Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis

Linda Onnasch, Technische Universität Berlin, Berlin, Germany, **Christopher D. Wickens**, Alion Science and Technology, McLean, Virginia, USA, **Huiyang Li**, University of Michigan, Ann Arbor, Michigan, USA, and **Dietrich Manzey**, Technische Universität Berlin, Berlin, Germany

Objective: We investigated how automation-induced human performance consequences depended on the degree of automation (DOA).

Background: Function allocation between human and automation can be represented in terms of the stages and levels taxonomy proposed by Parasuraman, Sheridan, and Wickens. Higher DOAs are achieved both by later stages and higher levels within stages.

Method: A meta-analysis based on data of 18 experiments examines the mediating effects of DOA on routine system performance, performance when the automation fails, workload, and situation awareness (SA). The effects of DOA on these measures are summarized by level of statistical significance.

Results: We found (a) a clear automation benefit for routine system performance with increasing DOA, (b) a similar but weaker pattern for workload when automation functioned properly, and (c) a negative impact of higher DOA on failure system performance and SA. Most interesting was the finding that negative consequences of automation seem to be most likely when DOA moved across a critical boundary, which was identified between automation supporting information analysis and automation supporting action selection.

Conclusion: Results support the proposed cost–benefit trade-off with regard to DOA. It seems that routine performance and workload on one hand, and the potential loss of SA and manual skills on the other hand, directly trade off and that appropriate function allocation can serve only one of the two aspects.

Application: Findings contribute to the body of research on adequate function allocation by providing an overall picture through quantitatively combining data from a variety of studies across varying domains.

Keywords: degree of automation, operator performance, workload, situation awareness, human-automation interaction, function allocation

Address correspondence to Linda Onnasch, TU Berlin, Marchstr. 12, F7, D-10587 Berlin, Germany; e-mail: linda.onnasch@tu-berlin.de.

HUMAN FACTORS
Vol. 56, No. 3, May 2014, pp. 476–488
DOI: 10.1177/0018720813501549

INTRODUCTION

It has been long known that automation can both hurt and benefit human performance (e.g., Bainbridge, 1983; Ephrath & Young, 1981; Kessel & Wickens, 1982; Rasmussen & Rouse, 1981; Sheridan, 2002; Wickens & Kessel, 1979, 1981; Wickens, Mavor, Parasuraman, & McGee, 1998; Wiener & Curry, 1980). This cost–benefit trade-off is particularly prominent when automation is imperfectly reliable. Automation infrequently fails, either due to hardware or software failures, or it fails to achieve desired outcomes simply because a functionality is used in circumstances for which it was not intended. For fielded automation, it is almost always the case that routine or "nonfailure" performance substantially exceeds unaided human performance and/or the automation assistance lowers workload. If it did not, the system would not be fielded or considered useful.

However, on those infrequent occasions when automation does fail, the effects on joint human-machine system performance may be catastrophic. These catastrophic effects may result from human's reduced monitoring of highly reliable automation at the time it fails, trusting it too much (Parasuraman & Riley, 1997) and losing situation awareness (Endsley & Kiris, 1995). This is sometimes described as a form of complacency (Parasuraman, Molloy, & Singh, 1993) or an automation-induced decision bias (Mosier & Skitka, 1996). Indeed, operators occasionally over-rely on automation and exhibit complacency because the highly (but not perfectly) reliable automation functioned properly for an extended period prior to this first failure (Parasuraman et al., 1993; Parasuraman & Manzey, 2010; Yeh, Merlo, Wickens, & Brandenburg, 2003).

Going beyond the issues of highly reliable automation, Endsley and Kiris (1995) and Miller and Parasuraman (2007) have pointed out that

also the "competence" of the automation must be considered. The more support an automated system provides, the higher the risk of adverse effects on human performance (e.g., complacency, loss of situation awareness, skill degradation), and the greater the likelihood of catastrophic consequences when it fails. This trade-off, in which *more automation yields better human-system performance when all is well but induces increased dependence, which may produce more problematic performance when things fail*, will be of critical importance to this review of the performance effects of different degrees of automation. We might refer to this conventional wisdom about automation as the "lumber jack effect"; as applied to trees in the forest, "the higher they are, the farther they fall." Of importance, the choice of whether or not, and to what degree, to automate a particular function should involve a trade-off between the benefits of reliable automation and the expected costs (true costs × probability of failure) of automation failures (Sheridan & Parasuraman, 2000).

The "routine-failure" trade-off is complicated by the fact that "automation" is not an all-or-none concept, as it was often assumed to be in the classic human–machine task allocation analyses (e.g., the "Fitts List"; for a critique of those analyses, see Dekker & Woods, 2002; Parasuraman, Sheridan, & Wickens, 2008). Instead, one can think of varying *levels of automation* as first put forth by Sheridan and Verplank (1978; see also Endsley & Kiris, 1995). This continuum can be jointly defined by the amount of automation autonomy and responsibility (highest at the highest level) and the amount of human physical and cognitive activity (highest at the lowest level). For example, at the highest level, the automation can perform a decision task completely autonomously; at a lower level, it can choose (and possibly execute) an option unless the human vetoes; and at an even lower level, it may simply offer the human a selection of options.

More recently, Kaber and Endsley (2004) and Wickens et al. (1998) put forth the idea that automation could also be categorized according to the *stage of information processing* that it accomplished. Elaborating on this concept, Parasuraman, Sheridan, and Wickens (2000) and Wickens et al. (1998) proposed a concept in which automation could *filter* information from the environment (Stage 1: information acquisition), *integrate* this information, as when forming an assessment based on several sources of information (Stage 2: information analysis), choose or *decide* on an action based on the assessment (Stage 3: decision and action selection), and *implement* the action via a typically manual activity (Stage 4: action implementation). Within each stage, varying levels could be defined. For example, as described earlier, Sheridan and Verplank (1978) define multiple levels at Stage 3. In so doing, automation can be said to offload, assist, or replace human performance at corresponding stages of human information processing (e.g., automation filtering at Stage 1, can assist human selective attention).

