Copy No._____

Defence Research and Development Canada
Recherche et développement pour la défense Canada

DEFENCE R&D DÉFENSE

Apparent Reliability

Conditions for Reliance on Supervised Automation

Ronald T. Kessel

Defence R&D Canada – Atlantic

Technical Memorandum
DRDC Atlantic TM 2005-155
July 2005

Canada

This page intentionally left blank.

Apparent Reliability

Conditions for Reliance on Supervised Automation

Ronald T. Kessel

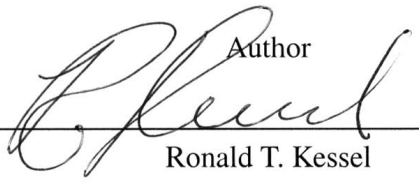
Defence R & D Canada – Atlantic

Technical Memorandum

DRDC Atlantic TM 2005-155

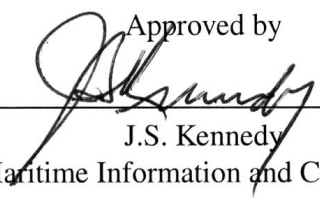
July 2005

Author



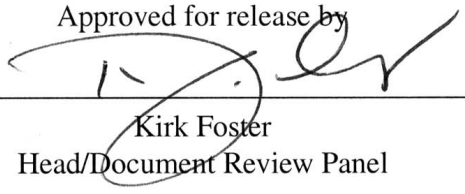
Ronald T. Kessel

Approved by



J.S. Kennedy
Head/Maritime Information and Combat Systems

Approved for release by



Kirk Foster
Head/Document Review Panel

Abstract

When experts are given automation for assistance, they naturally observe its actions and judge its reliability for themselves. The automation is supervised, for a time at least, when it must win the trust of its users before it can serve effectively through routine use. Lack of trust, on the other hand, results in under utilization or total disuse of the automation. This occurs so often in practice that researchers investigating trust in automation now speak of a prevailing "bias toward self reliance" among users. Here a quantitative analysis of apparent reliability is applied to supervised automation. It is shown that subjective reliability assessment imposes minimum standards of proficiency on the user of the automation. It takes a high-performance user, that is, to recognize the high-performance of automation in action. This is demonstrated here using the applied mathematics of operation analysis. Minimum performance standards for both user and automation are derived that must be met before it is plausible that the automation may appear reliable to its user. The standards are surprisingly stringent for critical applications, creating a natural barrier—or bias—against reliance that is unavoidable with supervised automation. These are conditions of feasibility that must be met to avoid failure in disuse for supervised automation. Special attention is given to automation within a constant false alarm rate (CFAR) concept of operations, and for command and control (C2) and situation awareness.

Résumé

Quand on fournit à des experts du matériel d'automatisation, il est naturel que ceux-ci en observent le fonctionnement et en jugent eux-mêmes la fiabilité. L'automatisation est donc supervisée, pendant un certain temps du moins, et doit gagner la confiance de ses utilisateurs avant de pouvoir servir de façon efficace sur une base régulière. Le manque de confiance, par ailleurs, occasionne une sous-utilisation de l'automatisation, voire son rejet total. Cette situation se produit si souvent que les chercheurs qui examinent la confiance dans le secteur de l'automatisation parlent maintenant d'un penchant marqué pour l'autosuffisance chez les utilisateurs. On applique ici une analyse quantitative de la fiabilité apparente à l'automatisation supervisée. Il est démontré que l'évaluation de fiabilité subjective impose des normes minimales de compétence à l'utilisateur de l'automatisation. En réalité, l'utilisateur doit être de haut niveau pour reconnaître le haut rendement de l'automatisation. On en fait la démonstration ici à l'aide des mathématiques appliquées de l'analyse

des opérations. Des normes minimales de rendement sont dérivées tant pour l'utilisateur que pour l'automatisation, et il faut les respecter avant qu'il ne soit plausible que l'automatisation semble fiable pour l'utilisateur. Les normes sont étonnamment strictes pour les applications critiques, ce qui pose un obstacle naturel à la confiance Ű ou un penchant naturel pour la méfiance Ű que l'on ne peut pas éviter dans le cas de l'automatisation supervisée. Certaines conditions de faisabilité doivent être respectées pour éviter l'échec de la désuétude de l'automatisation supervisée. Une attention particulière est accordée à l'automatisation dans un concept opérationnel où le taux de fausse alarme est constant (TFAC), de même qu'aux fins du commandement et du contrôle (C2) ainsi que de la connaissance de la situation (CS).

Executive summary

Background

Where automation is not trusted to replace a human expert, it is often introduced under human supervision. It is to be relied upon for assistance insofar as it appears to the supervisory expert to be reliable in its actions. Automation that fails to win trust will not be used. Disuse or marked under-utilization occurs often enough that it has been attributed to a prevailing "bias toward self reliance" among users. It has been prevalent in military applications like automatic target detection and recognition, and decision aids.

Principal results

The apparent reliability of automation is a prior condition for trust and reliance. Here it is shown that apparent reliability depends not only on the true reliability of the automation, but on the true reliability of the user as well. In effect, it takes a high-performance user to recognize reliability in automation. Minimum performance standards are derived for both human and automation. These are conditions of feasibility—the conditions under which it is plausible that supervised automation will appear reliable to its user. Thus the mode of human-machine interaction imposes hidden conditions of its own on human and machine that are no less important than the more obvious considerations, the cost of errors, stress and workload on the user, overall mission success, and so forth. Illustrations are drawn from autodetection (including DRDC's Remote Minehunting System (RMS) TDP), situation awareness, command and control, and other applications.

Significance of results

Longstanding troublesome issues in human-machine interaction are addressed, helping engineers to understand, avoid, and fix (by system redesign) the lapse into disuse even before a system is actually built. The analysis finds application wherever supervised automation might be considered for assistance in military or civilian applications. It should therefore play a role in many DRDC projects.

Ronald T. Kessel; 2006; Apparent Reliability;
DRDC Atlantic TM 2005-155; Defence R & D Canada – Atlantic.

Sommaire

Contexte

Quand on ne fait pas confiance à l'automatisation pour remplacer un expert humain, on l'introduit souvent sous la supervision d'une personne. On s'en sert dans la mesure où l'expert qui en assure la surveillance la juge fiable. Le matériel automatisé qui ne réussit pas à gagner la confiance est écarté. La mise au rancart ou une nette sous-utilisation sont des pratiques si courantes qu'on les attribue à une tendance marquée pour l'autosuffisance chez les utilisateurs.

Principaux résultats

Il est démontré ici que la fiabilité apparente dépend non seulement de la fiabilité de l'automatisation comme telle, mais aussi de la fiabilité de l'utilisateur. En réalité, l'utilisateur doit être de haut niveau pour reconnaître la fiabilité de l'automatisation. Des normes minimales de rendement sont dérivées tant pour l'utilisateur que pour l'automatisation. Ce sont des conditions de faisabilité — les conditions dans lesquelles il est plausible que l'automatisation supervisée semble fiable pour l'utilisateur. Ainsi, le mode d'interaction homme-machine impose à l'homme et à la machine ses propres conditions cachées qui n'ont pas moins d'importance que les considérations les plus évidentes, le coût des erreurs, du stress et de la charge de travail pour l'utilisateur, la réussite globale de la mission, et ainsi de suite. On cite des exemples tirés de l'autodétection, de la connaissance de la situation, du commandement et du contrôle et d'autres applications.

Signification des résultats

On s'intéresse à des enjeux délicats de l'interaction homme-machine qui ne datent pas d'hier, ce qui aide les ingénieurs à comprendre, à éviter et à corriger (par la restructuration des systèmes) la désuétude, même avant qu'un système ne soit fabriqué. L'analyse pourrait s'appliquer à n'importe quelle situation où l'on pourrait envisager l'automatisation supervisée d'applications militaires ou civiles. Elle devrait donc avoir une incidence sur de nombreux projets de RDDC.

Ronald T. Kessel; 2006; Fiabilité apparente : condition pour avoir confiance dans l'automatisation supervisée ;
DRDC Atlantic TM 2005-155; R & D pour la défense Canada – Atlantique.

Table of contents

List of figures

This page intentionally left blank.

1 Introduction

To *rely on automation* means to entrust some undertaking to its automatic functioning. It may be part of a military mission, surveillance for homeland defence, industrial process control, medical diagnosis, the flight of an aircraft, and so forth. Typically it is a task that people find monotonous to do themselves, though it might call on occasion for critical action to be taken. Automation is introduced to assist by performing the task automatically, though under supervision because the automation is not trusted to replace a person entirely. Thus the person responsible for performing the task, now the *user* of the automation, is to observe the automation in action and decide whether to rely on it or not. Automation that fails to win trust at this stage—or, what amounts to much the same thing, that never appears to be reliable—will not be used.

