challenge that military decision makers like to offer to developers is that the developers ought to come out into the field, to use the decision aid in the real world and discover its frustration first hand. Usually this is understood by developers to mean that the expert system has been making mistakes—something to be expected, but which they hoped would not be too annoying. But the decision makers' frustration can be traced to a deeper conceptual flaw with computer advice, a flaw by which computer advice will continue to fail until expert systems finally advance to their remote ideal of replacing the human decision maker. The flaw is traced here to the nature of the computer advice itself, to an insurmountable ambiguity about its meaning and relevance that is masked by its deceptive clarity. We begin with a thought experiment based on all of the assumptions (1) to (5) near the end of Section 2.3.

### 4.1 A thought experiment: in the decision maker's shoes

Let us imagine that personal circumstances call for the following decision: one's car is in need of repair. After researching options, one feels uncertain whether to fix the car or to buy a new one. If the car is repaired, the repairs may be followed later by still more repairs and inconvenience due in part to its age, but this option may still be more economical than the greater, though more predictable cost of buying a new car, as one must presumably do eventually in any case sometime in the future. Repairs postpone the expenditure of a purchase and they may permit a better vacation next summer, for instance. I assume that the reader wants to be rational (act for greatest expected benefit) in the decision to fix or buy, and that the reader feels responsible for personal decisions of this kind.

Now enters the developer with an expert system designed specially for such decisions. The developer confidently demonstrates the system using realistic examples and stories with details similar to those above, but cautions that the system may nevertheless be unreliable, "The system aids your decision; it does not replace you." From a list of two options, then, either to "fix" or "buy", the expert system, apprised of the situation, reports its decision: "buy" a new car with "medium" confidence.

In the decision maker's shoes, what are we to make of such advice? Does one feel a sense of relief, having been nudged closer toward settlement? Or does one feel a twinge of ambiguity because the advice invokes uncertainty about its interpretation, integration, or reliability? What impetus does the advice impart to deliberations about the course of action? Does it legitimately push one toward cognitive commitment, or does it launch one instead toward a new domain of questioning, regarding the trustworthiness of the expert system for instance?

These are admittedly leading questions, but they are intended to spur close observation of what by extension may be typical more generally for decision makers during the moments of decision when faced with computer advice. These are the kinds of questions that should be addressed by a guiding theory of decision aids, for it is in the enquiring mind of the decision maker that the advice exerts its force, either for better or worse.

At first glance the advice is clear: “buy” a new car, “medium” confidence. What more could one ask? But on second thought one discovers how stubbornly opaque it actually is. How should the automated information enter into deliberations about the best course of action? The direction of the advice is clear, but what weight is to be given to it? What relation does it have to the other facts in the case? What inferences can be drawn from the advice? In the decision maker’s place, some readers may ask for more detail about the expert system, about its demonstration and track record, as if the information in the thought experiment were inadequate, or what really amounts to the same thing, as if the computer advice, despite its clarity, lacked a certain quality that would make it useful. All of these questions point to an inadequacy in the advice, and to the unsettling of cognitive commitment rather than its settlement. Unsettling may not be a bad thing, but here again, developers must show in what way it is good if they are relying on it as the primary decision aid mechanism.

Judging from my own experience and that related by colleagues, both from thought experiments like this and when discussing prospective decision aids for military applications<sup>3</sup>, I find that most people will find that they do not know how to fit the computer advice into the puzzle to be solved, not because they are inexperienced in the decision space<sup>4</sup>, but because they frankly do not know what to make of the advice. More information or more knowledge about the performance of the expert system, about its credentials or the meaning of its advice, is required. Thus the computer advice spawns new ambiguities rather than settles those faced in the original decision. It adds a new dimension to the cognitive workload rather than reducing it.

## **4.2 The burden of unfamiliar deliberations and new ambiguity**

The added burden is depicted schematically in Fig. (3). The boundary of the decision makers’

---

<sup>3</sup> The application in this case was the integration of automatic algorithms to assist human operators in target detection and classification tasks for sea minehunting as implemented in Canada’s Remote Minehunting System (RMS) Technology Demonstration Project (TDP) [MacDonald, Dettwiler and Associates (2003)], and more speculatively for ship detection and classification for space-based surveillance and reconnaissance using synthetic aperture radar (SAR).

<sup>4</sup> By decision space or decision domain is meant the context surrounding the decision to be reached by the decision maker: the benefits, costs, and risks influenced, the information available, and the expertise required.

expertise signifies their accepted limits before the use of any decision aid. This is the expertise that qualifies decision makers to be authoritative and responsible in the first place. The boundary of the system developers' expertise, on the other hand, signifies their rather different accepted limits for the integration and testing of the expert system as decision aid for the decision maker. Overlap with the decision makers' domain occurs because developers would ordinarily acquire some of the decision makers' expertise in the course of developing the expert system. The boundary of the third domain signifies the limits of the new expertise that one requires to translate the computer advice into the context of the decision at hand, to interpret it as a telling cue that legitimately weighs into and sways the course of deliberations about the best course of action, an expertise that is necessary to purposefully translate the computer advice into the decision space. This domain overlaps the others somewhat, but covers significant new ground as well, which neither decision makers nor developers have—decision makers because they are not accustomed to expert systems, and developers by their own admission inasmuch as the expert system's performance is admittedly unknown, or known sometimes to fail. The decision makers' domain of expertise must be expanded into that new domain if computer advice is to fulfill a legitimate advisory role. This new ground represents the burden of computer advice.

Perhaps no one will deny that decision makers need additional training of some kind when a decision aid is first introduced, to expand their expertise in the direction of that required to make use of the computer advice. All readily admit, that is, that despite its clarity the computer advice must be augmented in some way, by context or understanding, for it to be useful. That training presumably includes explicit instructions from developers, regarding the attitudes and responses (fusion rules) that decision makers should adopt when the expert system disagrees about the best decision. When integrating an expert system, the goal of training seems on the one hand to be to demonstrate the expert system in a positive light, by showing that it has some ability in decision making, while on the other hand conveying its limitations as well, in order to instill a healthy skepticism<sup>5</sup>. Training may also include trials specially staged for developing tacit knowledge through experience<sup>6</sup>. This knowledge, its source, and the effort to acquire it, must all be considered in the developer's integration plan for the expert system as a decision aid. Such a plan is rarely spelled out in practice because complete knowledge regarding system performance and reliability is in fact nowhere to be found. Decision makers look to developers, and developers point in turn to the need for more