As an example, health care automation may (a) alert (call attention to) abnormal patient symptoms, (b) integrate these symptoms to form an intelligent diagnosis of the patient condition, (c) recommend a treatment or course of action based on the diagnosis, and (d) carry out the action as, for example, with an automated infusion pump. In applying this taxonomy, where any given stage can function at various levels, it is important to note the quasi-independence of levels across the various stages. Thus, for example, a totally automated diagnosis may be followed by a fully manual (physician chosen) course of action, just as a fully manual diagnosis may trigger an automated choice of treatment.

Considering that "more automation" can be represented both by higher levels within a stage, and, typically, later stages (which, in the literature, are typically preceded by automation at earlier stages), we assume, in the analysis given later, that these two dimensions (higher levels and later stages) increase the *degree of automation* (DOA; e.g., Manzey, Reichenbach, & Onnasch, 2012). More specifically, it is assumed that differences between automated (support) systems representing automation of different stages and levels can be described on an ordinal scale reflecting the amount of automated support that is provided. The main assumption underlying this concept as we define it asserts that assessment of "more versus less automation" can be based on dominance relationships,

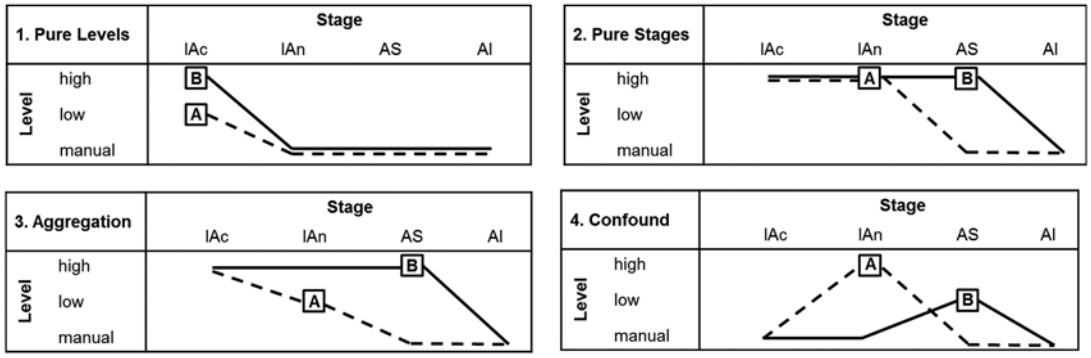

Figure 1. Four cases comparing degree of automation (DOA) across stages, that is, information acquisition (IAc), information analysis (IAn), action selection (AS), and action implementation (AI), and levels, that is, high, low, and manual. Two systems compared are represented by dashed (System A) and solid lines (System B). For Cases 1 to 3, System B always represents "more automation" in a distinct way (e.g., by higher levels or by higher stages). Case 4 represents a confound where "what is more automation" cannot be defined.

as long as the following three postulates are agreed on. That is, all other factors held equivalent, (a) a higher *level* of automation constitutes "more automation," (b) a later *stage* of automation constitutes "more automation," and (c) as a consequence, a combination of higher levels and a greater *number* of stages at which automation is implemented constitutes "more automation." As will be shown later, applying this reasoning enables an unambiguous rank ordering of automated systems that have been analyzed and compared in different studies and domains. It further seems to reflect the implicit or explicit assumptions that researchers in the field whose data are employed in the current analysis usually apply when comparing their systems in terms of some concept of "more or less automation."

We illustrate this within Figure 1, which, for simplicity, presents examples of the four-stage model of Parasuraman et al. (2000) with only three levels per stage. Each of the four cases contains two automation systems, A and B, which are compared within one experiment. The automation characteristic of each of these systems is characterized by a profile of levels across the stages. The first three cases also represent the three postulates given earlier. Case 1 ("Pure Levels") represents different levels within a stage. Case 2 ("Pure Stages") represents different stages at the same level. Case 3 ("Aggregation") represents an earlier stage and lower level versus a later stage and higher level. Case 4 ("Confound") represents an earlier stage and higher level versus a later stage and lower level (i.e., a "trade-off" between stages and levels).

We argue that, to the extent that the three postulates given earlier are agreed on, comparisons 1 to 3 clearly represent contrasts between systems with more (System B) versus less (System A) automation, as defined on an ordinal scale. These relationships characterize all of the studies we have reviewed in which authors have invoked a phrase like "more automation." Case 4 is an important exception. Here, there is a trade-off between later stages and higher levels. It is impossible to assess a relative DOA unless both stages and levels are expressed on an interval or ratio scale, and we have no confidence that this has or even can be done. But none of the studies analyzed later in this article involved such a comparison.

Thus, in our analysis, DOA is a useful ordinal metric explicitly available and used to compare two or more systems (or experimental conditions), specifically for the purpose of examining the trade-offs inherent in the lumberjack analogy.

Within this DOA concept, the discrete trade-off described earlier (i.e., automation supports better performance in routine situations but is problematic when automation breaks down) can be expressed as a more continuous trade-off, as illustrated in Figure 2. The two primary performance functions in this figure (heavy lines)

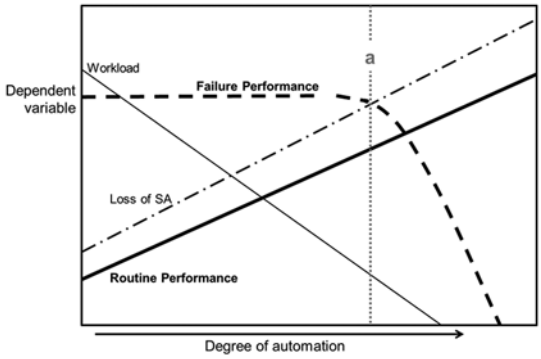

Figure 2. Trade-off of variables, with degree of automation.

indicate that, as the DOA increases, routine performance will improve but performance under failure will decline. This relationship is expressed intuitively by the lumberjack analogy. Prior research has found that the lumberjack analogy appears to apply to the continuum of automation reliability of alerting systems (Wickens & Dixon, 2007). We examine here the extent to which this may also apply to DOA. Furthermore, our interest lies in whether DOA also has a systematic impact on workload and situation awareness (Endsley & Kiris, 1995). Indeed, as discussed later, to the extent that loss of situation awareness may be due to both an increase in automation reliability and an increased DOA, it is plausible to assume that the lumberjack analogy may apply to the latter case (see also Wickens, 2008b).