Disuse or marked under-utilization is so commonly encountered [1]-[3] that it has been attributed to a "bias toward self reliance" among users [4]. Disuse has been especially prevalent in critical military applications like autodetection and target recognition. Everyone who is familiar with autodetection in real-world operations will know the frustration of continually adjusting the sensitivity of the automation during the course of a mission. The adjustments are made repeatedly, but satisfactory performance is never achieved. Either too many obviously false alarms are raised, or too many plausible target contacts are missed, or so it seems at least to the user. The automation is eventually abandoned because the user has more important things to be doing at such times.

Defence R&D Canada developed autodetection and target classification for their Remote Minehunting System (RMS) technology demonstration project (TDP), to provide a semi-autonomous sea minehunting capability. Among the project's many goals was autodetection and autoclassification of mine targets, operating for a time in supervised mode as a step (ideally) toward more complete automation of operations. These particular goals were ranked as high priority and high risk objectives; the risk being in large part the disuse found so often in other systems. The understanding and analysis of human-machine interaction would have to be advanced if the new system would avoid the same fate. This paper is one outcome of that work, but the analysis applies far beyond autodetection, to supervised automation more generally.

As many researchers have pointed out [5]-[11], the link between the reliability of automation and its use is ultimately a matter of trust. Although trust is admittedly an emotion or attitude like love, anger, or hatred, it must be reduced to a numerical quantity for the sake of analysis, to discover its dynamics and

influence through experiment, and to prescribe what constitutes "appropriate" trust. This numerical reduction brings with it all of the methodological and theoretical difficulties that quantitative psychology typically faces. Measurement and analysis require precise definitions, for instance. Many papers therefore begin with a survey of standard usage of the word "trust", though falling short of the force required for unambiguous treatment, and short of deriving quantitative rules of engineering for system design. Indeed, the connection between numerical "trust", reliant behaviour, and the operational gains (or losses) at stake, have so far remained qualitative, consisting of empirical correlations and tendencies (like the bias toward self reliance) rather than quantitative prescriptions that can steer developers away from failure before a system has actually been built and tested.

Things will prove very difficult for system developers if one must venture so far into the psychology of trust for the successful integration of automation into operations. But one can go a long way in analysis while stopping short of a psychology of trust. As suggested at the outset, trust and reliant behaviour in experts depend largely on the *apparent reliability* of supervised automation. Apparent reliability is the condition for trust and reliance to follow, and it is therefore a condition for supervised automation. The story of supervised automation therefore begins with apparent reliability for the system developer. Achieving so much would at least be an excellent first step toward purposeful system design.

Apparent reliability is not the full story behind reliant behaviour, of course, because there are many reasons to resort to automation other than its apparent reliability. One may rely on it as a measure of last resort when "anything is better than nothing", when one is incapacitated in some way, or rely on it out of a legal or moral obligation, and so forth. Indeed, it is when reliant behaviour runs counter to the apparent reliability that one must especially turn to psychology, to begin (as researchers have) to explain why users choose not to rely on apparently reliable automation, or why they choose to rely on apparently unreliable automation. To rely on apparently reliable automation, on the other hand, hardly calls for any explanation at all. It is the obviously rational thing to do. It is this rationally motivated reliance that we consider here.

The advantage of apparent reliability is that it is amenable to analysis using applied mathematics. As we shall see, it can be derived from the *true* reliabilities (the performance probabilities) of both the user and automation working independently against the same task. These true reliabilities are in principle objectively measurable by experiment or by the systematic monitoring of long-term observations. In target detection they would be the probabilities of de-

tection and false alarm [12][13], for both the user and the automation when they are working independently. It is shown that the subjective assessment of reliability by the user depends in part on his or her ability in the same task, and that the subjective assessment of reliability therefore imposes minimum standards of proficiency on the user, quite independently of any other operational considerations such as the costs and risks making errors, the user's workload and stress, and so forth—much of which is central to the quantitative psychology of trust and reliant behaviour. In addition, the automation must of course meet minimum standards of proficiency for it to appear reliable. The minimum proficiencies for both user and automation are derived here. They are the conditions of feasibility for supervised automation. As we shall see, the conditions can be very stringent for critical (high reliability) tasks.

1.1 Outline

The approach will be to first review the paradigm assumed here for supervised automation (Section 2). Then a mathematical analysis will be fit to that paradigm (Section 3). As with applied mathematics generally, much effort goes toward the problem definition, but the final results can be quite simple. The analysis is then applied to the simplest case of single-action tasks (Section 4). Quantitative conditions of feasibility are given, as well as examples to show how they may be generalized to many different applications. The implications for system design and improvement are then drawn in the remainder, to show that *difficult* tasks are *not* generally candidates for supervised automation (Section 5), to show what remediation is possible when the conditions of feasibility are not met (Section 6), and to show how the experience with DRDC's RMS TDP project confirms the analysis (Section 7) .

2 Paradigm assumed for supervised automation

By *supervised automation* it is meant that:

1. The automation is not trusted to replace the responsible expert in a particular task.

2. The expert has the option to execute the task him or herself, or to give the task to the automation.

3. The automation's reliability is an open question to be settled by the expert at some time during operations.

4. The expert relies on the automation when he or she judges it to be reliable.

5. The expert's judgement of reliability is based on his or her observation and opinion of the automation's actions or recommendations.

Subjective reliability assessments are often the norm in practice, either because perfect and complete verification of actions or recommendations is simply not available outside of controlled trials, or because the notion of verification may itself be ambiguous. In process control or flying an aircraft, for instance, the task amounts to adjusting many continuously variable parameters, and there may be debate and variability among experts about what constitutes appropriate action. One must speak then of the automation performing responsibly rather than of being obviously right or wrong. Reliability and verification must therefore be understood broadly here to mean that the actions of the automation (or the user) meets or fails the expectations of its users (or system analysts) in some important respects.

All automation does not fall under the five characteristics above, of course. Autonomous robots or unmanned vehicles may typically eliminate immediate human supervision of the automation, if only for a time. As suggested earlier, moreover, the users of automation may rely on automation for reasons other than its apparent reliability—as a measure of last resort, when one is incapacitated, not confident, inexpert, and so forth. These may be very important occasions for reliance, and they may be very important system-design considerations for particular applications, but they deviate from the paradigm for supervised automation that is envisioned here inasmuch as they are exceptional occasions, not routine, in which the user's reliance is motivated expressly by the *lack* of expertise, perhaps even its abdication, rather than by the positive assertion of expertise. The failure to clearly distinguish between the two modes of human-machine interaction, between reliance out of the strength of expertise and reliance out of weakness, is a source of confusion for system developers. The gains to be had from routine reliance on automation by capable experts, of instance, must be calculated quite differently from the gains during emergency states or other unpredictable transients. The mode of human-machine interaction envisioned here is admittedly idealistic. Reliance on automation because it appears to be reliable is the perhaps best outcome that system engineers could hope to see for the automation they propose, so the paradigm serves as a worthy target for system design.

This ideal mode of human-machine interaction, insofar as it is achieved in practice, does not itself guarantee that the automation will be beneficial to

the mission as a whole. Overall benefit depends on the particular context of the application, especially the costs and risks of error during operations. These are not considered here. Indeed, the point to be made is that the mode of human-machine interaction imposes conditions of feasibility entirely of its own making on both human and machine, quite apart from the bigger question of the operational benefit or effectiveness of the missions that the human-machine team are to carry out. Here, then, we consider the conditions of feasibility for supervised automation, but not proofs of its overall operational utility.

3 Mathematical analysis

If the user relies on his or her situation awareness and expert judgement to decide when the automation is right or wrong, or when it behaves responsibly or irresponsibly, then it means that the automation's true reliability asserts itself through the agreement or disagreement it finds in the user. If the user agrees with the automation's action or recommendation, then its apparent reliability increases and reliant behaviour is reinforced. If the user disagrees, its apparent reliability decreases and reliant behaviour is inhibited. Thus the force of the automation's true reliability in action—its apparent reliability, that is to say—consists of the agreements and disagreements between the automation and its user. Apparent reliability therefore follows at least two independent modes of inquiry and measurement:

1. **Agreement rate** a—Given the actions (recommendations) of the automation during operations, the user can agree or disagree with those actions (recommendations made). The agreement rate may be assessed by the probability that, given an action taken (recommendation) by the automation, the user agrees with that action (recommendation), which may be written as

$$a = P\left(\text{user agrees} \mid \text{automation executes action}\right). \tag{1}$$

2. **Miss rate** m—Given no actions (recommendations) on the part of the automation at times during operations, the user may believe that occasions for action (recommendation) have been missed by the automation. This may be assessed by the probability that, given that the user sees an occasion for action, the automation fails to take action,

$$m = P\left(\text{automation executes no action} \mid \text{user sees occasion for action}\right). \tag{2}$$

Other more specialized directions of inquiry are possible, perhaps to discriminate different types of disagreements when several options for action are possible, with some disagreements among those options being more critical than others. But (1) and (2) are generally applicable to many cases, and they are complete for many tasks.