---

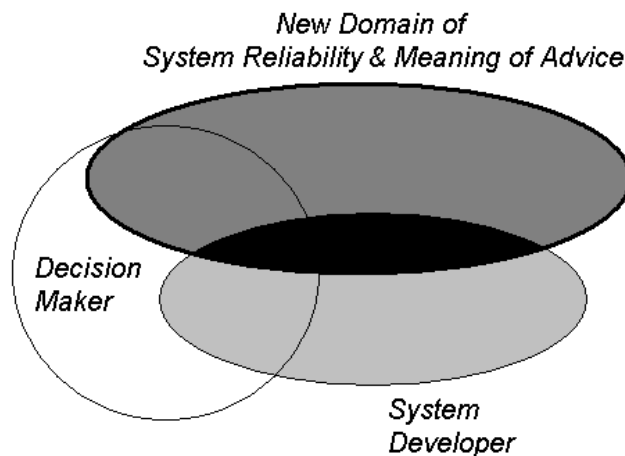
<sup>5</sup> Healthy skepticism is taught by explaining the origins of the computer advice in the inner workings of the expert system, which is the deconstruction of the advice to its mechanistic essence. This is a point to which we return in Sections 5.1 and 5.2 in connection with the meaning of the advice.

<sup>6</sup> Following knowledge management, it is usual to distinguish between *explicit* knowledge, which can be captured and communicated in words or symbols, and *tacit* knowledge, which can only be communicated by practice and experience (piano playing, for instance) [Polanyi (1962)]. The distinction is useful when discussing cause-and-effect mechanisms of decision aiding and training because it emphasizes that these can work in radically different but equally valid dimensions, through classroom instruction and hands-on training. The terms are used in this context here.



development and testing, but without indicating how much exactly is required, while merely cautioning against reliance on the expert system in the meantime.

The strength or weakness of computer advice becomes most evident in the instructions (fusion rules) given to the decision maker for handling decisions when the computer's automatic decision disagrees with the decision maker's. On what grounds should decision makers be swayed by the expert system to change their mind? An unconditional prohibition against trusting the expert system or sharing responsibility means that decision makers should never give way in cases of disagreement unless there is some meaningful information tacitly gleaned from the computer's advice, going beyond the bare report of its decision, by which the deliberations of a rational decision maker can be legitimately swayed. That prospect is examined in the next section.



**Figure 3.** This Venn diagram depicts the domains of expertise of the decision maker (qualifying them to be responsible in the original decision space), and the developer of the expert system (qualifying them to propose and integrate an expert system as a decision aid), and then a third domain of expertise that the decision maker must have in order to meaningfully integrate the decisions of the expert system as advice of the decision at hand.

## 5. Dissecting computer advice

---

In this section the flaw in computer advice is traced back to its components, to the automatic choice selection, and to the confidence indicator assigned to that choice. Confidence, in particular, is seen by developers as a way to qualify the reliability of the advice, especially when it is ostensibly an objective probability. Confidence reports are therefore given close attention here. The perspective assumed is again the decision maker's, though for the sake of argument it is allowed that the decision maker may sometimes have a developer's expert knowledge of the inner workings of the expert system.

### 5.1 Automatic choice selection

Focusing first on the choice selection alone, which is in fact the sole component of advice for many expert systems, there can be no doubt about the direction of the computer's decision: it points directly toward the option named. What remains questionable is the role that this decision vector legitimately plays within the decision maker's deliberations.

Taken at face value, the unadorned automated choice selection no more constitutes advice than a mystical oracle's bare directive. As Polanyi (1962; pp.28) points out, "An unasserted sentence is no better than an unsigned check; just paper and ink without power or meaning." The advice has a detached, unreal quality. It calls for trust or distrust, not for expert deliberation. The automatic decision may have been expertly derived from hidden information, or, what is more likely in complex decisions, it may have been drawn from mistaken perceptions, overlooked information, or simplistic heuristics. Decision makers simply do not know. They must instead make a willful resolution either to trust or to distrust the computer advice, just as those who consult a mystical oracle must do when they do not have contextual information about the character, reliability, and interests of the advisor.

On the other hand, for decision makers who have a developer's intimate knowledge of the inner workings of the expert system, the report of the expert system's decision may be deconstructed into its mechanistic essence. That is, the automatic decision may be explained by the values of certain numerical variables (possibly the confidence indicator itself) crossing predetermined thresholds, perhaps with one variable excelling others in a computational game of steeple chase between variables representing each allowable option, with final choice selection falling to the winner. The process (training) by which the expert system's sensitivities and thresholds have been set and its capacity to generalize to previously unseen conditions then becomes paramount for interpreting the significance of the automatic choice. So here ambiguity surfaces again, but in new form. Has the training of the expert system been

adequate? Does it generalize well to the present situation? Much the same ambiguity exists for both for the developer and the decision maker.

There are methods for addressing these questions of system training [Hastie et al. (2001)], but none is ultimately conclusive. More to the purpose here, they vault the decision maker into deliberations that take place on a still higher level of abstraction, within the domain of machine learning, probability, and statistics. Hence computer advice drives decision makers from their original decision space where their proper responsibilities lie, to matters regarding the inner workings of the expert system, and then further to advanced questions of machine learning. Or viewed another way, the ambiguity about computer advice is pushed to higher levels of abstraction without being eliminated, to the limits of one's reach, like a bubble under freshly hung wallpaper. The ambiguity may even be increasing along the way.

## **5.2 Including a confidence indicator**

Where developers have felt the burden of computer advice, they have typically included a confidence indicator as if to automatically answer questions like "What is the probability that the computer is right?" It may simply be a "low", "medium", or "high" indicator, or it may be a numerical value, possibly on a scale between 0 and 1, sometimes ostensibly a probability.

As with humans, reports of confidence tend to be subjective: more meaningful to the agent reporting confidence than to others receiving the report. But unlike a human's report of confidence, an expert system's derives from well-defined numerical algorithms whose steps can be precisely traced, at least by system developers. It may not be unrealistic for decision makers to be taught essentially the same "bottom-up" understanding of the confidence indicator, and it makes sense to do so because knowledge of its computational origin demystifies the confidence indicator to some extent. Thus, much as with the automatic choice selection, "high" confidence simply means that a particular internal decision variable of the expert system, derived from features of the input data, has crossed the threshold signifying "high", for instance. But here again, the deconstructed confidence indicator misses a much more important "top-down" understanding of its significance in light of the particular decision at hand. If the confidence indicator is to solicit a rational response, then it must be interpreted or calibrated in the decision maker's mind in light of the costs, risks and information cues pertaining to the decision it qualifies, which is something that a merely etiological explanation cannot give.

