

IMPERFECT DIAGNOSTIC AUTOMATION: AN EXPERIMENTAL EXAMINATION OF PRIORITIES AND THRESHOLD SETTING

Christopher D. Wickens, Stephen R. Dixon, and Nicholas Johnson
University of Illinois, Aviation Human Factors Division
Savoy, Illinois

Diagnostic automation, such as alarms, alerts, or automatic target recognition systems can vary in their reliability and threshold setting, the latter influencing the balance between misses and false alarms. This experiment examined the implications of both of these, when the system detected military targets in parallel with the human operator in an unmanned air vehicle simulator. Unlike previous investigations of this paradigm in a dual task setting, the automation diagnosis task was very difficult, and priority between this task and a concurrent task was explicitly varied, along with the threshold setting. The results revealed that: (a) reliability as low as 0.6 aided human performance. (b) the effects of the automation threshold shift could be partially modeled in terms of the classic reliance-compliance dimensions of automation dependence, at both task priority settings. (c) The priority effects were more pronounced with miss-prone automation. (d) directing attention to false-alarm prone automation actually degrades human-system performance.

INTRODUCTION

Across many domains, diagnostic automation, in which automation is asked to discriminate two (or more) states of the world represented by fuzzy data, can unburden the human operator. Such automation has two general functions: (1) in a multi task environment, it allows the operator to divert (usually visual) attention to concurrent tasks. The alarm or alert is typical here (Pritchett, 2001). (2) In single task situations, it may increase accuracy of detecting difficult-to-see signals and events (e.g., luggage screening decision aids, Goh, Wiegmann & Madhavan, 2005, or automatic military target recognizers; Maltz & Shinar, 2003; Yeh & Wickens, 2001). When examined within the framework of signal detection theory, with two states of the world to be discriminated, two features of such a representation are important, corresponding to the two parameters of signal detection: the sensitivity (d') and the response criterion (beta), or threshold setting of the diagnostic automation (e.g., the alert threshold).

First, concerning sensitivity, the parameter assesses how good automation is in resolving signals from non-signals, a term that corresponds to automation reliability. There are two important reasons why this reliability may drop well below

1.0. First, such systems may be asked to operate when the quality of data is poor, and algorithms are incapable of fully compensating (e.g., target recognition of fuzzy visual images [Maltz & Shinar, 2003] or medical diagnosis on the basis of imperfect symptoms). Second, such systems are often asked to make predictions in an inherently unpredictable world (e.g., weather predictions, air or ground vehicle collision predictions, Xu, Rantanen & Wickens, in press).

Second, regarding beta, or the *diagnostic threshold*, assuming that there is some asymmetry in signal versus no-signal events (in cost and/or probability), it matters how the threshold is set, and this can be determined in part from the “optimal beta” which trades off the costs of automation misses vs. false alarms. However if the human too can monitor the raw data, in parallel with automation (Sorkin & Woods, 1985), then setting of automation beta has important implications for human multi-task performance, as rooted in the “reliance-compliance” distinction of automation dependence proposed by Meyer (2001, 2004; see also Wickens, Dixon & Ambinder, 2006).

With regard to sensitivity (reliability), a recent meta-analysis integrated studies that compared performance of a human aided by imperfect diagnostic automation, with unaided human

performance detecting the same signals (Wickens & Dixon, in press). The study revealed that: (a) automation benefited overall performance relative to baseline, as long as its reliability was above about 0.70. At lower levels of reliability operators appeared to depend on automation even if they would be better off without it (the cement life-preserver analogy). (b) The linkage of performance to automation reliability was more pronounced in studies in which workload was high, either with difficult diagnostic tasks, or in dual task situations, suggesting greater automation dependence in these circumstances. (c) In dual task studies, performance of the automation task was much more degraded by loss of reliability, than was performance of the concurrent task. It is this third issue that is a major focus of the current study.

We hypothesize that one possible explanation for this insensitivity of concurrent task performance to the costs of imperfect diagnostic automation, is that the human operator intrinsically treats an automated task as “secondary” and a concurrent task as “primary”, thereby allocating the latter the necessary resources to preserve its performance at a constant level, even in the face of automation imperfections. However an alternative hypothesis is that people simply cannot re-allocate the resources from the automated task to the concurrent task even if explicitly requested to do so, as if separate resources are involved in each (Navon & Gopher, 1979). That is, performance would be insensitive to priority instructions. Since none of the studies evaluated explicitly manipulated task priority nor provided such instructions, we could not confirm which hypothesis was true, and therefore in the current experimental investigation we varied priorities, involving an unmanned air vehicle simulation with imperfect automated target recognition, and concurrent task, in an unmanned air vehicle simulation.