The assertion that the user relies on automation when he or she finds it to be reliable can be written mathematically as the dual condition

$$a > \alpha_R \text{ and } (1 - m) > \alpha_R, \tag{3}$$

in which α_R is some minimum threshold of reliability. Here it is assumed for simplicity that the threshold is the same for both modes of inquiry. This is plausible insofar as reliability exists as a property that can be treated in abstraction, apart from the particular technology or methodology in question. Much as medical research sets uniform standards of reliable evidence across many different research questions (the *significance* in hypothesis testing [14]), and much as academics set uniform passing grades across many fields of study, the users of automation may likewise impose roughly uniform standards across different forms of apparent reliability. The analysis could be carried out with two different thresholds if necessary.

3.1 Occasions for action and inaction

Of all the events that the real world presents to a mission or operation, reliability is concerned only with those that potentially lead to action. These include genuine occasions for action (call them *class 1* events), and occasions when no action is required but in which an agent might mistakenly take action nonetheless (call them *class 0* events). In problems of military target detection, for instance, class 1 events would be the coming in range of enemy forces, and class 0 events may be the coming in range of confusable target-like contacts (friendly or neutral forces, sensor noise, elements of the environment, and so forth). With autopilots and process control, class 1 events may be occasions for a corrective manoeuvre or change of settings, and class 0 events may be random fluctuations in operating conditions caused by turbulence or benign process fluctuations that in reality do not call for action. Let the prior probabilities of encountering each class of events be

$$\begin{aligned} p_1 &= P\left(\text{encounter class 1 event} \mid \text{realistic operations}\right), \\ p_0 &= P\left(\text{encounter class 0 event} \mid \text{realistic operations}\right). \end{aligned} \tag{4}$$

Here

$$p_0 + p_1 = 1 \tag{5}$$

because class 1 and 0 constitute the entire set of relevant events. In sustainable or monotonous operations, class 1 events are usually rare, making p_1 much less than p_0,

$$p_1 \ll p_0. \tag{6}$$

This is true of target detection in many military applications (submarine hunting, mine hunting, etc.), surveillance operations, manoeuvreless flight, and so forth.

In an objective analysis of operations, the probabilities p_1 and p_0 would be the relative frequencies of occurrence of class 1 and class 0 events. These might be estimated for a given application by operation or military analysts. Since we are speaking here of apparent reliabilities, however, they might also represent subjective probabilities of expectation in the user's mind. Objective and subjective probabilities are generally different, but their calculus is identical [15][16]. Thus, so far as p_1 and p_0 are concerned, we can change between objective and subjective probabilities and the subsequent analysis remains unchanged. Ideally, the user would "calibrate" his or her subjective probabilities to match the objective, through past experience, intelligence reports and briefings, coaching by superiors, and so forth. One could not go far wrong at the preliminary design stage to assume optimistically that both the analyst's objective and the user's subjective probabilities are the same. The distinction provides avenues for further analysis, but we do not pursue them here.

3.2 Human, machine, and joint performance probabilities

A number of other probabilities will be necessary for analysis. The list is long so we resort to tabulation. Most of these will be eliminated through a process of simplification. Since there are two agents (automation (A) and user (human, H)), and two reliability metrics (a and m), there are four situations to consider:

H	user acting independently
A	automation acting independently
AH	automation acts and is judged by human (for a)
HA	user acts, but automation does not act (for m)

The capitals on the left shall be used as superscripts to denote the situations to which probabilities pertain, and the subscripts 1 or 0 will be used to denote

class-conditional probabilities in the following way:

$$\begin{array}{rcl} P^H & = & P\left(H \text{ takes action} \mid \text{ any event }\right) \\ P_1^H & = & P\left(H \text{ takes action} \mid \text{ event class 1 }\right) \\ P_0^H & = & P\left(H \text{ takes action} \mid \text{ event class 0 }\right) \\ & & P^H = P_1^H p_1 + P_0^H p_0 \end{array} \tag{7}$$

$$\begin{array}{rcl} P^A & = & P\left(A \text{ takes action} \mid \text{ any event }\right) \\ P_1^A & = & P\left(A \text{ takes action} \mid \text{ event class 1)}\right. \\ P_0^A & = & P\left(A \text{ takes action} \mid \text{ event class 0 }\right) \\ & & P^A = P_1^A p_1 + P_0^A p_0 \end{array} \tag{8}$$

$$\begin{array}{rcl} P^{AH} & = & P\left(A \text{ takes action AND } H \text{ agrees} \mid \text{any event }\right) \\ P_1^{AH} & = & P\left(A \text{ takes action AND } H \text{ agrees} \mid \text{event class 1 }\right) \\ P_0^{AH} & = & P\left(A \text{ takes action AND } H \text{ agrees} \mid \text{event class 0 }\right) \\ & & P^{AH} = P_1^{AH} p_1 + P_0^{AH} p_0 \end{array} \tag{9}$$

$$\begin{array}{rcl} P^{HA} & = & P\left(\ H \text{ takes action AND } A \text{ takes none} \mid \text{any event }\right) \\ P_1^{HA} & = & P\left(\ H \text{ takes action AND } A \text{ takes none} \mid \text{ event class 1)}\right. \\ P_0^{HA} & = & P\left(\ H \text{ takes action AND } A \text{ takes none} \mid \text{ event class 0)}\right. \\ & & P^{HA} = P_1^{HA} p_1 + P_0^{HA} p_0 \end{array} \tag{10}$$

Note that the user and automation mistake class 0 events for class 1 with a probability of P_0^H and P_0^A, respectively. In a detection task, these would be the probabilities of false alarm for each agent working independently. By the same token, P_1^A and P_1^H would be the probabilities of detection. For autopilots in aircraft, P_0^H and P_0^A would be the probabilities of each agent effecting inappropriate mission-affecting action when no action was required, and P_1^A and P_1^H would be the probabilities of each independently taking appropriate action given an occasion for such action. Unlike the prior class probabilities p_1 and p_0, all of the above conditional probabilities denoted using a capital P are objectively measurable probabilities (relative frequencies of occurrence).

3.3 Apparent reliabilities

Recall that Bayes theorem for the conditional probability of outcome Y occurring given that outcome X has occurred may be written as [17]

$$P\left(Y \mid X\right) = \frac{P\left(Y \text{ AND } X\right)}{P\left(X\right)}, \tag{11}$$

in which $P\left(Y \text{ AND } X\right)$ is the probability of both outcomes X and Y occurring together, and $P\left(X\right)$ is the probability of outcome X occurring. In the notation above, the equivalent form

$$P\left(Y \mid X, \text{ any event}\right) = \frac{P\left(Y \text{ AND } X \mid \text{ any event}\right)}{P\left(X \mid \text{any event}\right)} \tag{12}$$

may be better. Outcome X may be that the automation takes action, for instance, and Y that the user agrees (i.e., takes the *same* action). Then a in (1) becomes

$$a = \frac{P^{AH}}{P^{A}} = \frac{P_1^{AH} p_1 + P_0^{AH} p_0}{P_1^{A} p_1 + P_0^{A} p_0}. \tag{13}$$

On the other hand, X may be the outcome that the automation takes no action, and Y that the user takes some action, in which case m in (2) becomes

$$m = \frac{P^{HA}}{P^{H}} = \frac{P_1^{HA} p_1 + P_0^{HA} p_0}{P_1^{H} p_1 + P_0^{H} p_0}. \tag{14}$$

These are the metrics by which apparent reliability can be judged using (3), to determine if the automation will appear reliable to its user.

3.4 Independence of user and automation

More can be said about the joint conditional probabilities, $P_{1,0}^{AH}$ and $P_{1,0}^{HA}$. They depend, of course, on the performance of each agent acting independently (on $P_{1,0}^{A}$ and $P_{1,0}^{H}$), but also possibly on the influence that one agent may exert on the other. The user's judgement can influence the automation's through the sensitivity (or other) adjustments the user might make on the automation, when the automation appears to be performing unreliably. Given noticeable disagreement or miss rates, for instance, the user typically has the option to adjust the sensitivity or inclinations of the automation to improve agreement or reduce misses. The actions of the automation would depend in some way on the judgement of the user while the adjustments are being made, and the joint conditional probabilities, $P_{1,0}^{AH}$ and $P_{1,0}^{HA}$ would be difficult, perhaps impossible, to estimate during those times. Outside of these adjustments, however, the actions of the automation are of course not influenced by the user's expert opinion about whether its actions are right or wrong. The automation then acts independently.

On the other hand, the automation may also exert an influence on the judgment of the user. If users lack confidence, are novice users, find their task very difficult, face new operating conditions, or feel incapacitated in some way, then they may take cues for action from the automation. Benefits can no doubt be had from automation this way, but reliance is then motivated by the lack of expertise rather than out of its strength as assumed here. Insofar as the user's judgment derives from his or her domain of expertise for doing the task at hand—insofar as the user is assessing the automation's reliability and not his or her own reliability, that is to say—the user's judgement will not be biased by the particular actions of the automation, but will be based instead

on an expert awareness of the occasion for action. The user is in this respect independent of the automation.