To illustrate the difficulty of interpretation, let us return to our thought experiment regarding the repair or purchase of a car in Section 4.1. Imagine now that the expert system reports its confidence using a number ranging from 0 (low) to 1 (high), rather than as "low", "medium", or "high", as imagined previously. At what confidence level should the decision maker's

view of the advice change from irrelevant to relevant? To be clear, imagine that the occasion for the decision is reenacted many times independently under what to the decision maker appear to be identical conditions except that the confidence indicator is deliberately faked each time, being artificially increased in small steps from 0.1 to 0.9 for each reenactment. At what confidence level should the decision maker's view of the advice change from irrelevant to relevant? The interpretation of the indicator is in fact thoroughly subjective, both on the part of the expert system producing it and the decision maker interpreting it. There is way to bridge the multiple gaps in translation, between the report of a subjective confidence measure and its subjective reception by the decision maker, and finally to its objective context amid the facts of the particular decision at hand.

### 5.3 When the confidence indicator is ostensibly a probability

Developers who have recognized the difficulty have tried to compute the expert system's confidence in a way that is objectively meaningful, by making it a probability of some kind such as the probability that the automatic decision is correct. The decision maker then faces the problem of interpreting what "probability" means within the context of the problem at hand. At first glance this is simple if one thinks of coin tosses and betting, but it is quite another matter in practice when thinking about the probability of future events or the truth of assertions about the state one's environment more generally.

The difficulty of the interpretation of probability has been a longstanding debate in probability and statistics [Jaynes (2003); Howie (2002); Polya (1968); Popper (1957); Good (1962, ch.2); Savage (1954); Bartlett (1933)] and is a source of confusion for decisions in practice [Wickens and Hollands (2000; ch. 2); Pitz (1980); Slovic et al. (1977)]. To venture into that debate would take us far from decision aiding, so here we only venture so far as to note that the debate is a competition between the objective frequency definition of probability, as in games of chance, and the subjective definition of probability is a numerical representation of the state of information or subjective belief<sup>7</sup>. Given that this debate continues, it follows that developers who claim that their expert systems report objective probabilities must 1) take special care when instructing decision makers about that meaning, in order to be sure their intended meaning is conveyed without misunderstanding, resulting then in erroneous bias; and

---

<sup>7</sup> The distinction between the types of probability is stated more precisely by Bartlett (1933). The objective frequency definition Bartlett prefers to call *chance*: "...given a particular premise, it may be possible to give a probability a value with which everyone readily agrees; this value is, moreover, independent of further data or premises unless these contradict the original premise, that is, all other knowledge is irrelevant. The probability may thus be said to be a chance." Bartlett gives a Keynesian definition of subjective probability: "...it will be thus be assumed sufficient to say that a probability  $P$  of a proposition or an event  $a$  on a set of premises  $d$ , is the degree of rational belief in  $a$  given  $d$ ." It is evident why developers hesitate to lead decision makers into the subtleties along this route. The debate apparently inclines now toward subjective meaning following objective rules [Jaynes (2003); Polya (1968)]. See Howie (2002) for a detailed historical review of the debate.

2) recognize that this instruction is by no means a simple task, but calls for considerable mathematical finesse and pedagogic skill—mental acrobatics as some say. Not that decision makers are incapable of them, but the instruction obviously enters into modes of rationality that decision makers will almost certainly find uncomfortable, if not totally alien, and therefore burdensome. Some developers nevertheless hold tenaciously to objective probability indicators, seeing these as the ultimate goal for expert systems to strive for, especially for decision aiding. We therefore present further counter arguments<sup>8</sup>.

### 5.3.1 The probability that the probability is right

Much as with any parameter estimate, the question naturally arises regarding a reported probability whether it is right or not. Is it possible, for instance, that a probability-computing algorithm, proven at one time to be correct according to an objective definition of probability, may nevertheless be wrong—very wrong—at a later time for a given situation because the state of information has changed? Consider automatically classifying ships at long range through remote sensing, for example. The probability that a given contact is a fishing vessel falls dramatically when one learns that there happens to be a ban against fishing in an area that is strictly enforced for conservation purposes. Such a ban may be a late development that system developers could not have anticipated when designing the expert system. It would at least be reasonable for decision makers to question whether such information had been taken into account if the expert system had been operational for some time.

Uncertainty about probability estimates is formally called ambiguity. It is known that the effect of this ambiguity is equivalent to applying a subjective corrective adjustment to the reported probability estimate [Einhorn & Hogarth (1985); Ellsberg (1961); Skyrms (1980); Yates & Zukowski (1976)]. Thus, in practice a probability estimate is not taken at face value, but it is adjusted in the mind in a way that accounts for the ambiguity surrounding it. In decision aiding, decision makers therefore naturally apply a corrective adjustment to the objective probability reported by the system, by guessing what was known and considered relevant at the time the expert system was programmed relative to the current state of information. When ostensibly objective statements of probability are adjusted this way they of course lapse into subjectivity, becoming neither more nor less effective than informal subjective statements of confidence.

Strictly speaking, it may be argued on the other hand that assigning probabilities to probability judgments is disallowed by the very definition of probability; that, properly

---

<sup>8</sup> Note that in the design of expert systems as human substitutes rather than decision aids the debate between subjective and objective probabilities is largely irrelevant because the expert system keeps its probability valuations to itself.

computed, a probability includes all of the uncertainty about the assertions it qualifies, including uncertainty about the computation of the probability itself [Savage (1954)]. If we accept the argument, then it follows that the decision maker ought to be instructed accordingly: to accept the reported probability at face value as an unassailable fact. In other words, the expert system merits absolute trust in its computation of probability because it captures all of the ambiguity of the decision at hand. By rights, the expert system (if not the developer) should then also carry the responsibility for any errors in judgment caused by errors in computations of that probability. This is a responsibility that few developers would accept, and for good reason. Continually changing information is typical of many decision problems, especially in the military. Whereas, if probability is a numerical representation of the state of information known about an assertion [Jaynes (2003)], then it necessarily changes as the available information changes. Being derived from preprogrammed algorithms, the probabilities potentially misrepresent the current state of knowledge, making it perfectly reasonable to doubt their reliability in practice, and therefore weakening the ostensibly objective probability to a state equivalent to subjective indicators<sup>9</sup>.

### 5.3.2 Utility of probability

Pressing further nonetheless, if we ignore the above and accept that decision makers have somehow navigated the meaning and ambiguity of the objective probability report, then we may ask how it is to be used to good effect. In classical decision making [von Neumann & Morgenstern (1944)], the probabilities are matched with the costs of error to estimate the expected cost of operation; the best decision option being that which minimizes the expected cost of operation.