Of course, the reliability of diagnostic automation can be reduced in one of two ways, as a function of the threshold setting. This reduction can be manifest to varying degrees by either increasing misses or increasing false alarms. Work by Meyer (2001, 2004) and subsequent elaboration by Dixon and Wickens (in press) has revealed two different, partially independent “syndromes” of performance effects engendered by varying the threshold.

Compliance, the state reflected when the alarm sounds, is affected by the false alarm rate, and will be degraded as this FA rate increases. The most prominent manifestation of reduced compliance is the (1) “cry wolf” effect, whereby both true and false alarms are responded to more slowly, or not at all. **Reliance**, the state reflected when the alarm is silent, is affected by miss rate. A high miss rate, reducing reliance forces the vigilant operator to spend more time inspecting the raw data, at the expense of attending to other tasks in order to assure that automation has not missed a signal. This has two behavioral symptoms characterizing the reliance syndrome: (2) those infrequent automation misses **are** better detected than when automation miss rate is low (and reliance is high). This is related to reduced complacency (3) concurrent task performance will degrade, as resources are re-allocated to raw-data monitoring.

These three effects have generally been replicated in dual task paradigms, as the automation alert threshold is varied, although effect 3 is not always obtained, and sometimes false alarm prone automation actually hurts concurrent task performance more than miss-prone automation (Dixon, McCarley & Wickens, 2006). However, in all these studies, the automation task has been assumed, implicitly to be of secondary importance. Furthermore, in most studies the unaided automation task has been relatively easy to perform in single task conditions, with the purpose of automation being to better support time sharing (rather than aiding difficult discrimination). Thus a second purpose of the current investigation was to assess if similar reliance-compliance syndromes were observed when the automation task is emphasized, as well as when it is a difficult unaided task. To do so, we varied the threshold setting to produce miss-prone (low reliance) versus false-alarm prone (low compliance) automation, in addition to varying priority between the automated and the concurrent task. Three effects were explicitly examined:

1. To choose between the two attention re-allocation hypotheses, we examined whether priority instructions had a robust effect on performance of both (the automated and concurrent) tasks, or a minimal effect.

2. To examine reliance and compliance, we were interested in the extent to which the general reliance-compliance syndrome was expressed when the automation task was explicitly emphasized (as well as whether it was replicated when the concurrent task was emphasized).
3. The potential relation **between** the two factors, threshold setting and priorities had several different interpretations. If additivity is observed, it suggests the robustness of the reliance-compliance distinction across different attention sets. On the other hand, the predicted form of an interaction that might be observed, is that the effects of priority would be more pronounced on the indices of reliance (particularly concurrent task performance) than those of compliance; this because reliance effects are themselves very much mediated by the allocation of resources between the automated and the concurrent task(s).

METHODS

The UAV simulation employed is described in detail in Dixon and Wickens (in press), and Wickens et al. (2006). Pilots flew a simulated UAV to various waypoints, manipulating a joy stick and, when detecting a ground target on their look-down 3D image display, they zoomed in for closer inspection. These ground targets or **targets of opportunity (TOO)** were camouflaged and difficult-to-detect, and this task was therefore supported by an imperfect automated detection system. We label this to TOO task. Concurrently with flying the UAV and monitoring its course, pilots had to monitor a display of parameters of system health and detect departures that indicated a failure. This was the “concurrent task”. In prior research (Dixon & Wickens, in press), we had automated this system failure (SF) detection task, and rendered the TOO task as the concurrent task. Here we reversed the assignment. In a between subjects-design, pilots were asked to emphasize either the system monitoring (SF) task, or the automation-aided TOO task. Emphasis was reinforced with monetary payoffs. The latter task had an automation reliability of 0.60, and was

implemented with a beta (detection threshold) to either produce a high miss rate or a high false alarm rate. Our interest was in establishing whether, when the automated task was explicitly designated as **primary**, its performance would no longer suffer the costs of imperfect automation. The baseline data of unaided manual performance was estimated from the results of two prior studies (Dixon & Wickens, in press; Wickens, Dixon, Goh & Hammer, 2005), using the same paradigm. Eight participants were assigned to each condition, and each participant flew ten waypoint legs. Participants were paid bonuses of up to \$20 for adhering to priority instructions.