Thus, Figure 2 also depicts the hypothetical trade-off between the two secondary variables, workload and loss of situation awareness (the two lighter lines). With a higher DOA, the workload imposed by the automated task is progressively reduced, almost by definition, since if the automation is doing more cognitive/physical work, the human is doing less. This holds at least if the automation is properly designed and does not provide new effortful challenges and tasks related to its engagement and monitoring (e.g., Kirlik, 1993; Wiener, 1988). If this is granted, the automation enables the human to allocate more attention to other concurrent tasks (Wickens, 2008a); but if the human does so (i.e., exploits the lower workload to enhance overall productivity), the resulting reduction of attention to the tasks served by automation could have consequences expressed in the loss of situation awareness (LSA), that is, loss of awareness of the state of the system supported by automation (e.g., lack of altitude awareness in the autopilot-controlled cockpit) or the state of the automation itself (i.e., poor mode awareness of the flight management system; e.g., Sarter, 2008).

Even though there is broad consensus in the understanding of the concept of situation awareness (SA) as it has been defined by Endsley's (1988) three levels model, the operational definitions used to assess SA in different studies are considerably diverse. In the context of the present research, we consider both direct as well as indirect measures as indicators of SA. Direct indicators are derived from conventional methods to assess SA, like the situation awareness global assessment technique (SAGAT; Endsley, 2000). Indirect measures of (a loss of) SA include any performance consequences in interaction with automation that point to a lack of information sampling, a lack of understanding, or a lack of correctly anticipating the behavior of the automation (e.g., errors of omission or commission; Mosier & Skitka, 1996).

The hypothetical trade-offs depicted in Figure 2 are critical for task allocation because these trade-offs may not be linear, and in some cases a "flat" function may allow strong recommendations for the optimal task allocation (Wickens, 2008a). For example, if the costs of imperfect automation (mediated by LSA) remain flat up to a high DOA (as shown in Figure 2), then the recommended DOA would be at Point a in the figure: maximum routine performance and lowest workload, without sacrificing failure performance.

Earlier research contrasting human performance with and without automation support has focused only on what has been referred to as "out-of the-loop unfamiliarity" effects without varying the levels or stages of automation (e.g., Crossman, 1974; Ephrath & Young, 1981; Kessel & Wickens, 1982; Wickens & Kessel, 1979, 1980, 1981). These studies provide evidence for automation-induced performance consequences but do not allow for any conclusion about the relationship to different degrees of automation. The latter issue attracted little research until the

early 1990s (for early examples, see, e.g., Crocoll & Coury, 1990; Layton, Smith, & McCoy, 1994). Yet since then at least a limited number of studies have become available that have collected empirical data on effects of two or more different DOAs on workload and/or SA (e.g., Endsley & Kiris, 1995; Kaber, Onal, & Endsley, 2000; Lorenz, Di Nocera, Röttger, & Parasuraman, 2002a; Sarter & Schroeder, 2001). The pattern of results of these single studies provides a somewhat mixed picture. Whereas some studies support the existence of the trade-off as defined by better routine performance but worse performance when automation fails (e.g., Sarter & Schroeder, 2001), others do not find this effect (Lorenz et al., 2002a) and still others suggest that medium levels of automation provide the best choice in terms of maintaining SA and return-to-manual performance (Endsley & Kiris, 1995) or provide an even more complex pattern of effects (Endsley & Kaber, 1999). However, due to differences in DOA levels considered, and a generally limited statistical power, the effects of single studies are inconclusive.

A more valid overall picture might be revealed by quantitatively combining data from a variety of studies across varying domains (e.g., process control, aviation), an approach analogous to a classic meta-analysis (Fadden, Ververs, & Wickens, 1998; Horrey & Wickens, 2006; Rosenthal, 1991; Wickens, Hutchinson, Carolan, & Cumming, 2013). The purpose of the current investigation is to provide such meta-analysis (a) by aggregating data from studies that compared different degrees of automation, (b) by examining the extent to which they show the postulated trade-off between normal operations and failure conditions as the DOA was manipulated, and (c) if possible, by identifying factors that may mitigate or moderate this trade-off.

METHOD

In a first step we looked for relevant studies to be included in this analysis. Sources used for this purpose included databank searches (e.g., PsycINFO), analyses of tables of contents of relevant journals (e.g., *Human Factors, Ergonomics, International Journal of Human-Computer Interaction*) and conference proceedings for the years 1990 to present, and direct contact of colleagues to identify relevant technical reports or other examples of references that were not available through publishers. Only studies that compared at least two different degrees of automation defined by the postulates given earlier, for example, either by varying the stage of automation or the number of stages or by varying the level of automation within a stage, with respect to at least one relevant performance measure, were included. Consequently, a total of 18 studies were identified and integrated in the analysis (see Table 1).

The second step included a proper quantification of the independent variable (i.e., DOA) and dependent variables (i.e., performance, workload, and SA data) as a basis for our meta-analysis approach. For each single study, the DOAs analyzed were converted into rank data with an increasing rank (beginning by rank = 1) reflecting an increasing DOA via either stages or levels corresponding to the logic described earlier. We note that none of the studies contrasted conditions with higher level/earlier stage with lower level/later stage (or vice versa), which would not be easy to rank due to lacking unambiguous a priori criteria for cross-stage comparisons of levels. Manual performance conditions were always assigned a rank of 0. Rankings were provided by one of the authors (LO) and double-checked by two of the coauthors (CW, DM).