If the costs of error are the same for all options, then the probabilities alone are sufficient. The best decision option is simply that with the highest probability of being correct. This is the role that most developers apparently envision for an expert system's report of probability. The most probable option is then the best, and its degree of preferment is in proportion to the degree to which its probability exceeds that of all others. Note that in the mere ranking of options, objective probability accomplishes no more than any subjective numerical indicator of confidence likewise does. Hence the added meaning of objective probability must lie in the difference between the selected and other options, in the degree to which one option is preferred above the others. But the expert system typically reports the probability of only the best option, much as if relative comparisons were superfluous or too much of a burden for decision makers. The omission of the probability of all options, for whatever reason, suppresses precisely the information that would be required to convey the probability's

---

<sup>9</sup> This weakening in practice is in part the case made more generally against objective probabilities in the debate about probability [Jaynes (2003); Polya (1968)].

objective significance.

More generally, the costs of errors are of course not equal between all decision options, and the most probable option is therefore not necessarily the best. When classifying rare but dangerous targets in high-clutter scenarios, for instance, the cost of one type of error (missed targets) makes it rational to accept numerous errors of another (false alarms), whereas it is irrational (if not vain) for the decision maker to insist on being merely correct. If one aimed always to be correct much more often than one is wrong—a form of equal-cost strategy as above—then one should mechanically label every contact in a high-clutter area as “clutter”. Errors would be rare because targets are rare. If one contact in a hundred were a target, for instance, then one would score 99 % accuracy this way, entirely without effort. But the strategy is ridiculous in view of the high cost of mistaking a genuine target for mere clutter. One errs deliberately on the safe side, tolerating errors of one kind owing to the cost of errors of another kind. This is the gist of decision theory and commonsense.

For probabilities to accomplish more than subjective confidence indicators, they must therefore be fitted into a complete calculus of expected costs, with the calculus being carried out to completion either roughly within the decision maker’s mind, or more exactly by the expert system on behalf of the decision maker. In military decision problems, however, the costs of error are difficult if not impossible to estimate. They may include the cost of injury or loss of life, or the cost to a larger strategy of a single operation. Costs of error may furthermore be changing from one moment to another, as the conditions surrounding a mission change<sup>10</sup>. Fresh appraisals of costs are required for every situation, as one learns more about the situation over time or as external factors change the definition of mission success (loss or gain of resources, new time constraints, etc.). Human decision makers are given briefings to roughly calibrate their subjective sense of costs and risks, or they set about gathering information, whereas expert systems must be fed cost-relevant data directly by system operators. Such on-site management of an expert system is rarely undertaken in practice, both because the costs of errors are often admitted to be incalculable, and because the objective of the expert system was in any case merely to aid rather than to replace the human expert.

Some readers may complain that this digression about probability goes too far, that decision makers are unlikely to derive insight for a particular decision by careful study of the probabilities and costs of all options, and that here, if anywhere, we are imposing a significant cognitive burden that is ever more abstract and further removed from the decision makers proper domain of expertise, all of which is true. That is precisely the point to be made about

---

<sup>10</sup> The operational contrast between military and other decision spaces such as industrial or medical has been elaborated elsewhere, dealing primarily with the changing nature of mission information, costs, risks, and objectives [Allen & Kessel (2002); Van Der Wal (1998); Cox & Lloyd (1984)].

probabilities in computer advice. To become useful in an objective rational way they require decision makers venture far into unfamiliar dimensions of deliberation and ambiguity that promise no practical resolution.

Some may furthermore wonder if we have not gone so far as to dismiss the utility of numerical probabilities altogether. But this would again be a confusion of application, between expert systems as autonomous agents or advisors. An autonomous agent's personal numerical probability estimates are of course useful to itself, and, as Jaynes (2003) demonstrates, rational behaviour can be emulated in a robot using a well-defined numerical calculus of probabilities. It is when an advisor attempts to communicate its own probability assessments using a number whose absolute value is to be objectively meaningful to a decision maker that the utility of a numerical probability comes into question. This is particularly evident, for instance, when experts in critical operations like nuclear power generation give advice about expanding operations, by building more nuclear plants, and they are therefore asked to assess the risks of a single operation or plant. The communication of that risk, as a numerical statement of probability, is in fact very difficult to convey in a meaningful way, whether to politicians, the general public, or to university professors. Rational people interpret the same statements of risk in radically different ways, according to their subjective inclinations to optimism or pessimism, or to personal interests. Probability and likelihood are very difficult to convey in a meaningful way. Much the same is true for computer advice<sup>11</sup>.

### 5.3.3 Non-classical decision making

Finally, it bears mentioning that for command-style decisions it is known that decision makers rarely use the calculus of expected costs [Klein (1989); Klein et al. (1993); Hutchins (1995); March (1972); March (1976); Pitz (1980); Slovic et al. (1977)]. They instead adopt a more naturalistic process, like "satisficing"—selecting the first option that satisfies certain critical constraints in a sequential rehearsal of options in the mind [Klein (1989)]. Alternate decision strategies should be of great interest to system developers, to design more natural decision aids. They are not pursued here. The point to be made, rather, is that if the reported

---

<sup>11</sup> Pearl (1988; pp. 21-23) gives an illustration of the utility of probabilities, in which the disposition of mines in a mine field are conveyed by a probability map, generated by numerical computation from the density of mines rather than by precise mine locations. The probability map is obviously better for soldiers making their way across the minefield than no map at all. Notice in this case, however, that 1) it is the relative probabilities across the map, not the absolute value of probability at isolated points that are important for directing the soldiers (they follow the *lowest* probabilities as much as possible); and 2) that the probability map is derived from information that the soldiers never had, or could not be expected to remember, so they are well advised to trust the map. This is quite a different situation than the computer advice considered here, as summarized in Section 2.3. Imagine the dismay of the soldiers, for instance, if the developers of the map cautioned them *not* to trust it. Trust features centrally in the remedies for computer advice proposed later in Section 6.



probability is used in a decision maker's deliberations at all, then it is human nature to use it in a different way than in a rigorous calculus of expected costs, raising doubt about the advantage of an ostensibly objective probability in computer advice. Developers who advocate objective probabilities should furthermore be prepared to modify the decision makers' cognitive habits, which is no doubt much more psychologically intrusive than anyone imagines a decision aid should be.