RESULTS

Table 1 provides the results of the two major tasks, and the four conditions (two factors) of interest, and also includes the baseline data from the prior study.

Table 1. Experimental data. MA = Miss-Prone. FAP = False Alarm Prone.

MEASURE	PRIORITY (EMPHASIS)				Baseline
	TOO Task		SF Task		
	MP	FAP	MP	FAP	
<u>TOO (Auto-task)</u>					
RT (sec)	3.70	5.00	9.00	9.90	7.00
Error	0.08	0.19	0.14	0.10	0.30
<u>SF (concurrent)</u>					
RT (sec)	1.40	2.20	1.30	1.00	3.50

For the TOO detection (automated task) RT, reflecting compliance, the ANOVA revealed an expected effect that task emphasis improved (shortened) TOO detection by an average of 5 sec ($F_{1,7} = 32.44, p < .01$). There was no significant effect of threshold, although RT in the false-alarm (FA) prone condition was 1.5 sec longer, reflecting a “cry wolf” effect (an effect that was marginally significant when the TOO-emphasis condition was analyzed separately, $p < .07$). There was no significant interaction.

For TOO error rate (proportion of missed targets), neither priority nor threshold had a significant effect, but the interaction between the

two was significant ($F_{1,27} = 5.16, p < .05$). Under TOO emphasis instructions, FA prone automation hurt performance (reflecting the “cry wolf” effect), but under SF (concurrent task) emphasis this effect vanished. Another way of considering this interaction is that, in the FA prone condition, emphasizing the automated TOO detection task, actually produced **more** errors of detection.

The SF task, reflecting reliance, was consistently performed at perfect accuracy, so only RT effects were analyzed. Here priority affected RT in the expected direction ($F_{1,27} = 15.35$) with shorter RT to the task when it was emphasized. As with the TOO task there was no significant effect of threshold, but the interaction was significant ($F_{1,27} = 15.53, p < .01$). This interaction revealed the expected lengthening of concurrent task RT under miss-prone automation (reflecting the allocation of visual attention to the TOO task), when the concurrent task was emphasized, but a less anticipated lengthening of RT under FA-prone automation when the automated task itself was emphasized. This was in spite of the fact that in this condition we also saw degraded performance of the automated TOO task (increased error) reflecting the particularly lethal combination of emphasis on a FA-prone automated task.

While the system failure task was not hurt by miss-prone automation, a second measure of residual attention, a measure of memory failure for mission critical information, **was** substantially and significantly ($F_{1,27} = 54.95$) degraded in this direction.

Interestingly, as shown in Table 1, across all three dependent variables, performance was consistently better (TOO Ac and SF RT), or no worse (TOO RT) than baseline performance with no automation, in spite of the relatively low (0.60) level of reliability.

DISCUSSION

We set out to examine three effects, related to hypotheses to be tested. First, we asked whether priority instructions would induce a pronounced tradeoff in task performance. If so, then the failure to find that concurrent tasks had been much degraded by automation imperfections in prior studies (Wickens & Dixon, in press), could be

attributed to the implicit emphasis on the concurrent task in those studies. Finding in the current data that such a tradeoff was both significant in both tasks, and substantial in magnitude (for the automated TOO task), allows us to confirm the above interpretation. Resources can be shared and re-allocated between the two tasks, if requested.

Second, we asked the extent to which the reliance-compliance “syndrome” was replicated when the automated task was explicitly emphasized. Here the answer was somewhat equivocal. While the TOO task showed the expected pattern of “compliance” effects (“cry wolf” degraded detection performance: detection accuracy and speed loss under FA-prone automation), concurrent task performance was actually **more**, rather than less disrupted by FA-prone automation. In this regard, the effect on concurrent task performance actually replicates patterns of data found elsewhere (Dixon, McCarley & Wickens, 2006), that suggest, in contrast to the pure reliance-compliance model, that FA-prone automation is quite disruptive to concurrent tasks.