To bring the variety of dependent measures and definitions used in the studies to a comparable level, we defined "metavariables" that were broad enough to group the data while still representing a clear definition of the concept in question (e.g., SA). As our main focus of the present study was on performance costs and benefits of automation support, we differentiated between *primary task performance* when the automation functioned properly (metavariable *routine primary task performance*, reflecting joint performance of operator and system together) and performance when there was a complete automation breakdown, that is, when operators had to resume the automated task and perform it manually again after some time of reliable automation support (metavariable *return-to-manual primary task performance*). The routine primary task performance metavariable, for example, could be realized within the

TABLE 1: Kendall's Tau for the Single Studies on the Six Metavariables With Resulting Overall Kendall's Tau and Statistics of One-Tailed *t* Tests

Study	Routine Primary Task Performance (TP)	Return-to-Manual Primary TP	Routine Secondary TP	Return-to-Manual Secondary TP	Subjective Workload	Situation Awareness
Calhoun et al. (2009)	–.816		0			0
Crocoll & Coury (1990)	.707					
Cummings & Mitchell (2007)	0					0
Endsley & Kaber (1999)	.637	.025			.804	.597
Endsley & Kiris (1995)		–.837			0	–.837
Kaber & Endsley (2004)	.6	0			–.598	.258
Kaber et al. (2000)	.316	–.408			–.775	–.632
Li et al. (in preparation)	1				–1	–1
Lorenz et al. (2002a)	.333	–.333	0	0	0	
Lorenz et al. (2002b)	.816	.333				
Manzey et al. (2012)	.913	–.816	.913		–.913	–.707
Metzger & Parasuraman (2005)	0	0	0	0	0	0
Reichenbach et al. (2011)	1	–1	0	0	0	0
Röttger et al. (2009)	.816		0		–1	
Rovira et al. (2007)	.837		.707		–.333	
Sarter & Schroeder (2001)	1					
Sethumadhavan (2009)			.707			–.913
Wright & Kaber (2005)	0				.913	
Overall τ	.509	–.337	.291	0	–.242	–.294
t-crit	1.341	–1.397	1.415		–1.363	–1.372
t	4.027	–2.176	2.024		–1.284	–1.809
p	.0005*	.031*	.042*		.056	.049*

single studies as fault identification time in a monitoring task (e.g., Lorenz, Di Nocera, Röttger, & Parasuraman, 2002b), the decision accuracy in interaction with an automated decision aid (e.g., Rovira, McGarry, & Parasuraman, 2007) or the out-of-target error when the main task was to maintain certain values in a dynamic task (e.g., Manzey et al., 2012). Nevertheless, all these measures represented operators' performance when working together with a reliable automation support and were therefore subsumed under the same metavariable. For defining the return-to-manual primary task performance metavariable, the same measures as for routine

performance were considered for a given study but for a situation where the operator needed to perform the primary task manually after a complete automation breakdown of the automation.

Workload measures were assessed in two different ways: As a performance variable we defined the metavariable *secondary task performance* (if the study used a multitask environment) again for routine and return-to-manual performance, respectively. A second metavariable was operators' *subjective workload*, typically quantified by the NASA-TLX (Hart & Staveland, 1988) as used, for example, by Endsley and Kaber (1999).

The *situation awareness* metavariable merged any direct and indirect indicators that pointed to LSA when working together with automation. As direct indicators we considered the outcome of techniques that are designed to directly ask for SA like SAGAT (Endsley, 1988, 2000) or questionnaires such as the Situational Awareness Rating Technique (Taylor, 1990). As indirect evidence for a possible loss of SA we considered all sorts of operators' errors that might be attributed to a loss of SA due to an overtrust in automation or a lack of proper understanding. Such errors can include, for example, mode errors (Sarter, 2008) or errors of omission or commission (Mosier & Skitka, 1996), that is, errors where operators failed to respond to a critical situation if the automation failed to alert them properly or where operators followed incorrect advice of automation without detecting this failure. When participants committed these kinds of error, we interpreted this as evidence for deficient SA as they did not realize that the automation had made a mistake.

Departing from the classic meta-analysis approach we assigned rankings for every metavariable within a single study according to significant effects found with regard to DOA (a priori, a posteriori). This was done as effect sizes (e.g., Hedges's *g*) were only rarely reported in the original studies and therefore could not be used for analysis, without eliminating many studies from consideration. Furthermore, any other estimates of effect sizes based on the *F* ratios for multiple conditions reported in the studies would not be able to capture the ordinal aspect of data, which is of particular relevance for our question. Although unconventional, this approach of data aggregation is in line with the basic idea of meta-analysis (e.g., Rosenthal, 1991), where no particular statistical method is defined for this "analysis of analyses." It is also in line with other authors who also departed from the classic approach for similar reasons (e.g., Hutchins, Wickens, Carolan, & Cumming, 2013; Wickens & Dixon, 2007; Wickens et al., 2013; Wickens, Hooey, Gore, Sebok, & Koenicke, 2009).

Different rankings were assigned when there was a significant effect between two DOA conditions ($p < .05$). In case of nonsignificant effects between different degrees of automation, we assigned tied ranks. For example, in case a study comparing the impact of three different DOAs on routine primary task performance revealed all pairwise comparisons between the DOA conditions as significant, the condition showing the worst performance was assigned rank 1, the condition with the second-best performance rank 2, and rank 3 was assigned to the condition with best performance. However, when only one condition differed in terms of superior performance compared to the other two conditions, the best condition was assigned rank 3 and the other two conditions were assigned tied ranks, in this case rank 1.5.

When a metavariable was measured by more than one dependent variable within a study (e.g., error of omission and SAGAT for SA), the rankings of the single variables were integrated into one "overall ranking." With this approach we were able to integrate data from various studies assessed in numerous ways to examine the tradeoff when automation degree increased and to identify trends on a descriptive level.