## 6. Realistic integration of expert systems

---

We saw in Section 3 that the mechanisms behind computer advice, whether modeled anthropomorphically as advice between humans, or modeled instrumentally as one information cue among many, faces hurdles that developers have failed to recognize and address. Then in Section 4, we saw that computer advice introduces new dimensions of unfamiliar deliberation and ambiguity that increase the decision maker's cognitive workload with no promise of improved quality in decisions. From the decision maker's perspective, then, one perfectly rational response is to simply ignore the advice, which many do in fact, in this way eliminating confusion about responsibilities, satisfying the developer's caution against trusting the expert system, and bypassing the supervenient burden of computer advice. But the decision aid has failed entirely.

The problem lies in the developers' unqualified caution against trusting the expert system, operating principle number (1) assumed at the end of Section 2.3. An unconditional prohibition against trust admits no plausible cause-and-effect mechanism for computer advice to operate, at least none have so far been identified under such conditions. On the other hand, by enabling trust in deliberately limited ways, the developer opens the door to the benefits of automated decisions that otherwise remain blocked.

Trust in the expert system is of course problematic. It is not trusted because its performance is unknown, or is known to be inadequate under some realistic operating conditions. In this respect the developers' caution against trust is fully justified. When it comes to operational decision aiding, however, the computer advice has no mechanism to exert positive influence on the decision maker. The challenge for system developers is to discover legitimate principles of trust for admittedly imperfect expert systems.

### 6.1 Trust management

One hesitates to introduce new terms in a field of study already loaded with jargon, but it may nevertheless be useful to coin the term *trust management*, meaning the purposeful allotment of trust and responsibility among all agents, human and computer, in human-machine systems, in order to ensure that contradictions and confusion such as those plaguing computer advice are eliminated from operational systems. Unconditional cautions against trust amount to a simple negative form of trust management: there is simply to be no trust at all.

The emotion of trust is typically avoided by developers, or dismissed as an elusive, incidental human factor unworthy of serious analysis, despite researchers' work in human factors

repeatedly highlighting its central role [Muir (1987); Muir & Morray (1996); Waern et al. (1995); Waern & Ramberg (1996); Lee & Morray (1992), Chu & Rouse (1979)]. The importance of trust is reiterated here, but now in light of the preceding critique of computer advice, with a view toward system design rather than human psychology.

Two rational principles of purposeful trust management are given below. These are generalizations of system-design strategies proposed by Allen & Kessel (2002) for autodetection systems, which can be viewed as a form of computer advice. There may be other principles for managing trust. The examples illustrate the concept of trust management, and how imperfect expert systems might be turned to good use in practice. They are by no means exhaustive of the possibilities.

## **6.2 Measure of last resort**

It is rational for decision makers to trust computer advice whenever they believe the expert system may outperform themselves, perhaps because they feel incapacitated by fatigue, stress, or injury, or believe that the decision falls beyond their expertise or ability to respond. The expert system is then an instrument of last resort under abnormal conditions, when decision makers are deficient in ways that are recognizable to themselves. The change from routine decision aid to a measure of last resort is significant. Deference to the advice is automatic, rational, and responsible. There is no burden of interpreting the significance of the advice. Training amounts to practicing emergency measures rather than close observation of the expert system and long training to discover its strengths and weaknesses. And use of the expert system may follow a predetermined script devised by developers with operational results in mind. The clarity of this operating principle contrasts starkly with the ambiguity and contradiction surrounding decision aids thus far. Note that this change is more operational than technological. Thus expert systems that currently make too many errors for use in routine decision aiding may be turned to better use as an emergency measure, for even mediocre expert systems may outperform a human who is incapacitated in some way.

The principle of last resort may be used in non-emergency situations as well, to make decisions between humanly indistinguishable options. In maritime surveillance, for instance, the human decision maker may find that there are more ship contacts than can be investigated more closely using available resources and time. The contacts may be indistinguishable “dots” on a radar screen or a map, for instance, with all contacts being equally likely candidates for the target sought. Left to the decision maker, a feasible subset of contacts for further prosecution could only be chosen at random, whereas, turning to an expert system, a more probable subset could be selected on the basis of its independent analysis of each contact. Thus the expert system would serve as an intelligent filter, reducing the list of

otherwise indistinguishable contacts to a feasible number, making better use of limited resources<sup>12</sup>. Note that the inability to discriminate between options is also an incapacity of sorts, though it is not necessarily an emergency.

More generally, then, the mechanism of aiding proposed here is that decision makers must first of all have sufficient self-awareness to recognize that they are incapacitated in ways that jeopardize matters, which in turn makes it rational to transfer control to a designated alternate—the expert system. This makes sense to both decision makers and developers so long as the risks of operation using the doubtful expert system are perceived to be less than the risks of operating with an incapacitated decision maker.

Note that the expert system, when serving as a measure of last resort, serves as a temporary substitute for the human, and not as a decision aid in the advisory sense. This illustrates an important point that can be made more generally about the integration of expert systems: that expert systems, like automation generally, only become useful when substituting for humans in some way, and that the integration of expert systems in an ostensibly advisory role remains an untenable half-measure, at least until clear mechanisms of aiding have been discovered.

### **6.3 Balancing responsibilities and abilities**

To say that an expert system is untrustworthy, for whatever reason, is equivalent to saying that the responsibilities envisioned for it are beyond what is known of its abilities. And if untrustworthiness prohibits its use as a substitute for the decision maker, and then causes it to fail again in an advisory role for the reasons given earlier, the commonsense solution is to find other responsibilities for which the expert system *can* be trusted. In autodetection, for instance, this might mean reassigning the detection algorithms from their originally intended role of performing like the human operator, as if they bore full responsibility for the detection role, to a low-responsibility task such as merely improving the operator's vigilance [Kessel (2001); Allen & Kessel (2002)]. Thus an algorithm that proved untrustworthy duplicating the operator might, with straightforward sensitivity adjustments, be turned to good use enhancing the operator's vigilance.

Much the same can be done for expert systems more generally. That is, if the expert system cannot be trusted to substitute the human decision maker, then the developer is obliged to find tasks of lesser responsibility in which the expert system can be trusted. An expert system might be used to reduce information overload, for instance, by selectively suppressing irrelevant information, rather than making decisions about relevant information. This change

---

<sup>12</sup> Such filters are being considered by the Maritime Information & Knowledge Management (MIKM) Group at DRDC Atlantic for use by the Canadian Forces maritime Operation Centers (MOCs) for the surveillance of Canada's coastal waters.

of roles works in two ways: 1) irrelevance-detection tends to be simpler than relevance-detection, and 2) irrelevant information is more prevalent than relevant information, naturally giving larger, less expensive test data sets for exhaustive performance testing and verification of the expert system before integration. An example is the automatic region-of-interest detector proposed in Kessel (2002), designed to identify and suppress vacant signals from analysis, which is both simpler and easier to comprehensively evaluate than most algorithms designed to detect particular classes of targets (in this case mines on the seafloor).