The finding that a FA-producing threshold setting hurts the concurrent task, was not observed when the SF (concurrent task) was emphasized, so this invites us to inspect the interaction between the two independent variables (hypothesis 3) which was, indeed significant for two of the three dependent variables. Generally speaking, the **form** of the interaction was as predicted. That is, all three DV’s showed the expected priority-induced tradeoff of performance when the threshold setting led to miss-prone automation, and therefore, when visual capacity necessary to monitor the raw data was at a premium (see Wickens, Dixon, Goh & Hammer, 2005, for direct link of visual capacity to visual scanning in this context). This priority tradeoff was less consistent with FA-prone automation, suggesting that visual capacity is a key player underlying task priority instructions here.

Another observation highlighted by the pattern of interactions, is the counter-intuitive phenomenon that directing attention **to** the automated task, when it is given a FA-prone threshold setting, actually degrades performance (accuracy) on that task. What we assume has happened here, is that such attention becomes focused not just on the raw data, but on the properties of the alarm system itself. Its noticeably

salient and poor functioning here, highlights its unreliable properties to the human detector, and amplifies the mistrust experienced, leading to the strong “cry wolf” behavior.

A final observation in the present data not to be overlooked, was the overall improvement in performance of 0.60 reliable automation, on both the automated task and the concurrent task, relative to that for a non-automated baseline (right column of Table 1). Why we obtained improvement here, whereas based on the meta-analysis conducted earlier by Wickens and Dixon (in press) one would predict that a 0.6 reliability level would degrade that performance, can be related to one particular feature of the current experiment that was unique compared to the other studies examined. The automated task was both demanding **and** carried out under multi-task conditions. This feature thereby amplified the demand for resources in an unaided condition, and substantially lowered the level of performance possible there.

ACKNOWLEDGMENTS

The authors wish to acknowledge the research support of contract ARMY MAD 6021.000-01 from Micro-Analysis and Design and the Army Research Laboratory. Dave Dahn and Marc Gacy were the scientific/technical monitors. The opinions expressed in this chapter are those of the authors and do not necessarily reflect those of the US Army.

REFERENCES

- Dixon, S.R., McCarley, J., & Wickens, C.D. (2006). How do automation false alarms and misses affect operator compliance and reliance. *HFES Proceedings*.
- Dixon, S.R. & Wickens, C.D. (in press). Automation reliability in unmanned aerial vehicle flight control. *Human Factors*.
- Goh, J., Wiegmann, D., & Madhavan, P. (2005). Maltz, M., & Shinar, D. (2003). New alternative methods in analyzing human behavior in cued target acquisition. *Human Factors*, 45(2), 281-295.
- Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors*, 43(4), 563-572.
- Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, 46(2), 196-204.
- Navon, D. & Gopher, D. (1979). On the economy of the human processing system. *Psychological Review* 86, 254-285.
- Pritchett, A. (2001). Reviewing the role of cockpit alerting systems. *Human Factors and Aerospace Safety*, 1, 5-38.
- Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors, a signal detection analysis. *Human-Computer Interaction*, 1, 49-75.
- Wickens, C.D. & Dixon, S.R. (in press). Is there a magic number 7 (to the minus 1)? The benefits of imperfect diagnostic automation. *Theoretical Issues in Ergonomics Science*.
- Wickens, C.D., Dixon, S.R., & Ambinder, M.S. (2006). Workload and automation reliability in unmanned air vehicles. In N. J. Cooke, H. Pringle, H. Pedersen, & O. Connor (Eds.), *Advances in human performance and cognitive engineering research, Vol. 7, Human factors of remotely operated vehicles* (pp. 209-222). Elsevier Ltd.
- Wickens, C.D., Dixon, S.R., Goh, J. & Hammer, B. (2005). Pilot dependence on imperfect diagnostic automation in simulated UAV flights. *Proceedings of the 13th Annual International Symposium of Aviation Psychology*.
- Xu, X., Wickens, C.D., & Rantanen, E.M. (in press). Effects of conflict alerting system reliability and task difficulty on pilots' conflict detection with cockpit display of traffic information. *Ergonomics*.
- Yeh, M., & Wickens, C. D. (2001). Attentional filtering in the design of electronic map displays: A comparison of color-coding, intensity coding, and decluttering techniques. *Human Factors*, 43(4), 543-562.