In a third step we described the relationship between the DOAs and the different metavariables by computing Kendall's tau (correlation between rank orderings) to see if the DOA had an impact on a certain class of metavariable. Kendall's tau was used as an alternative analysis to product moment correlations as we only had rank orderings as data bases. With this analysis it was possible to determine and test for a monotonic relation between two dependent variables (e.g., DOA and workload). Furthermore, Kendall's tau

does not make the implicit assumption of equidistance between different rankings, which would not have been the case for our data.

To further abstract the results, we computed an overall Kendall's tau for every metavariable across studies and tested with one-tailed *t* tests if this correlation was different from zero in the hypothesized direction. In doing so, we defined each Kendall's tau, computed for every study, as a certain manifestation of the variable in question (e.g., routine primary task performance). With this last step, we could also examine various instances of the trade-off: For example, do the routine and failure aspects of performance trade off? How strongly is decreased failure response coupled with LSA? Do workload and LSA trade off?

RESULTS

Table 1 shows the correlations of DOA on the six metavariables for the single studies (Kendall's tau) and the computed overall Kendall's tau for every metavariable including statistics of one-tailed *t* tests.

Primary Task Performance

A total of 16 studies provided data for the routine primary task performance metavariable. In terms of Kendall's tau, a vast majority of these studies indicated a strong positive correlation of DOA and routine performance. This is in accordance with the anticipated benefit of automation support with increasing DOA when automation functioned properly. Data of one study only resulted in a negative correlation, and an additional three studies revealed no evidence for a relation of DOA and routine primary task performance. Looking at the amount of studies with positive taus and the strength of these correlations supports the hypothesized benefit of automation support with increasing DOA. This interpretation is also backed up by a significant overall rank correlation across studies ($\tau = .51$, $p < .001$).

For an assessment of the impact of DOA on return-to-manual primary task performance, data of nine studies were available. Five of these studies reported effects that resulted in a negative Kendall's tau, whereas only three others showed no evidence for the hypothesized negative impact of DOA when participants had to resume the formerly automated task because of an automation breakdown. This general trend was reinforced by a negative overall Kendall's tau averaged across the nine studies that was significantly different from zero ($\tau = -.34$, $p = .03$).

Taken together, results for primary task performance support the hypothesized lumberjack effect as the routine and failure aspects of performance trade-off with increasing automation complexity. Further in line with the hypothesized trade-off (Figure 2) is the fact that eight out of the nine studies that assessed both aspects of performance (routine and return-to-manual) showed a higher (more positive) correlation of DOA with routine than with failure performance, and the single exception (Metzger & Parasuraman, 2005) showed a zero correlation in both cases.

Workload

Workload was evaluated on a performance level and on a subjective level. Eight studies provided data for the metavariable routine secondary task performance. Three out of these eight studies revealed a strong positive correlation of DOA and performance; that is, operators showed better results when supported by higher degrees of reliable automation. This was also supported by a significant overall Kendall's tau ($\tau = +.03$, $p = .04$). However, this interpretation is challenged by the fact that five out of the eight studies revealed no connection between DOA and performance in terms of zero correlations. Therefore, the results have to be interpreted with caution.

In contrast to primary task performance, secondary task performance did not seem to be affected by surprising automation breakdowns, as there was no evidence for an impact of DOA on return-to-manual secondary task performance. However, only three studies reported data for this variable, so the explanatory power of this result is rather low.

Concerning the impact of DOA on subjective workload, the 12 studies that reported data for this metavariable provided a quite complex pattern of results. Two of these studies (Endsley & Kaber, 1999; Wright & Kaber, 2005) reported data that revealed strong positive relations

between DOA and subjective workload ($\tau = +.80$, $\tau = +.91$). In contrast, six studies provided a reversed pattern with strong negative Kendall's taus, and the remaining four studies showed no correlation at all.

However, because the majority of the data provided negative correlations, overall Kendall's tau also showed a negative albeit weak trend ($\tau = -.24$, $p = .05$) that supports the often stated argument that higher degrees of automation reduce operators' workload. Nevertheless, because of the different results of the single studies, further research is needed to ensure the proposed interpretation.

Situation Awareness

We hypothesized that one of the costs concerned with higher degrees of automation would be associated with LSA. Eleven studies reported data for this metavariable. Whereas five studies, such as Endsley and Kiris (1995) and Manzey et al. (2012), did report a potential loss of SA with increasing automation, four other studies did not find an impact of DOA (Calhoun, Draper, & Ruff, 2009; Cummings & Mitchell, 2007; Metzger & Parasuraman, 2005; Reichenbach, Onnasch, & Manzey, 2011), and data of the two remaining studies even produced positive correlations of DOA on SA. Due to this, the hypothesized negative trend was not as strong as expected, with an overall Kendall's tau of –.29, but was still significantly different from zero ($p = .04$).

Taking a closer look at the single studies, it is striking that the two studies with the highest lumberjack trade-off (routine vs. return-to-manual primary task performance; Kaber et al., 2000; Manzey et al., 2012) were also two of the four studies that yielded comparatively strong negative correlations between DOA and SA (values of –.71 and –.63, respectively). This is in accordance with the assumption that higher DOAs increase the risk of out-of-the-loop unfamiliarity issues reflected in a loss of SA as well as with negative performance consequences in case an operator unexpectedly needs to resume manual control of an automated task (Endsley & Kiris, 1995). Similarly, the strongest negative correlation between DOA on SA was found in the study conducted by Li, Wickens, Sarter, and Sebok (in preparation), which at the same time showed the greatest automation benefits for routine primary task performance as well as the greatest decrease in subjective workload of all studies. Therefore, it seems that routine primary task performance and workload, on one hand, and the potential loss of SA, on the other hand, directly trade off and that appropriate function allocation can serve only one of the two aspects.