The independence of the automation and user is admittedly an idealization of real-world human-machine interactions. But it closely follows the paradigm for a supervised automation as it is usually envisioned. Independence will be used in the next section to simplify and progress the analysis.

4 Single-action tasks

Single-action tasks are the simplest to analyze. They amount to the raising of an alarm of some kind, in which the normal or predominant condition of operation is the non-alarmed state. It might be the detection of a military target, dangerous conditions in process control, or a positive test-result for a rare condition in medicine. The single-action task is archetypal of multi-option tasks because it illustrates the dynamics of supervised automation in their simplest form, and because any multi-option task can in principle be subdivided into a series (decision tree) of progressive single-action tasks. To make the discussion concrete we shall speak in terms of target detection, bearing in mind that the single-action task applies to many different applications as we shall see later.

In target detection it is customary to represent a detector's (human or machine) ability using a receiver-operating characteristic (ROC) curve [12][13], as shown in Fig.(1). The horizontal axis is the probability of false alarm—the probability of mistaking a non-action event (class 0, clutter) for an action event (class 1, target). These are P_0^A and P_0^H given above for the automation and user respectively. The vertical axis is the probability of detection—the probability, that is, of correctly treating an action event as an action event. These are P_1^A and P_1^H above for automation and user respectively. A detector's performance is characterized then by a curve of ordered pairs, $\left(P_0^A, P_1^A\right)$ or $\left(P_0^H, P_1^H\right)$, which encompass the full range of the agent's ability as the aversion to missed targets or false alarms is changed. The point at which the detector operates best is determined by the likelihood of encountering action events (class 1), and the aversion to missed targets and false alarms [12], all of which typically change from mission to mission, or throughout the course of a single mission. It is the operating point of the automation on its ROC curve that users typically change when they adjust its sensitivity for instance. The ROC curves for high-performing detectors lie in the upper left corner of the ROC plane, where detection is high and false alarms are low. The ROC curve applies in fact to any single-action task, not just detection tasks.

In a single action task, to agree about the need for action is the same as agree-

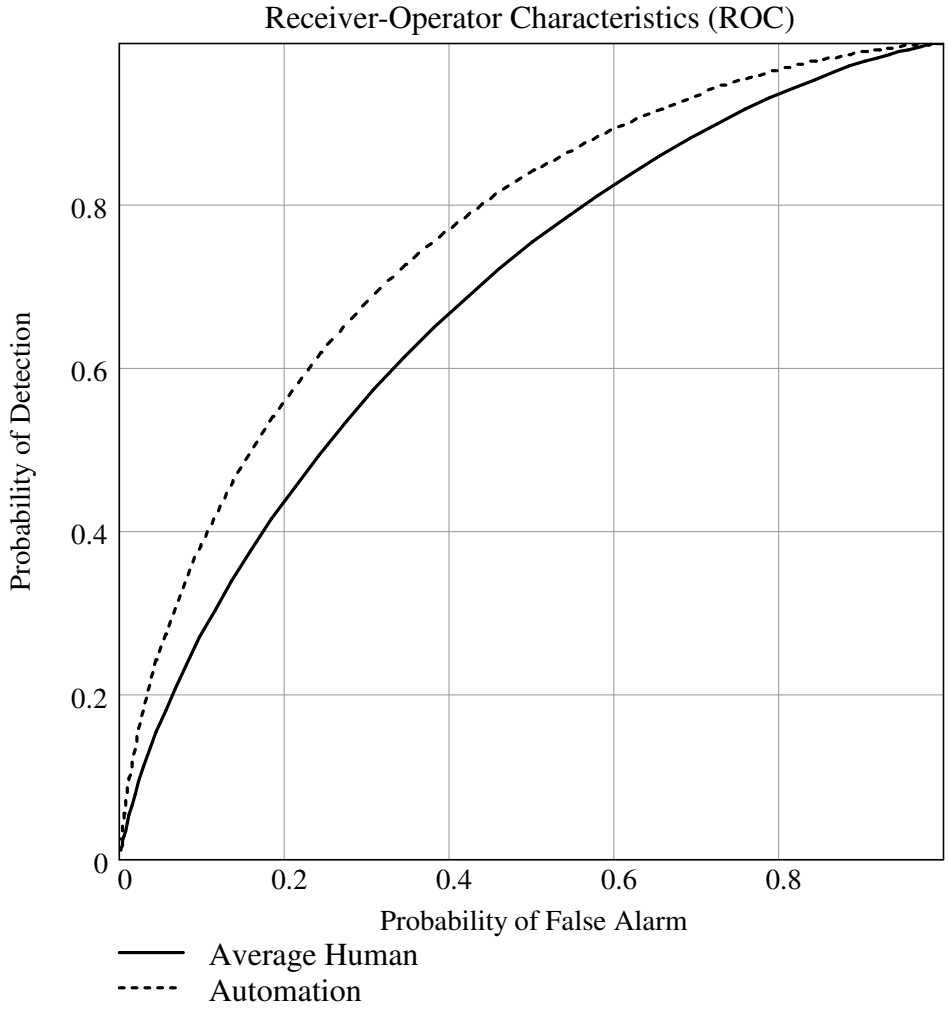


Figure 1: *The performance in an uncertain single-action task is best represented using the receiver-operating characteristic (ROC) curve. Human performance is compared here with automation in a difficult target recognition task. The further the curve ventures into the upper left corner of the graph, the better the performance. In the case illustrated, the automation is a better detector than the average human, for instance.*

ing about the particular action that is required, hence the joint probabilities introduced above are

$$\begin{array}{ll} P_1^{AH} = P_1^A P_1^H, & A \text{ and } H \text{ correctly recognize class 1 event,} \\ P_0^{AH} = P_0^A P_0^H, & A \text{ and } H \text{ incorrectly recognize class 0 event,} \\ P_1^{HA} = P_1^H \left(1 - P_1^A\right), & A \text{ misses class 1 event that } H \text{ correctly sees,} \\ P_0^{HA} = P_0^H \left(1 - P_0^A\right), & A \text{ misses class 0 event that } H \text{ mistakenly sees.} \end{array} \tag{15}$$

These are simple products because the user and automation are assumed to make their decisions about action independently, even when working together in a human-machine team as discussed earlier in Section 3.4.

Using (15) in a (13) and m (14) gives

$$a = \frac{P_1^H P_1^A p_1 + P_0^H P_0^A p_0}{P_1^A p_1 + P_0^A p_0}, \tag{16}$$

and

$$m = \frac{P_1^H \left(1 - P_1^A\right) p_1 + P_0^H \left(1 - P_0^A\right) p_0}{P_1^H p_1 + P_0^H p_0}. \tag{17}$$

These are estimates of the apparent reliability for supervised automation in a single-action task.

4.1 Minimum standard of performance imposed on user

The supervisor-supervised mode of human-machine interaction assumes that the user will recognize good reliability when he or she sees it in action. Let us examine how well this assumption applies in reality by testing the apparent reliabilities a and m when the automation happens to be perfectly reliable.

When the true reliability of the automation is *perfect*, the automation gives perfect detection of all targets with no false alarms,

$$P_1^A \to 1 \text{ and } P_0^A \to 0. \tag{18}$$

If that perfect reliability were evident to the user, then it would mean that $a \to 1$ and $m \to 0$. Notice, however, that substituting perfect automation (18) into (16) and (17) gives less than perfect apparent reliability,

$$a = P_1^H \tag{19}$$

and

$$m = \frac{P_0^H p_0}{P_1^H p_1 + P_0^H p_0}. \tag{20}$$

This is because perfect reliability is being assessed subjectively, by an imperfect user.

A good illustration of this how this might appear in practice comes from target detection for mine hunting. In the search for mines on the seafloor, in a large portion of the seafloor where there happens to be much clutter but no mines, perfect automation would raise no alarms, because there are no mines present. The supervisory user, on the other hand, owing to his or her imperfect ability against so much clutter, would be inclined to raise a number of alarms, all of them false. In our analysis, the apparent miss rate m in (20) of the automation would be noticeably high. Thus the automation would appear to the user to be missing prospective targets, which is a very serious apparent failure to the mind of a responsible user, and the automation would be judged unreliable.

More generally, if the apparent reliability of perfectly reliable automation is to be greater than α_R as in (3), then (19) and (20) give

$$P_1^H > \alpha_R \quad \text{and} \quad P_1^H > P_0^H \frac{p_0}{p_1} \frac{\alpha_R}{1 - \alpha_R}. \tag{21}$$

These are constraints *on the user* in order for perfectly reliable automation to appear reliable to them. They are therefore necessary conditions for appropriate expertise-motivated reliance on supervised automation.