Note that role re-definition for an expert system, from human-style decisions to low-responsibility mechanical decisions, is intended to make the expert system function autonomously, without decision-by-decision supervision as in the decision aid. Here autonomy does not mean replacing the decision maker, but simply filling a secondary lower-responsibility role better suited to the abilities of the expert system.

## 7. Summary and conclusions

---

It was shown that computer advice imposes new cognitive burdens of deliberation and ambiguity on decision makers, remote from their proper decision domain, that system developers likewise cannot navigate. This naturally inclines decision makers to ignore the computer advice. Those who remain determined to make use of the computer advice face a dilemma: either to independently acquire the expertise needed to purposefully interpret and apply the computer advice, through close observation of a particular expert system's behaviour under many different conditions, an intensive dubious task, or they must simply trust the expert system, following its directives implicitly. Developers cannot in good conscience recommend either course. The remedy instead is a return to the drawing board, for developers to recast the operational integration of the expert system along clear principles of cause and effect, relating means (computer decisions) to ends (better decisions). If we keep to computer advice, to a mode of operation, that is, in which the expert system recommends and the human acts, the remedy lies in establishing a rational principle of trust.

The crucial role of trust is the central message for developers to take away from the discussion. Trust is both a precondition for the decision maker to attend to the computer advice and a short cut around the cognitive burden that computer advice otherwise entails. Two examples of principles for system integration were given here: 1) using the expert system as a measure of last resort, when the decision maker is incapacitated in some way; and 2) giving the expert system simpler, low-responsibility tasks for which it is trustworthy. In either case the expert system is given limited autonomy, though never as a replacement for a capable human decision maker. On the other hand, if trust in an expert system cannot be justified, then neither can its integration into critical operations, not even as a halfway measure like decision aiding.

These conclusions follow from a critique of computer advice based on two minimum standards for decision aids (Section 3.2); namely, that a good decision aid must satisfy 1) the decision maker's subjective judgment of what constitutes a good decision, and 2) the developer's objective expectation of better quality decisions. It was shown that the common implementation for expert systems as decision aids violates the first inasmuch as decision makers, if they attend to computer advice at all, face the dilemma noted above, and the second inasmuch as developers have not identified a plausible cause-and-effect mechanism by which computer advice creates the expectation of better decisions. Two commonly held but fallacious views of expert systems were identified during the discussion: 1) that the decisions of an unreliable expert system constitute constructive advice for the capable decision maker, and 2) that a developer's knowledge about the design and workings of the expert system is sufficient for the advice to become useful. Regarding the first, it remains for developers to

formulate a clear mechanism by which doubtful automatic decisions assist a human expert, without which the decision aid is based on no more than wishful thinking. Regarding the second, it should be clear that, against real-world decisions rather than staged demonstrations of an expert system operating as a human substitute, system developers face much the same dilemma with computer advice as decision makers do. These flaws in principle are borne out by the unsatisfactory experience with computer advice in practice.

It was Leonardo Da Vinci who said, “Those who devote themselves to practice without theory are as mariners who go to sea in ships without rudder or compass.” The remark is proven by decision aids. Developers have clearly placed practice ahead of theory. Indeed, taking one step back it is remarkable that new technology, known to be unreliable in significant respects, has nevertheless been invited into the process of making critical decisions. Perhaps in no other field of system design—industrial process control, transportation, or banking—has new technology been given such generous concessions. There may be three reasons for this: 1) there prevails a naïve spirit of optimism regarding computer automation, which pushes expectations beyond realistic promise; 2) there has been a reluctance by system developers to recognize the crucial human dimensions of decision making (advice, trust, and responsibility) when integrating expert systems as decision aids; and 3) the integration of decision aids has been driven in large part by a sense of urgency which argues that in critical decisions every little bit helps, in this way encouraging implementation before understanding. Thus if the remedies for decision aiding considered here, or the incidental benefits considered elsewhere by the human-factors community (increased vigilance, de-biasing, etc.), strike the reader as disappointing or meager relative to the original promise imagined for expert systems, then one may ask if that disillusionment is due to unfounded optimism for expert systems. Expectations become realistic, on the other hand, when they are founded instead on plausible principles of cause and effect, relating means to ends as argued here.

It is important to reiterate that this critique of computer advice does not extend to expert systems generally, but only to their use in advisory roles as decision aids. Indeed, on the positive side for expert systems it is clear that they can be very useful, but only insofar as they can be trusted to fill autonomous roles, without decision-by-decision supervision, much as envisioned for expert systems all along. This has been proven by many operational examples in automatic flight control, process control, and so forth. It is the halfway measure, with the expert system serving in an ostensibly advisory role because it cannot be trusted to replace a human, that offers no benefit. The failure to distinguish between integration as human substitute and as decision aid may be the biggest confusion surrounding the operational use of expert systems. System developers therefore face renewed challenges on two fronts: to discover productive mechanisms for making operational use of imperfect expert systems, of which the mechanisms proposed here are only examples, and to produce expert systems that can be trusted to replace the human decision maker.

## 8. References

---

Allen, N.J.M. and Kessel, R.T., (2002) The roles of human operator and machine in decision aid strategies for target detection, *Proceedings of the NATO Human Factors and Medicine (HFM-084)* RTO Meeting Proceedings MP-088, ISBN 92-837-0031-7, October 2003, 13 pages.

Bartlett, M.S., (1933) Probability and chance in the theory of statistics, *Proc. Roy. Soc. A*, Vol. 141, pp.518-534

Chalmers, B.A., Burns, C.M., and Bryant, D.J., (2000), Preliminary work domain models to support naval command and control, *Proc. 7<sup>th</sup> Annual Specialists' Meeting*, TTCP MAR TP-1, Defence Research Establishment Atlantic, Nova Scotia Canada, 16-20 Oct 2000.

Chu, Y., and Rouse, W.B., (1979), Adaptive allocation of decisionmaking responsibility between human and computer in multitask situations, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-9, pp. 769-778

Cox, I. J. and Lloyd, L. J. (1984). *Artificial-Intelligence Systems in Antisubmarine Warfare: Results of a Pilot Study with Expert Systems*. North Atlantic Treaty Organization (NATO) Supreme Allied Commander Atlantic (SACLANT) Underwater Research Center, Memorandum SM-176.