Moderating Factors

In a next step we tried to identify possible factors that might moderate potential trade-offs between the different measures. We looked for aspects that some studies had in common, especially those that strongly supported the trade-off hypothesis, but also differentiated them from other studies. As one such variable, we focused on the critical distinction between automation that supported situation assessment by providing automated information acquisition and analysis (Stage 1 or Stage 2) versus that which supported the selection and execution of action (Stages 3 and 4). This distinction of assessment versus action is an ubiquitous one that underlies many facets of human performance (Wickens, Hollands, Banbury, & Parasuraman, 2012). For this analysis only those studies that varied the DOA across the assumed critical boundary from information analysis support to action selection support were included (i.e., Crocoll & Coury, 1990; Cummings & Mitchell, 2007; Manzey et al., 2012; Reichenbach et al., 2011; Rovira et al., 2007; Sarter & Schroeder, 2001). Examining these studies exclusively, we found that when DOA was varied across this boundary, the pattern of the lumberjack analogy trade-off was substantially amplified. Calculated for these six studies separately, the overall Kendall's tau correlation of DOA with routine performance was +.68, higher than the overall correlation of +.51 for all studies (see Table 1); and the overall correlation with return-to-manual performance was –.90, much more negative than the overall correlation of all studies, a value of –.34.

We also examined how other variables such as prior experience with failures or subject experience might have modulated the trade-off. Concerning the training participants received, we looked especially for the possible impact of

"first failure automation training," which has been found to affect human performance with automation (Bahner, Huper, & Manzey, 2008). Yet none of the integrated studies applied such kind of training. The only systematic differences in training were related to practice time before the experiment started. However, this difference is hard to evaluate since training time usually depends on the complexity of the experimental task.

In all but four studies students served as participants. In one study participants were recruited from Air Force personnel but still were novices for the experimental task (Calhoun et al., 2009). Three studies were conducted with experts as the experimental simulation was very realistic (control of unmanned aerial vehicles, ATC, pilots). Nevertheless, these differences did not moderate the reported lumberjack trade-off effects of DOA.

Also, most studies used multitask settings but differed in the number of concurrent tasks (two or three). Yet four studies represented single-task studies (Crocoll & Coury, 1990; Endsley & Kaber, 1999; Endsley & Kiris, 1995; Kaber et al., 2000). One could assume that this differentiation might be important, especially for variables such as workload or SA. However, the amount of secondary tasks did not seem to make a difference.

Another variable that was examined in detail was the nature of the display of the automated process. The rationale for this focus was twofold. The emerging literature that clear, intuitive, or "ecological" displays of the state of automated processes can support a proper response to automation failures (Bennett & Flach, 2011; Burns et al., 2008; Seppelt & Lee, 2007). The linkage between displays and SA support on one hand, and our finding that suggests that LSA might be related to return-to-manual performance issues. Although this relation too did not emerge from our post hoc analysis of the data, we are certainly reluctant to conclude that effective displays do not support off-nominal response via the mediating role of SA because of the relatively low power of our assessment.

DISCUSSION

Overall, the results fairly conclusively confirm the lumberjack hypothesis with regard to the DOA. "Conventional wisdom" has now been transformed into "statistical wisdom." Thus, the pattern underlying the DOA confirms the general pattern that had previously been observed regarding the presence or absence of automation. Automation helps when all goes well, but leaving the user out of the loop can be problematic because it leads to considerable performance impairment if the automation suddenly fails. And this risk appears to increase with increasing DOA. The data presented in Table 1 further suggest that this effect is linked to raised issues of LSA with increasing DOA. However, due to a lack of statistical power, this latter conclusion needs to be treated with caution.

The most promising account is suggested by the final post hoc analysis reported earlier. When DOA moves across the critical boundary from information acquisition and information analysis to action selection, the latter alleviating the human from some or all aspects of choosing an action, then the human is much more vulnerable to automation "failures." Actively choosing actions manually (the generation effect; Slamecka & Graf, 1978) supports SA in a way that supports the manual performance in case of automation breakdown. When that choice is removed, the automation failure response suffers. Thus, the distinction between situation assessment and action support is critically important in automation, just as the simple dichotomy is in other aspects of human factors and cognitive engineering, such as cognitive task analysis (Hoffman, Crandall, & Shadbolt, 1998), predicting multitask performance (Wickens, 2008b), and predicting transfer of training (Osgood, 1949).

This finding also qualifies and specifies earlier claims that medium levels of automation would represent an optimum choice with respect to primary performance improvements and workload reductions by, at the same time, reducing unwanted performance consequences in terms of LSA and difficulties of return-to-manual performance (Endsley & Kiris, 1995). The direct trade-off between DOA-related consequences on primary task performance and return-to-manual performance, respectively, suggests that there is no clear optimum of automation support. Each

step of increase of DOA seems to be associated with an increase of the risk of return-to-manual performance decrements, meaning that there is no specific DOA below which automation-induced performance benefits can be increased without any performance costs. This renders doubts in any simple design recipes like "medium DOAs are best." However, the strength of the trade-off is important particularly if the border between information and action support is crossed. That is, the general recommendation of preferring "medium levels of automation" where the human is kept somehow "in the loop" can now be turned into a more specific one: If return-to-manual performance issues are of serious concern, human operators should be kept involved at least to some extent in decision and action selection as well as action implementation. However, even if in this case risks of return to manual might not be fully excluded, they can probably be kept to a comparably low level.

One limitation of the present research is the comparably small number of studies available for this analysis and the need to just consider rank data with respect to scaling DOA and performance effects. Our approach of using rank orders based on dominance orderings of three features, that is, stages, levels, and number of stages, allowed neither for quantifying the DOA on a ratio or interval scale nor for resolving trade-offs between stages and levels. This made it difficult to yield clear statistical conclusions for all of the findings and limits the conclusiveness of results with respect to the formal characteristics of the observed trade-offs (To what extent are they linear?). However, based on the limited current knowledge and available data, the rank order approach applied to represent DOA seemed to be the only way to yield a quantitative input for our meta-analysis. Clearly, much more psychophysical and controlled experimental research is needed before more distinct metric DOA scales may be developed. A second limitation is the possibility that we might have underestimated the trends within any particular study, with the relatively coarse dichotomous "grain size" by which effects were coded (significant vs. nonsignificant). In doing so, we collapsed across quantitative measures of the size of an effect that might have added precision to the coding. Taking these limitations in mind, the overall pattern of raw effects and statistical results provides a first quantitative summary of the state of knowledge about performance consequences of stages and levels of automation which, together with the remaining questions concerning possible moderating factors, certainly offers an invitation for future research.