4.1.1 Example: Sea minehunting

These constraints can be plotted on the user's receiver-operator characteristic (ROC) curve. Let us assume that the user's standard of reliability is $\alpha_R = 0.8$, which is perhaps low for critical tasks but will serve for illustration. In other words, the user expects to see greater than 80 % reliability against events that he or she believes to be action events (class 1). And let us assume that the user's expectation of encountering an action event (class 1) is $p_1 = 0.20$, which would typically be rather high for realistic military detection as in minehunting or submarine hunting. Recall that $p_0 = 1 - p_1 = 0.80$ by (5). Figure (2) shows the dual constraints (21) on the user for these nominal conditions. What is remarkable is how expert and tightly constrained the user's performance must be for supervised automation to be feasible in single-action tasks. The user, as supervisor and judge of reliability in the automation, must have very high probabilities of detection P_1^H, and very low probabilities of false alarm P_0^H. Otherwise even very good automation will appear unreliable to the user

of the automation, and it will lapse into disuse during routine operations. The constraints on the user become more stringent as the reliability threshold α_R increases, and as the likelihood p_1 of encountering genuine class 1 events decreases.

Figure (2) also shows the ROC curve measured elsewhere [18] for human operators discriminating man-made mine-sized objects on the seafloor from naturally occurring mine-sized clutter (mostly boulders) in sonar imagery. This recognition problem is so challenging that the average user's performance falls well outside the constraints (21). This recognition task is therefore not a candidate for automated assistance in a supervised mode of human-machine interaction. In practice the automation would lapse into total disuse, which has in fact been the experience with many automatic target recognition for tasks. It must be emphasized that it is the user's imperfections, not the automation's, that create the barrier against reliance in this case.

4.1.2 Example: Autopilots

If we think instead of instances where supervised automation has been relied on by experts in critical applications, such as a pilot's routine use of an autopilot, it is evident that the automation is being entrusted in those cases with tasks that the user finds easy to perform, such as for manoeuvreless flight across long distances. Being reliable performers themselves, the users are capable of recognizing reliability in automation, which is the precondition for responsibly passing control to the automation for a time in a supervisor-supervised mode of human-machine interaction.

The constraints (21) assume a single-action task, which is admittedly simplistic for the many actions pilots might perform. One could go back to redefine the agreement and miss rates, a in (13) and m in (14), and perhaps the probabilities (7) to (10) as well, to derive more specialized conditions of feasibility for complex tasks like flight and process control. But the results would differ in complexity, not in their essential conclusions, namely, that the mode of human-machine interaction places quantitative conditions on the reliability of the *user*, and these conditions become increasingly stringent as the task becomes more critical (require high reliability α_R), and as the occasions for critical action become rare (p_1 becomes small).

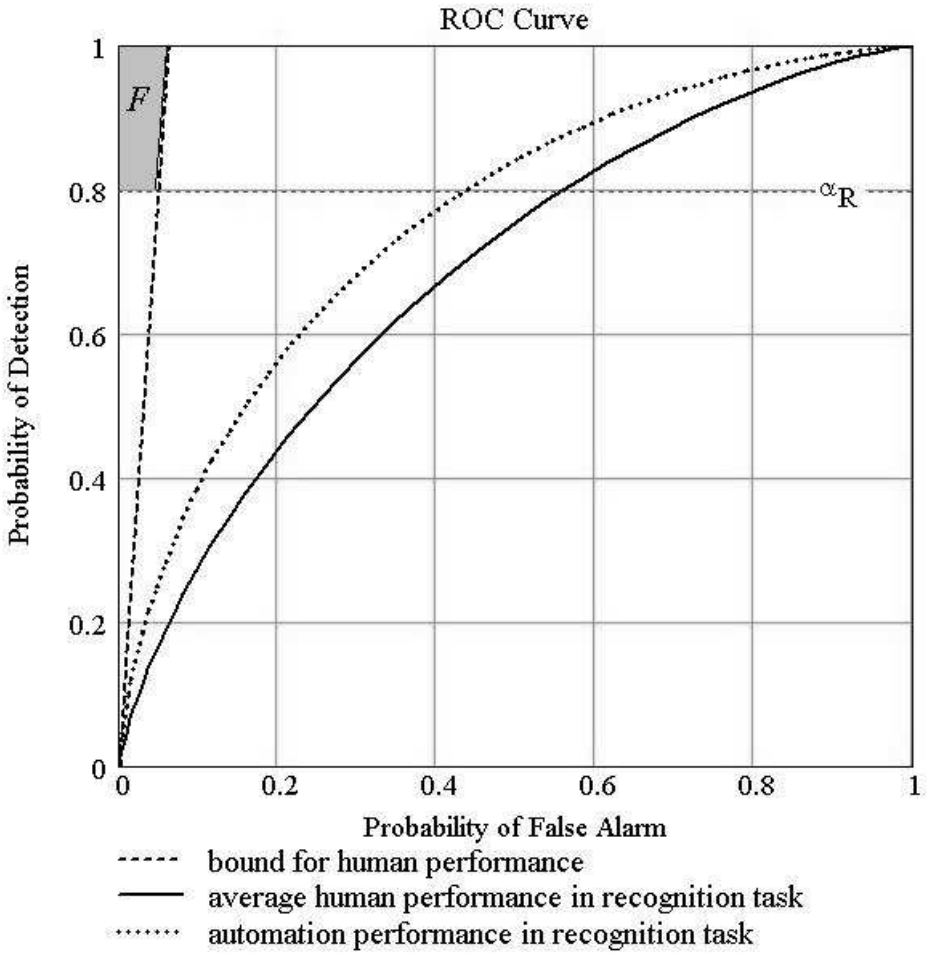


Figure 2: *The user's performance must fall within the region marked F for supervised automation to be plausibly feasible. The region is bounded by the dual constraints given by (21). The apparent reliability is greater than* $\alpha_R = 0.80$, *and occasions for action are relatively rare* $p_1 = 0.20$. *In Section 4.2 it is shown that the performance of the supervised automation must also fall in the region* F.

4.2 Minimum standard of performance imposed on automation

The perfect user is the ideal supervisor for automation inasmuch as the reliability is correctly judged by the user without bias or error. Here we reverse the previous analysis, assuming that the user is perfect and the automation imperfect at the task, as some developers may assume in fact out of deference to the expert user. We assume the perfect user here to discover how reliable imperfect automation must be for it to appear reliable, to set constraints now on the performance of the automation for it to justifiably win trust and reliance, by virtue of a perfectly correct assessment of its true reliability.

For the perfect user, $P_1^H \to 1$ and $P_0^H \to 0$, and the apparent reliabilities a (16) and m (17) become

$$a = \frac{P_1^A p_1}{P_1^A p_1 + P_0^A p_0} \tag{22}$$

and

$$m = 1 - P_1^A. \tag{23}$$

Rearranging these much as before, we have constraints on the performance of the automation

$$P_1^A > P_0^A \frac{p_0}{p_1} \frac{\alpha_R}{(1-\alpha_R)} \quad \text{and} \quad P_1^A > \alpha_R \; . \tag{24}$$

These have the same form as the constraints (21) for the user.

It is important to point out that a in (22) is not the maximum possible agreement rate in practice. Nor is m in (23) the minimum possible miss rate. Higher a or lower m are possible in practice given imperfect users of imperfect automation, since the user and automation happen to agree in cases when they are *both* mistaken. But the system designer should not try to maximize that kind of agreement. The intent is not to maximize the apparent reliability, as if apparent reliability were valuable for operations in its own right. The intent here, rather, is to estimate how large the apparent reliability will be when it is correctly assessed, and to use this ideal limiting case as a template for feasible system design. If the condition (24) is not satisfied, for instance, then it means that the apparent reliability could only cross the reliability threshold α_R in part by accident, by capitalizing on the simultaneous errors of both user and automation, which cannot be the basis for good system design. The condition (24) is therefore a necessary condition if the apparent reliability is to be a sound cue for reliance for the user.

4.2.1 Example: Sea minehunting continued

To illustrate, Fig.(2) includes the ROC curve reported elsewhere for a pattern-recognition algorithm [19] for performing the same mine-clutter discrimination task [18] used earlier in the example of Section 4.1.1. Here again, the ROC curve for the automation falls well outside the upper-left corner zone. Hence it would as a rule appear unreliable to the perfect user, and reliable to the imperfect user only by accident. Thus it should not be used in a supervised mode of human-machine interaction. The fact that the automation outperforms the user is not enough to make the automation apparently reliable in this case as shown in [18].