Dasarathy, B.V., (1994), *Decision Fusion*, IEEE Computer Society Press, Los Alamitos California

Dasarathy, B.V., (1998), Trainable Decisions-in-decisions-out fusion system, *Proc. SPIE Conference on Sensor Fusion: Architectures, Algorithms, and Applications*, Vol. 3376, pp.35-43

Dasarathy, B.V., (2003), *Multi-sensor, multi-source information fusion: architectures, algorithms, and applications*, SC149 Short Course Notes, SPIE, The International Society for Optical Engineering, April 2003

Dewey, J., (1938) *Logic: The Theory of Inquiry*, republished Irvington Publications (1982)

Duda, R.O., Hart, P.E., and Stork, D.G., (2001) *Pattern classification*, 2<sup>nd</sup> Ed., Wiley-Interscience Publications, New York

Einhorn, H.J., and Hogarth, R.M., (1985). Ambiguity and uncertainty in probabilistic inference, *Psychological Review*, 92, pp. 433-461



- Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms, *Quarterly Journal of Economics*, 75, pp. 643-669
- Gadamer, H.G. (1982) *Reason in the age of science*, translated by F.G.Lawrence, MIT press, Cambridge, pp.71.
- Good, I.J. (1952), Rational Decisions, *J. Roy. Statistical Soc.*, B14, pp. 107-114
- Good, I.J., (1962) *Good Thinking: The foundations of probability and its application*, University of Minnesota Press
- Green, D.E., and Swets, J.A., (1988), *Signal detection theory and psychophysics*, Peninsula Publishing, Los Altos California
- Hastie, T., Tibshirani, R., and Friedman, J., (2001) *The elements of statistical learning: data mining, inference, and prediction*, Springer-Verlag, New York
- Hopple, G. W. (1986). Decision Aiding Dangers: The Law of the Hammer and Other Maxims. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-16, No. 6, pp.948-964.
- Howie, D., (2002) *Interpreting probability: controversies and developments in the early twentieth century*, Cambridge University Press, Cambridge
- Heraclitus (500 b.c.) (reprinted 2001) *Fragments: the collected wisdom of Heraclitus*, translated by B. Haxton, Viking Penguin, New York.
- Hutchins, E., (1995) *Cognition in the wild*, MIT Press, Cambridge Mass.
- Jaynes, E.T., (2003) *Probability Theory: The logic of science*, Cambridge University Press
- Johannsen, G., Runsdorp, J.E., and Sage, A.P., (1983) Human system interface concerns in support system design, *Automatica*, Vol. 19, pp. 595-603
- Kessel, R.T. (2001). *On-Screen Alarms in Computer-Aided Detection Systems: Signal Processing, Human Factors, and System Design*. Defence Research Establishment Atlantic Technical Memorandum DREA TM 2001-184, Nov 2001.
- Kessel, R.T. (2002) Using sonar speckle to identify regions of interest and for mine detection, *Proc. of the SPIE, Optical Engineering Society*, Vol. 4742, AeroSense April 2002.
- Keyes, J., (1989) "Why expert systems fail", *AI Expert*, Nov 1989, pp. 50-53

- Klein, G.A., (1989) "Recognition primed decisions", *Advances in Man-Machine Systems Research*, Vol. 5, pp. 47-92
- Klein, G.A., Orasanu, J, Calderwood, R., and Zsombok, C.E., (1993) *Decision making in action: models and methods*, Ablex Publishing Corp, Norwood, New Jersey
- Langer, E., (1994), The illusion of calculated decisions, In R.C. Schank and E. Langer, Eds., *Beliefs, Reasoning, and Decision Making*, Lawrence Erlbaum, New Jersey
- Lee, J. and Moray, N., (1992) Trust, control strategies and allocation of function in human-machine systems, *Ergonomics*, Vol. 35, pp. 1243-1270
- Levi, I., (1997). *The Covenant of Reason: Rationality and the Commitments of Thought*, Cambridge: Cambridge University Press.
- MacDonald, Dettwiler and Associates, (2003) "Way-ahead Report for a Semisubmersible-based Remote Minehunting Capability for the Canadian Navy, Volume 1 - Development Report", Contractor's Report, Remote Minehunting System (RMS) Technology Demonstration Project (TDP), Defence R&D Canada – Atlantic, DN0446, April 2003.
- March, J.G. (1972), Bounded rationality, ambiguity, and the engineering of choice, *Bell Journal of Economics*, Vol 9, pp. 587-608
- March, J.G., (1976), The technology of foolishness, in *Ambiguity and choice in organizations*, March J.G., and Olsen Eds., pp. 69-81
- Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *Int. J. Man-Machine Studies*, vol. 27, pp.527-539.
- Muir, B.M., and Moray, N. (1996) Trust in automation. Part II. Experimental studies of trust and human intervention in a process of control simulation, *Ergonomics*, Vol. 39, pp. 429-460
- Norman, D.A., (1990) The 'problem' with automation: inappropriate feedback and interaction, not 'over-automation', *Phil. Trans. R. Soc. Lond.*, B 327, pp. 585-593
- Parasuraman, R. (1997), Humans and automation: Use, misuse, disuse, abuse, *Human Factors*, Vol. 39, pp. 230-253

- Pearl, J. (1988), *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, Morgan Kaufmann Publishers, Inc, San Mateo.
- Pierce, C.S. (1878), *Philosophical writings of Peirce*, Buchler J. Editor, Dover Publication Inc., New York, 1987
- Pitz, G.F., (1980) The very guide of life: The use of probabilistic information for making decisions, Chapter 5 in *Cognitive processes in choice and decision behaviour*, T.S. Wallsten Ed., Lawrence Erlbaum Associates, New Jersey
- Poincare, H., (1952) *Science and Hypothesis*, Dover Press, New York
- Polanyi, M. (1962), *Personal knowledge: towards a post-critical philosophy*, The University of Chicago Press, Chicago
- Pollack, M.E., Hirschberg, J., and Webber, B., (1982) User participation in the reasoning processes in expert systems, *Proc. National Conference on Artificial Intelligence AAI-82*, pp.358-361
- Polya, G., (1968) *Patterns of plausible inference, Volume II*, Princeton University Press, London
- Pomranky, R., Dzindolet, M.T., and Peterson, S., Violations of expectations: a study of automation use, *Proc. 6<sup>th</sup> ICCRTS, Collaboration in the Information Age*, June 19-21, 2001, Command & Control Research Program (CCRP)
- Popper, K., (1959) The propensity interpretation of probability, *British Journal for the Philosophy of Science*, Vol. 10, pp. 25-42
- Ripley, B.D., (1996), *Pattern recognition and neural networks*, Cambridge University Press, Cambridge
- Savage, L.J., (1954) *The Foundations of Statistics*. 1972 edition, New York: Dover.
- Silverman, B.G., (1992), Human-Computer Collaboration, *Human-Machine Interaction*, Vol. 7, pp.165-196
- Skitka, L.J, Mosier, K.L, and Burdick, M., (1999). Does automation bias decision making?, *Int. J. Human-Computer Studies*, 51, pp. 991-1006
- Skyrms, B., (1980). Higher order degrees of belief, Chapt.6 in *Prospects for pragmatism: Essays in memory of F.P. Ramsey*, D.H. Mellor Ed., Cambridge University Press.