ACKNOWLEDGMENTS

This research was sponsored by NASA-Johnson Space Center Grant NNX09AM81G. Barbara Wolford was the scientific-technical monitor. The authors wish to thank Angelia Sebok and Nadine Sarter for their support in managing and monitoring the research contract under which this review was carried out.

KEY POINTS

- Increasing DOA supports routine system performance and workload.
- Increasing DOA negatively affects failure system performance and SA.
- Negative consequences of automation are most likely when DOA moves from Stage 2 to Stage 3 automation.

REFERENCES

Bahner, E. J., Huper, A. D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and training experience. *International Journal of Human-Computer Studies*, *66*, 688–699.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*, 775–779.

Bennett, K. B., & Flach, J. M. (2011). *Display and interface design: Subtle science, exact art*. Boca Raton, FL: CRC Press.

Burns, C. M., Skraaning, G., Jamieson, G., Lau, N., Kwok, J., Welch, R., & Andresen, G. (2008). Evaluation of ecological interface design for nuclear process control: Situation awareness effects. *Human Factors*, *50*, 663–698.

Calhoun, G. L., Draper, M. K., & Ruff, H. A. (2009). Effect of level of automation on unmanned aerial vehicle routing task. In *Proceedings of the Human Factors and Ergonomics Society 53rd annual meeting* (pp. 197–201). Santa Monica, CA: Human Factors and Ergonomics Society.

Crocoll, W. M., & Coury, B. G. (1990). Status or recommendation: Selecting the type of information for decision aiding. In *Proceedings of the Human Factors and Ergonomics Society 34th annual meeting* (pp. 1524–1528). Santa Monica, CA: Human Factors and Ergonomics Society.

Crossman, E. R. F. W. (1974). Automation and skills. In E. Edwards & F. Lees (Eds.), *The human operator in process control* (pp. 1–24). London, UK: Taylor & Francis.

Cummings, M. L., & Mitchell, P. J. (2007). Operator scheduling strategies in supervisory control of multiple UAVs. *Aerospace Science and Technology*, *11*, 339–348.

Dekker, S. W. A., & Woods, D. D. (2002). MABA-MABA or abracadabra? Progress on human-automation co-ordination. *Cognition, Technology, and Work*, *4*, 240–244.

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). In *Proceedings of the National Aerospace and Electronics Conference (NAECON)* (pp. 789–795). New York, NY: IEEE.

Endsley, M. R. (2000). Direct measurement of SA: Validity and use of SAGAT. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness: Analysis and measurement* (pp. 147–173). Mahwah, NJ: Lawrence Erlbaum.

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*, 462–492.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, *37*, 387–394.

Ephrath, A. R., & Young, L. R. (1981). Monitoring vs. man-in-the-loop detection of aircraft control failures. In J. Rasmussen & W. B. Rouse (Eds.), *Human detection and diagnosis of system failures* (pp. 143–154). New York, NY: Plenum.

Fadden, S., Ververs, P., & Wickens, C. D. (1998). Costs and benefits of head-up display use: A meta-analytic approach. In *Proceedings of the 42nd annual meeting of the Human Factors and Ergonomics Society* (pp. 16–20). Santa Monica, CA: Human Factors and Ergonomics Society.

Hart, S. G., & Staveland, L. E. (1988). Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam, Netherlands: Elsevier.

Hoffman, R. R., Crandall, B., & Shadbolt, N. (1998). Use of the critical decision method to elicit expert knowledge: A case study in the methodology of cognitive task analysis. *Human Factors*, *40*, 254–276.

Horrey, W. J., & Wickens, C. D. (2006). Examining the impact of cell phone conversations on driving using meta-analytic techniques. *Human Factors*, *48*, 196–205.

Hutchins, S. D., Wickens, C. D., Carolan, T. F., & Cumming, J. M. (2013). The influence of cognitive load on transfer with error prevention training methods: A meta-analysis. *Human Factors*, *55*, 854–874.

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*, 113–153.

Kaber, D. B., Onal, E., & Endsley, M. R. (2000). Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Human Factors and Ergonomics in Manufacturing*, *10*, 409–430.

Kessel, C. J., & Wickens, C. D. (1982). The transfer of failure-detection skills between monitoring and controlling dynamic systems. *Human Factors*, *24*, 49–60.

Kirlik, A. (1993). Modeling strategic behavior in human-automation interaction: Why an "aid" can (and should) go unused. *Human Factors*, *35*, 221–242.

Layton, C., Smith, P. J., & McCoy, C. E. (1994). Design of a cooperative problem-solving system for en-route flight planning: An empirical evaluation. *Human Factors*, *36*, 94–119.

Li, H., Wickens, C. D., Sarter, N. B., & Sebok, A. (in preparation). *An investigation of automation degree in assisting robotic arm control*. Technical report in preparation.

Lorenz, B., Di Nocera, F., Röttger, S., & Parasuraman, R. (2002a). Automated fault-management in a simulated spaceflight micro-world. *Aviation, Space, and Environmental Medicine*, *73*, 886–897.

Lorenz, B., Di Nocera, F., Röttger, S., & Parasuraman, R. (2002b). Varying types and levels of automation in the support of dynamic fault management: An analysis of performance costs and benefits. In D. de Waard, K. A. Brookhuis, J. Moraal, & A. Toffetti (Eds.), *Human factors in transportation, communication, health, and the workplace* (pp. 517–524). Maastricht, Netherlands: Shaker.

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, *6*, 57–87.

Metzger, U., & Parasuraman, R. (2005). Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. *Human Factors*, *47*, 35–49.