4.3 Constant false-alarm rate (CFAR) concept of operations (CONOPs)

In high-clutter/rare-target applications, the calculable costs of operation are dominated by the cost of prosecuting false alarms. It is usual then to require that detection operations yield a predetermined, tolerably low probability of false alarm P_{FA} [20]. This is widely used in practice for military surveillance, in the detection of targets by sonar and radar for instance. In such applications, for the sake of preliminary analysis, it is plausible to assume that both user and automation have roughly equal probabilities of false alarm, and that these are in turn roughly equal the predetermined false alarm rate P_{FA},

$$P_0^A = P_0^H = P_{FA}. \tag{25}$$

Let us furthermore assume optimistically that both user and automation have high probabilities of detection at this probability of false alarm,

$$P_1^H \to 1 \text{ and } P_1^A \to 1. \tag{26}$$

Using (25) and (26) in the apparent reliabilities a (16) and m (17), gives

$$a = \frac{p_1 + P_{FA}^2 p_0}{p_1 + P_{FA} p_0} = \frac{p_1 + P_{FA}^2 (1 - p_1)}{p_1 + P_{FA} (1 - p_1)} > \alpha_R, \tag{27}$$

and

$$m = 1 - a < 1 - \alpha_R. \tag{28}$$

Equation (27) is a condition now on the CFAR detection task itself inasmuch as the prior expectation p_1 of encountering a target is determined by one's prior knowledge of the enemy's inclination to deploy targets, and the allowable P_{FA}

is fixed by the amount of clutter one faces and the cost of prosecuting false alarms.

The left side of the condition (27) is plotted in Fig.(3). Notice that α_R is greater than 0.9 in the region *below* the diagonal contour marked 0.9. Given the logarithmic scales in the figure, this implies a ten-times rule of feasibility for supervised automation within a CFAR concept of operations,

$$p_1 > 10 \times P_{FA}. \tag{29}$$

The rule is an optimistic constraint because it assumes that both user and automation have high probabilities detection (26). In practice one would like to see the condition (29) hold with a factor somewhat bigger than 10.

Expressed in words, the ten-times rule (29) says that the prior expectation of encountering a genuine target (class 1) must as a rule be over ten times greater than the expectation of seeing a false alarm ($P_{FA}p_0 \rightarrow P_{FA}$) . If this is not true, then the automation will always appear unreliable to the user, either raising too many false alarms, missing too many likely targets, or both. This is in fact the experience with automatic detection algorithms that have lapsed into disuse.

It should be clear that (29) is not a condition of feasibility for the CFAR concept of operations in its own right. It is rather the feasibility of introducing automation in a supervised role within the CFAR concept of operations. It is therefore an engineering rule of thumb by which to decide if a given CFAR application is a candidate for the introduction of supervised automation. If (29) is not satisfied, that is to say, then the system developer cannot count on the apparent reliability of automation to motivate the user's reliance on that automation.

4.4 Situation awareness and command and control (C2)

Situation awareness has been a topic of military interest in command and control. The greater the situation awareness, that is, the fewer the errors in a commander's judgement, and the more effective are his or her decisions. So much may be obvious. A precise concept of situation awareness is nevertheless difficult to define. (See the thorough survey by Breton and Rousseau [21].) For our purposes it is enough to say that situation awareness constitutes the information from which decisions are made, and, in particular, the information from which a fundamental single-action decision is repeatedly made, namely, whether to continue with the present course of action or to change it. The

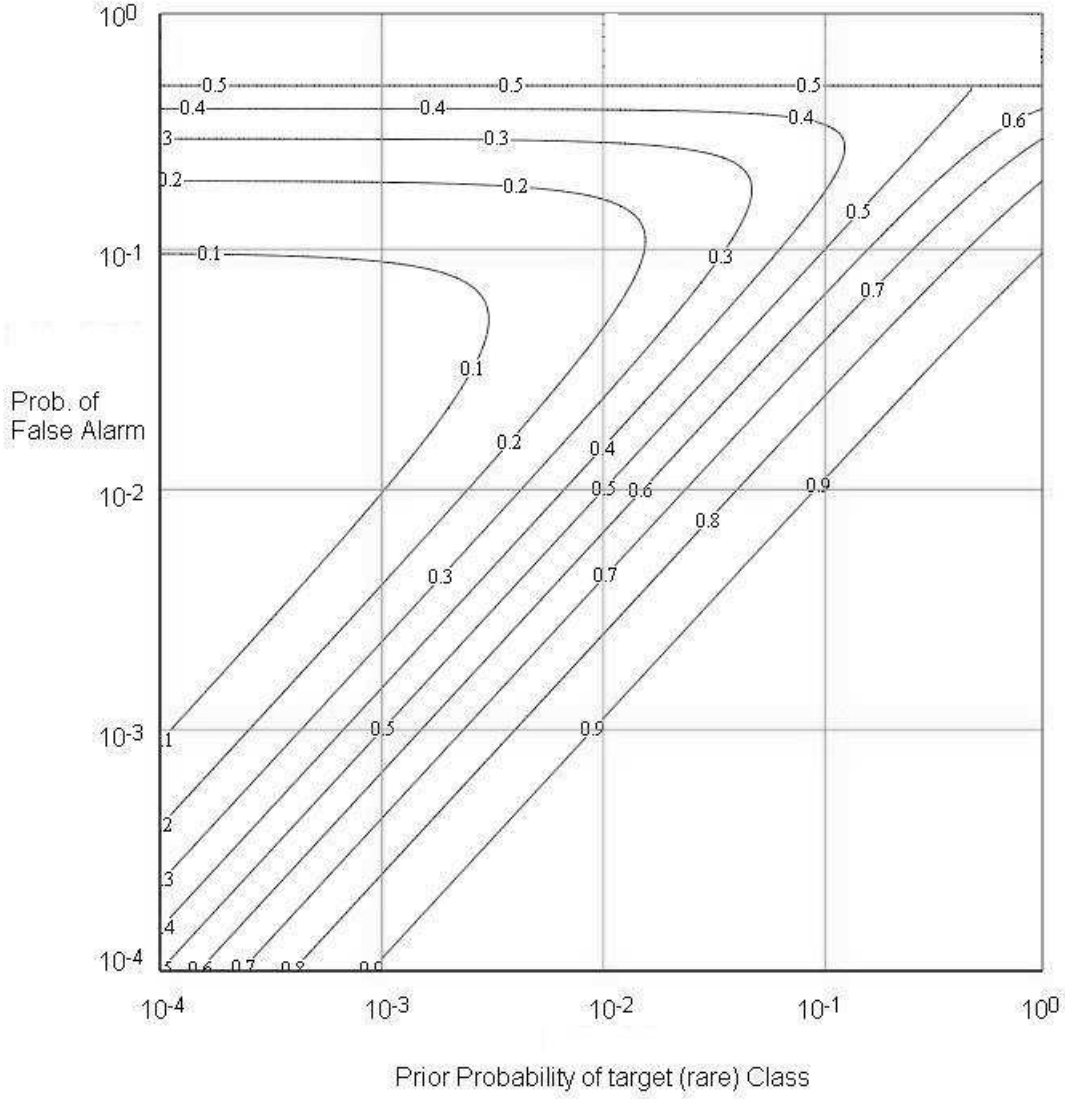


Figure 3: *The apparent reliability (left hand side of the condition of feasibility (27)) is plotted here for a CFAR concept of operations. The $\alpha_R = 0.90$ contour gives the ten-times rule for feasible supervised automation (29).*

single action, then, is simply to recognize occasions for responsible action. It may be a change in stance or attitude toward events, the raising of an alarm, the commitment of resources, a change in course or speed, and so forth.

p_0 is therefore the probability that the continuation of one's present stance or behavior is appropriate, and p_1 the probability that its continuation is inappropriate and a change is required. And P_0^H is the probability of mistakenly changing one's current stance or behavior given that it continues in reality to be appropriate, whereas P_1^H is the probability of correctly recognizing (detecting) the need for change. It may be possible for operations analysts to actually estimate p_1, $P_{1,0}^H$, and α_R for certain well-defined mission types, but their values are not needed for theoretical treatment.

Let us speculate, for instance, that at some time in the future a system developer will invent an automatic commander of sorts—automation for making critical command-style decisions in real time and initiating corresponding action. This seems to be the goal of many system developers, in fact, at least for certain elements of command and control like real-time threat assessment, mission planning, resource management, and so forth. Now a supervisory commander is expected to rely on this new automated assistant when it appears to be reliable. The question that our hypothetical system developer must ask, then, is whether the new automation will plausibly appear reliable to its human supervisor from observation of its behavior, where reliability means the ability to correctly recognize occasions on which commands ought to be made. At this stage the developer does not ask whether the correct commands are issued, but only whether the occasions for issuing commands are recognized by the automation.

The constraints derived earlier for users (21) are now constraints on the commander who is supervising and potentially using the automation. These constraints define, in effect, the level of situation awareness that the supervisory human commander must have for it to be possible for him or her to recognize reliable situation awareness in the automation. Good situation awareness means that occasions for issuing commands are easily recognized $\left(P_1^H \text{ is high}\right)$, and falsely motivated commands are easily avoided $\left(P_0^H \text{ is low}\right)$. And very good situation awareness would be a necessary condition for appropriate reliance of the human commander on the automation.

This example shows how general the present analysis can be. Its generality follows from its use of probabilities—the cornerstone for operations analysis of all kinds. Indeed, what occurs hypothetically above for auto-commanders occurs in effect more generally for supervised automation whenever a responsible expert chooses to substitute him or herself for a time with automation.