- Sloane, S.B., (1991) The use of artificial intelligence by the United States Navy: Case study in failure, *AI Magazine*, Spring 1991, pp.80-92.
- Slovic, P., Fischhoff, B., and Lichtenstein, S., (1977), Behavioural decision theory, *Ann. Rev. Psychol.* Vol. 28, pp. 1-39
- Sorkin, R.D., and Woods, D.D., (1985), Systems with Human Monitors: A signal detection analysis, *Human-Computer Interaction*, Vol. 1, pp. 49-75
- Suh, C-K. and Suh E-H., (1993) Using human factor guidelines for developing expert systems, *Expert Systems*, Vol. 10, pp. 151-156.
- Van Der Wal, A.J., (1998) The importance of soft computing methods for military observation systems, *Proc. Third International FLINS Workshop*, Sept 1998, pp. 163-170
- von Neumann, J., and Morgenstern, O., (1944) *Theory of games and economic behaviour*, Princeton University Press
- Wærn, Y. Häggglund, S., Ramberg, R., Rankin, I., Harrius, J. (1995). Computational Advice and Explanations - Behavioural and Computational Aspects. In: K. Nordby, P.H. Helmersen, D.J. Gilmore and S.A. Arnesen (eds.), *Human-Computer Interaction INTERACT'95*, Chapman & Hall, Oxford, 1995.
- Waern, Y., and Ramberg R. (1996), People's perception of human and computer advice, *Computers in human behaviour*, Vol 12, pp. 17-27
- Wickens, C. D. and Hollands, J. G. (2000). *Engineering Psychology and Human Performance, third edition*. New Jersey: Prentice-Hall.
- Woods, D.D. (1985). Cognitive technologies: the design of joint human-machine cognitive systems, *AI Magazine*, pp. 86-91
- Woods, D. D. (1986). The Design of Decision Aids in the Age of "Intelligence", *Proceedings of the 1986 IEEE International Conference on Systems, Man, and Cybernetics*, pp.398-401.
- Woods, D. D. and Roth, E. M. (1988). Aiding Human Performance II: From Cognitive Analysis to Support Systems. *Le Travail humain*, tome 51, no. 2, pp. 139-172.

Woods, D.D. (1996). Decomposing Automation: Apparent simplicity, real complexity, in *Automation and Human Performance Theory and Applications*, R. Parasuraman and M Mouloua Editors, Erlbaum, London

Wooldridge, M., and Jennings, N.R., (1998), Pitfalls of agent-oriented development, *Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98)*

Yates, J.F., and Zukowski, L.G., (1976), Characterization of ambiguity in decision making, *Behavioral Science*, Vol. 21, pp. 19-25

## Distribution list

---

### Internal Distribution List

DRDC Atlantic TR 2003-241

- 1 – Director General
- 1 – Deputy Director General
- 6 – Document Library
- 1 – Head/MICS
- 1 – R. T. Kessel (author)
- 1 – B. Chalmers
- 1 – B. McArthur
- 1 – D. Chapman
- 1 – B. Campbell
- 1 – T. Hammond
- 1 – L. Lapinski
- 1 – LCdr B. MacLennan
- 1 – N. Allen
- 1 – D. Hopkin
- 1 – J. Fawcett

**External Distribution List  
DRDC Atlantic TR 2003-241**

- 1 – DRDKIM
- 1 – DRDKIM (unbound copy)
- 1 – DRDC
- 3 –DRDC Toronto
  - Attn: R. Pigeau
  - S. McFadden
  - J. Crebholder
- 5 – DRDC Valcartier
  - Attn: R. Breton
  - R. Rousseau
  - M. Gauvin
  - E. Bosse
  - A.C. Bourry-Brisset
- 2 – DRDC Ottawa
  - Attn: C. Helleur
  - G. Geling
  
- 1 – Commanding Officer,  
MARLANT, Trinity
  
- 1 – Commanding Officer,  
MARPAAC, Athena
  
- 1 – Michael Matthews, Ph.D.,  
Humansystems Inc.  
2<sup>nd</sup> Floor, 111 Farquhar St.  
Guelph, Ontario  
N1H 3N4
  
- 1 – V. L. Myers,  
NATO Undersea Research Centre  
Viale San Bartolomeo 400, 19138 La Spezia (SP), ITALY

This page intentionally left blank.





13. **ABSTRACT** (a brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual).

Where expert systems are not trusted to replace human decision makers, they have instead been proposed as decision aids. By reporting its automatic decisions the expert system naturally helps the human arrive at better decisions, or so the reasoning goes. On further analysis, one quickly discovers that the usual vision for decision aids entails contradictions that are confusing if not disabling. Computer advice invokes unfamiliar deliberation and ambiguity about its interpretation within the context of a particular decision and about system performance, for which neither decision makers nor system developers are prepared. Here we examine the problematic nature of computer advice, examining in particular the mechanisms (or lack of them) by which a computer's decision becomes helpful to a human decision maker, identifying common fallacies in the decision-aid concept, setting minimum standards that decision aids must meet if they are to be useful, pointing to the crucial role of trust, and illustrating how system developers can integrate expert systems following rational principles of trust. Decision makers who have known the frustration of using expert-system decision aids should find confirmation of their experience and material for defence against system developers who do not appreciate the implications of the systems they propose, and developers should be redirected by this work toward more realistic and profitable principles of decision-aid design.

14. **KEYWORDS, DESCRIPTORS or IDENTIFIERS** (technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus. e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus-identified. If not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title).

Expert systems, decision aids, computer advice, human-machine interaction, human-machine systems, human-computer interaction, knowledge management, intelligent systems, artificial intelligence

This page intentionally left blank.

## **Defence R&D Canada**

**Canada's leader in defence  
and national security R&D**

## **R & D pour la défense Canada**

**Chef de file au Canada en R & D  
pour la défense et la sécurité nationale**



**[www.drdc-rddc.gc.ca](http://www.drdc-rddc.gc.ca)**