Miller, C. A., & Parasuraman, R. (2007). Designing for flexible interaction of humans and automation: Delegation interfaces for supervisory control. *Human Factors*, *49*, 57–75.

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 201–220). Mahwah, NJ: Lawrence Erlbaum.

Osgood, J. (1949). The similarity paradox in human learning: A resolution. *Psychological Review*, *47*, 419–427.

Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: A review and attentional synthesis. *Human Factors*, *52*, 381–410.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." *International Journal of Aviation Psychology*, *3*, 1–23.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse and abuse. *Human Factors*, *39*, 230–253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model of types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics–Part A*, *30*, 286–297.

Parasuraman, R., Sheridan, T., & Wickens, C. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making*, *6*, 140–160.

Rasmussen, J., & Rouse, W. B. (1981). *Human detection and diagnosis of system failures*. New York, NY: Plenum.

Reichenbach, J., Onnasch, L., & Manzey, D. (2011). Human performance consequences of automated decision aids in states of sleep loss. *Human Factors*, *53*, 717–728.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

Röttger, S., Bali, K., & Manzey, D. (2009). Impact of automated decision aids on performance, operator behaviour and workload in a simulated supervisory control task. *Ergonomics*, *52*, 512–523.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, *49*, 76–87.

Sarter, N. B. (2008). Investigating mode errors on automated flight decks: Illustrating the problem-driven, cumulative, and interdisciplinary nature of human factors research. *Human Factors*, *50*, 506–510.

Sarter, N. B., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors*, *43*, 573–583.

Seppelt, B. D., & Lee, J. D. (2007). Making adaptive cruise control (ACC) limits visible. *International Journal of Human-Computer Studies*, *65*, 192–205.

Sethumadhavan, A. (2009). Effects of automation types on air traffic controller situation awareness and performance. In *Proceedings of the Human Factors and Ergonomics Society 53rd annual meeting* (pp. 1–5). Santa Monica, CA: Human Factors and Ergonomics Society.

Sheridan, T. B. (2002). *Humans and automation: Systems design and research issues*. New York, NY: John Wiley.

Sheridan, T. B., & Parasuraman, R. (2000). Human vs. automation in responding to failures: An expected-value analysis. *Human Factors*, *42*, 403–407.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Technical report). Cambridge, MA: MIT, Man Machine Systems Laboratory.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning & Memory*, *4*, 592–604.

Taylor, R. M. (1990). Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design. In *Situational awareness in aerospace operations* (AGARD-CP-478; pp. 3/1–3/17). Neuilly Sur Seine, France: NATO-AGARD.

Wickens, C. D. (2008a). Multiple resources and mental workload. *Human Factors*, *50*, 449–455.

Wickens, C. D. (2008b). Situation awareness: Review of Mica Endsley's 1995 articles on situation awareness theory and measurement. *Human Factors*, *50*, 397–403.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*, 201–212.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2012). *Engineering psychology and human performance* (4th ed.). New York, NY: Pearson.

Wickens, C. D., Hooey, B. L., Gore, B. F., Sebok, A., & Koenicke, C. S. (2009). Identifying black swans in NextGen: Predicting human performance in off-nominal conditions. *Human Factors*, *51*, 638–651.

Wickens, C. D., Hutchinson, S., Carolan, T., & Cumming, J. (2013). Effectiveness of part task training and increasing difficulty training strategies: A meta-analysis approach. *Human Factors*, *55*, 461–470.

Wickens, C. D., & Kessel, C. J. (1979). The effects of participatory mode and task workload on the detection of dynamic system failures. *IEEE Transactions on Systems, Man and Cybernetics*, *9*, 24–34.

Wickens, C. D., & Kessel, C. J. (1980). Processing resource demands of failure detection in dynamic systems. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 564–577.

Wickens, C. D., & Kessel, C. J. (1981). Failure detection in dynamic systems. In J. Rasmussen & W. B. Rouse (Eds.), *Human detection and diagnosis of system failures* (pp. 155–169). New York, NY: Plenum.

Wickens, C. D., Mavor, A. S., Parasuraman, R., & McGee, P. (1998). Airspace system integration: The concept of free flight. In C. D. Wickens, A. S. Mavor, & J. P. McGee (Eds.), *The future of air traffic control: Human operators and automation* (pp. 225–245). Washington, DC: National Academy.

Wiener, E. L. (1988). Cockpit automation. In E. L. Wiener & D. C. Nagel (Eds.), *Human factors in aviation* (pp. 433-461). San Diego, CA: Academic Press.

Wiener, E. L., & Curry, R. E. (1980). Flight-deck automation: Promises and problems. *Ergonomics*, *23*, 995–1011.

Wright, M. C., & Kaber, D. B. (2005). Effects of automation of information-processing functions on teamwork. *Human Factors*, *47*, 50–66.

Yeh, M., Merlo, J. L., Wickens, C. D., & Brandenburg, D. L. (2003). Head up vs. head down: Costs of imprecision, unreliability, and visual clutter on cue effectiveness for display signaling. *Human Factors*, *45*, 390–407.

Linda Onnasch is a research fellow at the Department of Psychology and Ergonomics, Berlin Institute of Technology, Germany, where she earned a master in psychology in 2009. She is currently working on a PhD addressing issues of human redundancy in supervisory control.

Christopher D. Wickens is a professor emeritus of aviation and psychology at the University of Illinois and is a senior scientist at Alion Science and Technology, Boulder, Colorado.

Huiyang Li is a PhD candidate in industrial engineering at the University of Michigan. She earned her BS in electrical engineering from Peking University and her MS in applied psychology from the Chinese Academy of Sciences. Her areas of interest include function allocation, attention management, and multimodal interface design in complex systems.

Dietrich Manzey is a university professor of work, engineering and organizational psychology in the Department of Psychology and Ergonomics, Berlin Institute of Technology, Germany. He earned his PhD in experimental psychology at the University of Kiel, Germany, in 1988 and his habilitation in psychology at the University of Marburg, Germany, in 1999.

Date received: November 16, 2012
Date accepted: July 18, 2013