Responsible self-substitution depends on the apparent reliability of the automation, which depends first of all on its apparent reliability at recognizing occasions for action. Thus the conditions on user (21) and automation (24) derived for single-action tasks apply for any task whatever when class 1 events are defined to be occasions for action, and class 0 events to be occasions for which no action is required but might be mistaken for class 1 events. It is in this respect, at least, that the single-action task applies to all possible applications of supervised automation.

5 Supervised automation for easy and difficult tasks

We have seen that the conditions of feasibility for supervised automation are that the ROC curves for both user and automation must lie far in the upper left corner of the ROC plane (Fig.(2)), and that the constraint becomes more stringent for critical tasks (high reliability α_R) in which critical events (class 1) are rare (small p_1). The stringent conditions mean that both user and automation must be high performers in the task to be automated. High performance in a task means that it is performed with low error rates and high certainty, which is as much as to say that the task is in some important respect *easy* for the agent performing it. Expressed in words, then, conditions of feasibility (21) and (24) mean that both the user and the automation must find the task *easy* to perform for that task qualify as a candidate for improvement by supervised automation.

Put another way, tasks that either user or automation find *difficult* to do—that is, tasks that are susceptible to significant error rates or high uncertainty—are not candidates for improvement by supervised automation. Indeed, they encounter one of two modes of failure:

1. If the task is difficult for the automation but easy for the user, then the automation will appear unreliable and not be relied upon. This is the most obvious diagnosis for disuse of automation. It naturally sends system developers back to the drawing board to somehow improve the performance of the automation. But the diagnosis is not always correct.

2. It may be that the task is simply too difficult for the user, in which case the user cannot recognize reliability in the automation by observing and "expertly" judging its actions. In difficult tasks, the automation will appear to be unreliable, no matter how genuinely reliable it may be. When this happens, the system developers must look for alternate non-supervisory

modes of human-machine interaction, or for non-subjective means of reliability assessment for the users of the automation.

The second mode of failure may be a very common for supervised automation. It is in difficult and critical tasks that the drive to introduce automation, even un-trusted automation, is greatest owing to a sense of necessity and urgency to provide automated assistance of some kind. Any automation is better than nothing, or so the reasoning goes. Yet it is in precisely these difficult tasks that supervised automation will not function. The task is too difficult for the user to recognize reliable automation when he or she sees it in action. The main results to be taken from this analysis are the quantitative conditions under which the second mode of failure occurs.

6 Making supervised automation feasible

If a system analyst or developer finds that the conditions (21), (24), or (29) are not met, then it means that automation should not be introduced into operations in a supervised role because reliance on it cannot be expected. But the matter does not end there. If supervised automation fails for one application, it may nevertheless serve with little change in another application in which the constraints are satisfied. If the ten-times rule (29) is not satisfied for a CFAR application, for instance, then it is because the probability of false alarm P_{FA} is too high relative to the expectation p_1 of seeing targets. By redefining what constitutes a target and clutter, however, one might change the inequality (29) dramatically.

For example, conditions of feasibility might be violated in a minehunting application, when discriminating between mines and mine-sized rocks, for instance. But they might be satisfied when discriminating more generally between "objects of interest" (merely mine-like) and other clutter. If the operator faces a monotony of uninteresting sonar data, for instance, then the automation might screen the data for the operator, suppressing large portions of it, but passing on all objects of interest for further consideration by the human user independently. Apparently unreliable autodetection might then be turned to good operational use in object-of-interest-detection. In terms of the present analysis, the target class is expanded this way, making p_1 larger, while also reducing the reliability threshold α_R because the automation no longer bears the full weight of the mine recognition task. Both changes increase the likelihood of meeting the ten-times rule (29).

Much the same remedy applies very generally to the other conditions (21) and (24). The system developer must consult with the operations analyst to identify operationally useful tasks, to define their class 1 and 0 events and their expectations($p_{1,0}$), to estimate the performance probabilities of human and machine against these classes $\left(P^H_{1,0} \text{ and } P^A_{1,0}\right)$, and to do this as many times as it takes to find an operationally useful task for which the conditions of feasibility (21) and (24) are satisfied. These tasks would be the candidates for supervised automation.

7 Experience with DRDC's Remote Minehunting System (RMS) Technology Demonstration Project (TDP)

This approach has been loosely followed in DRDC's RMS TDP. Among its many goals, autodetection and autoclassification of mine targets in sonar imagery were ranked as a high priority, high risk objectives. The development team quickly realized that good estimates of human and automation performance in the mine detection task would be necessary, so an aggressive program of ground-truth signature collection[22] ran in parallel with the preliminary algorithm development. In this way a large data base of target and clutter imagery was compiled for controlled testing of automation and human performance; the results being ROC curves $\left(P^A_0, P^A_1\right)$ and $\left(P^H_0, P^H_1\right)$ for several different class 1/0 definitions (see [18] for example).

Experience gained through the development cycles of the project confirm the conclusion of the present analysis. It became evident, for instance, that mine (class 1)/clutter(class 0) discriminations in real time on the basis of single sonar "snap-shots" (as opposed to multiple looks or a systematic route-survey program for change detection) were difficult for both human sonar operators and automation to make reliably; and that the task was simplified somewhat by making man-made(class 1)/natural(class 0) discriminations instead [19], in this way broadening the target class, in effect increasing p_1 (though its value was unknown), while at the same time decreasing the critical emphasis of the task, thereby decreasing α_R. The change in class definition seemed necessary as experience was gained in working with autodetection algorithms and the database of contact imagery.

An even more dramatic change of classes was also made for a region-of-interest

detector [23], in which the alternate task is to detect non-vacant regions of the seafloor—to discriminate, that is, between energetic signatures (any feature on the seafloor whatsoever, class 1) from perfectly non-energetic signatures (featureless gravels or sands, class 0). Class 1 is then very broad, including any mine-sized patch of the seafloor that exhibits any feature whatsoever, which therefore potentially contains an object of interest and warrants further analysis by human or another machine. This is a way of automatically reducing (if necessary) the amount of data transmitted though a band-limited radio link, for instance, or of reducing the amount of imagery scanned by human or automation for greater focus and efficiency. The region-of-interest detector was ultimately integrated as a filter to make other specialized autodetection algorithms run more efficiently. It was integrated in an unsupervised mode, but the manner of its integration is especially relevant here. It operated for a time in a supervised mode, with the algorithm developers acting as supervisors while they turned their efforts more efficiently on other more advanced elements of autodetection. And it was only after the region-of-interest detector was judged to be reliable, through observation of its action, that it was integrated in its final unsupervised form. This detector therefore illustrates the advantage of changing class definitions, to increase the automation's apparent reliability, in this case making a supervised mode of operation productive to the point of final integration without supervision. It also illustrates the need to identify *easy* tasks for supervised automation inasmuch as region-of-interest detection is a task that humans (in this case the developers) find very easy to do correctly.

The RMS TDP completed its final demonstrations with a strong showing by autodetection algorithms, but without determining the extent to which they would be relied upon by system users in practice. Some developers suspected that the automation may have outperformed the user, but that its performance was nevertheless inadequate to win the users' (or the developers') trust in practice. The present analysis shows that this could be true, yet the superior reliability would not be evident to the user, or to anyone else observing a system demonstration because the mine-recognition task is by nature difficult and uncertain for the observer to do. The quantitative standards of feasibility derived here remain to be applied to the automation now operating in the RMS, or to its further advances made by DRDC Atlantic's Mine and Torpedo Defence Group.

8 Conclusions

The concept of apparent reliability addresses the feasibility of supervised automation but stops short of the larger psychological issues of trust and reliant

behavior. It was shown that the performance of a user shapes the way in which reliability will be assessed, and that this imposes conditions on the ability of a user of the automation to recognize reliable automation when they observe it in action. Indeed, it takes a high performing user to recognize high performance automation. This was shown analytically for the a single-action task, for which minimum performance limits were derived for user (21) and automation (24). These are the conditions of feasibility for supervised automation—the conditions, that is to say, under which supervised automation will plausibly appear reliable to its user. The conditions are particularly stringent for critical tasks (low p_1 but high α_R), which suggests that the bias that users have against reliance on automation is actually the natural consequence of the supervisor-supervised paradigm for human-machine interaction. It was shown how these conditions can guide the design and integration of automation, helping engineers and analysts to predict and remedy instances of dysfunctional (disused) automation before the systems are built.

The implications are significant. Most notable is the proof that the subjective reliability assessment of automation places stringent dual conditions of its own on the performance of both the automation *and* the user, not just on the automation as usually assumed. Much the same may be true of trust in other modes of human-machine interaction. The particular mode envisioned imposes feasibility constraints on user and automation alike. Also of note is the proof that critical tasks that users find difficult to do correctly, with certainty—tasks, that is, for which automation may at first glance seem to be especially necessary—are precisely the tasks for which supervised automation is *not* feasible. The automation will as a rule appear unreliable to the user, and it will never be relied upon, except possibly as a measure of last resort. Other non-supervisory modes of human-machine interaction, or at least non-subjective means for reliability assessment, must be considered for such tasks.

References

1. M. T. Dzindolet (2002), "The perceived utility of human and automated aids in a visual detection task,", *Human Factors*, **44**, pp. 74-94.

2. R. Parasuraman (1997), "Humans and automation: Use, Misuse, Disuse, Abuse," *Human Factors*, **39**, pp. 230-253

3. D. A. Wiegmann, A. Rich, and H. Zhang (2001), "Automated diagnostic aids: the effects of aid reliability on users' trust and reliance," *Theoretical Issues in Ergonomic Science*, **2**, pp. 352-367.

4. M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck (2003), "The role of trust in automation reliance," *International Journal of Human-Computer Studies*, **58**, pp. 697-718.

5. M. T. Khasawneh, S. R. Bowling, X. Jiang, A. K. Gramopadhye, and B. J. Melloy (2004), "Investigating Human Perception and Trust Due to Changes in Hybrid Inspection System Parameters," *Proceedings of the Industrial Engineering Research Conference,* Houston Texas, May, 2004.

6. J. D. Lee and K. A. See (2004), "Trust in Automation: Designing for Appropriate Reliance", *HUMAN FACTORS*, **46**, No. 1, Spring 2004, pp. 50–80.

7. J-Y. Jian, A. M. Bisantz, and C.G. Drury (2000), "Foundations for an empirically determined scale of trust in automated systems," *International Journal of Cognitive Ergonomics*, **4**, 53-71.

8. M. S. Cohen, R. Parasuraman, and J. T. Freeman (1998), "Trust in decision aids: a model and its training implications," Proc. 1998 Command & Control Research & technical Symposium, DoD C4ISSR Cognition Research Program.

9. B. G. Silverman (1992), "Human-computer collaboration", *Human-Computer Interaction*, 1992, **7**, pp. 165-196.

10. J. Lee and N. Moray (1992), "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, **35**, pp. 1243-1270.

11. B. M. Muir (1987), "Trust between humans and machines, and the design of decision aids," *Int. J. Man-Machine Studies*, **27**, pp.527-539.

12. H.L. Van Trees (1968), *Detection, estimation, and modulation theory*, John Wiley and Sons, New York.

13. D. E. Green and J. A. Swets (1988), *Signal detection theory and psychophysics*, Penninsula Publishing, Los Altos California

14. R.E. Walpole and R.H. Myers (1978) *Probability and Statistics for Engineers and Scientists*, 2nd Edition, MacMillan Publishing, New York.

15. E.T. Jaynes (2003) *Probability Theory: The logic of science*, Cambridge University Press

16. D. Howie (2002) *Interpreting probability: controversies and developments in the early twentieth century*, Cambridge University Press, Cambridge

17. A. Papoulis (1965), *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill Book Co., 1965.

18. R.T. Kessel and V.L. Myers (2005), "Discriminating man-made and natural objects in sidescan sonar imagery: human versus computer recognition performance", Proc SPIE, Defence & Security Symposium, Orlando, Florida, Mar 2005, Automatic Target Recognition XV.

19. R.T. Kessel (2003), "Texture-based discrimination of man-made and natural objects in sidescan sonar imagery", Proceedings of the SPIE, Vol.5096, AeroSense 2003, Orlando, Florida, pp.21-25, April 2003

20. V.H. Poor, *An introduction to signal detection and estimation*, Springer-Verlag, London, 1988

21. R. Breton and R. Rousseau (2003), "Situation Awareness: A review of the concept and its measurement," Defence R&D Canada, Technical Report, DRDC Valcartier TR 2001-220.

22. M.V. Trevorrow, A.M. Crawford, J. Fawcett, R.T. Kessel, and T. Miller (2001), Synopsis of survey data collected during Q-260 Maple 2001 Sea-trials with CFAV and NRV Alliance, Defence R&D Canada, Technical Memorandum, DREA TM-2001-172, Nov 2001.

23. R.T. Kessel (2002), "*Using sonar speckle to identify regions of interest and for mine detection*," Proceedings of the SPIE, Vol. 4742, AeroSense 2002 Conference, Orlando, Florida, April 2002.

9 Internal Distribution List

DRDC Atlantic TM 2005-155

1 – Director General

1 – Deputy Director General

3 – Document Library

1 – Head/MICS

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

1 – B. Nguyen

1 – J. Crebholder

1 – J. Hiltz

1 – J. Daniels

1 – J. Sildam

Total: 21

10 External Distribution List

DRDC Atlantic TM 2005-155

1 – DRDKIM

1 – DRDKIM (unbound copy)

1 – DRDC

3 –DRDC Toronto

Attn: R. Pigeau

S. McFadden

5 – DRDC Valcartier

Attn: R. Breton

R. Rousseau

2 – DRDC Ottawa

Attn: C. Helleur

G. Geling

1 – Commanding Officer,

MARLANT, Trinity

1 – Commanding Officer,

MARPAC, Athena

1 – Ronald T. Kessel (author)

NATO Undersea Research Centre

Viale San Bartolomeo 400, 19138 La Spezia (SP), ITALY

1 – V. L. Myers,

NATO Undersea Research Centre

Viale San Bartolomeo 400, 19138 La Spezia (SP), ITALY

Total: 17

This page intentionally left blank.

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (the name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's report, or tasking agency, are entered in section 8.)

DRDC Atlantic

2. SECURITY CLASSIFICATION (overall security classification of the document including special warning terms if applicable).

UNCLASSIFIED

3. TITLE (the complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S,C,R or U) in parentheses after the title).

Apparent Reliability: The conditions for reliance on supervised automation

4. AUTHORS (Last name, first name, middle initial. If military, show rank, e.g. Doe, Maj. John E.)

Ronald T. Kessel

5. DATE OF PUBLICATION (month and year of publication of document)	6a. NO. OF PAGES (total containing information Include Annexes, Appendices, etc).	6b. NO. OF REFS (total cited in document)
July 2005	39 (approx.)	22

7. DESCRIPTIVE NOTES (the category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered).

Technical Memorandum

8. SPONSORING ACTIVITY (the name of the department project office or laboratory sponsoring the research and development. Include address).

Defence R&D Canada – Atlantic
PO Box 1012
Dartmouth, NS, Canada B2Y 3Z7

9a. PROJECT OR GRANT NO. (if appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant).	9b. CONTRACT NO. (if appropriate, the applicable number under which the document was written).
11hd	

10a ORIGINATOR'S DOCUMENT NUMBER (the official document number by which the document is identified by the originating activity. This number must be unique to this document.)	10b OTHER DOCUMENT NOs. (Any other numbers which may be assigned this document either by the originator or by the sponsor.)
DRDC Atlantic TM-2005-155	

11. DOCUMENT AVAILABILITY (any limitations on further dissemination of the document, other than those imposed by security classification)

(X) Unlimited distribution
( ) Defence departments and defence contractors; further distribution only as approved
( ) Defence departments and Canadian defence contractors; further distribution only as approved
( ) Government departments and agencies; further distribution only as approved
( ) Defence departments; further distribution only as approved
( ) Other (please specify):

12. DOCUMENT ANNOUNCEMENT (any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected).

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).

When experts are given automation for assistance, they naturally observe its actions and judge its reliability for themselves. Thus the automation is supervised, for a time at least, when it must win the trust of its users before it can serve effectively through routine use. Lack of trust, on the other hand, results in under utilization or total disuse of the automation. This occurs so often in practice that researchers investigating trust in automation now speak of a prevailing “bias toward self reliance” among users. Here a quantitative analysis of apparent reliability is applied to supervised automation. It is shown that subjective reliability assessment imposes minimum standards of proficiency on the user of the automation. In effect, it takes a high-performance user to recognize the high-performance of automation in action. This is demonstrated here using the applied mathematics of operation analysis. Minimum performance standards for both user and automation are derived that must be met before it is plausible that the automation may appear reliable to its user. The standards are surprisingly stringent for critical applications, creating a natural barrier—or bias—against reliance that is unavoidable with supervised automation. These are conditions of feasibility that must be met to avoid failure in disuse for supervised automation. Special attention is given to automation within a constant false alarm rate (CFAR) concept of operations, and for command and control (C2) and situation awareness.

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 it not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title).

TRUST IN AUTOMATION, HUMAN-MACHINE INTERACTION, AUTOMATIC TARGET RECOGNITION, DECISION AIDS, SUPERVISED AUTOMATION, REMOTE MINEHUNTING SYSTEM (RMS) TDP

This page intentionally left blank.

Defence R&D Canada

Canada's leader in defence
and National Security
Science and Technology

R & D pour la défense Canada

Chef de file au Canada en matière
de science et de technologie pour
la défense et la sécurité nationale

www.drdc-rddc.gc.ca