2009

Conflict Alerts and False Alerts in En-Route Air Traffic Control: an Empirical Study of Causes and Consequences

Christopher D. Wickens

Stephen Rice

David Keller

Jamie Hughes

Shaun Hutchins

Repository Citation

Wickens, C. D., Rice, S., Keller, D., Hughes, J., & Hutchins, S. (2009). Conflict Alerts and False Alerts in En-Route Air Traffic Control: an Empirical Study of Causes and Consequences. *2009 International Symposium on Aviation Psychology*, 196-201.
https://corescholar.libraries.wright.edu/isap_2009/82

CONFLICT ALERTS AND FALSE ALERTS IN EN-ROUTE AIR TRAFFIC CONTROL: AN EMPIRICAL STUDY OF CAUSES AND CONSEQUENCES

Christopher D. Wickens
Alion Science: MA&D Operations, Boulder, Colo.
Stephen Rice
New Mexico State University
David Keller
New Mexico State University
Jamie Hughes
New Mexico State University
Shaun Hutchins
Alion Science: MA&D Operations, Boulder, Colo.

We analyzed the extent to which a high false alert rate of the conflict alerting (CA) system in five ATC facilities was the cause of a “cry-wolf” effect, whereby true alerts of a pending loss of separation were associated with either controller failure to respond or a delayed response. Radar track data surrounding 497 CA’s were examined and from these we extracted information as to whether the alert was true or false, whether a trajectory change was (response) or was not (non-response) evident, whether a loss of separation occurred, and the controller response time to the CA. Results revealed an overall 47% false alert rate, but that increases in this rate across facilities was not associated with more non-responses or delayed responses to true alerts, or loss-of-separation. Cry-wolf appeared to be absent. Instead, desirable anticipatory behavior indicated that controllers often responded prior to the conflict alerts.

In June 2006, the National transportation and Safety Board documented a series of accidents - controlled flight into terrain, and mid-air collisions - in which the minimum safe altitude warning (MSAW) and conflict alerts, respectively, announced a pending collision (NTSB, 2006). However, controllers failed to respond or intervene to prevent the accident. Furthermore, anecdotal evidence from a specific accident (midair collision of two Cessna aircraft near San Diego), and from other interviews with controllers (Ahlstrom & Pasnjwani, 2003), suggested the prevalence of controller experience of the “cry wolf” effect (Breznitz, 1983). The “cry wolf” effect is a general syndrome whereby excessive alarms, many of them seemingly unnecessary to the operator (e.g., “false alarms” or “false alerts”), lead to a distrust in the alarm system, and a disregarding of (or late response to) some true alarms.

Linking this well observed phenomenon to the findings of missed alerts in the NTSB study suggests that there may be a causal connection between the two. When examining false alerts in predictive collision alerting systems, certain features of time-dependence (Kuchar, 2000) make these different from other alerts such as cockpit engine warnings (Dixon & Wickens, 2006). In particular, inherent in any dynamic system in a noisy environment subject to cross winds, turbulence, and pilot control inputs, is the problem that prediction becomes less accurate with increasing look-ahead time. Furthermore, an alert may be “false” for two reasons; it may actually predict a loss of separation but extrapolation of the trajectory indicates that an LOS will not occur; or it may correctly predict an LOS, but a subsequent trajectory change (in response to a controller instruction) is implemented so that no LOS is observed. Finally, true “misses” are very rare in CA systems; but these are more often manifest as delayed alerts. Clearly if the alert is delayed so long that there is little time to maneuver the aircraft away from the separation loss, such an event can be seen as equivalent to a miss.

A general conclusion from research which has examined false alerts, when humans can monitor the data in parallel appears to be that, while misses may be catastrophic in a system in which there is no human backup to monitor the raw data in parallel, in systems that allow such parallel human-machine monitoring (Parasuraman, 1987), false alarm-prone systems may often be worse (Dixon & Wickens, 2006; Dixon, Wickens & McCarley, 2007; see Wickens, Levinthal, & Rice, in press, for a summary). This may be particularly true in high workload multi-task circumstances since a false-alarm prone system may not only cause ignorance of true alerts, but, when such alerts are responded to, this response can be quite disruptive of concurrent tasks; either as a result of carrying out the unnecessary alarm-triggered action or of the need to cross check the raw data to establish that the alarm was indeed false. In a further argument for a higher threshold, in many predictive alerting systems such as the conflict alerts studied here, a higher threshold does not necessarily translate to more missed events, but only to a later alerting of true events (a much less catastrophic outcome). Indeed if this alerting look-ahead time still provides adequate time for humans to respond, then the benefit of reducing false alarms would more than offset the shorter time period between the alert and the occurrence of the forecast event (e.g., the pending collision).

The purpose of the current study was to seek evidence for the FA-caused cry-wolf phenomenon from live or "naturalistic" data across five air traffic control facilities in which controllers responded to mid-air conflict alerts (CA's), and across which the CA false alert rate varied. Such live ATC data have never before been examined in this fashion; although it parallels the analysis of weather forecasters (Barnes, et al., 2006), pilots (Bliss, 2004), and health care practitioners (Xiao, et al., 2004), responding to imperfect alerting systems. In this process we must first examine performance of the CA system itself, to assess a FA rate, and then examine the influence of differences in this rate on behavior of the controller, and performance of the controller-CA (human-automation) system as a whole.

In the current research, we addressed the hypothesis that, assuming there to be variability in false alert rate across ATC facilities, those with the higher FA rate, would show greater evidence for "cry wolf" behavior: later responses, and/or more non-responses. In addition, we examine other aspects of controller response to CA's that are either true or false; in particular considering the properties of the alerting system that may lead to a loss of separation, and/or lead to desirable anticipatory behavior.

Methods: CA system analysis

The CA system (FAA, 2003) is designed to predict when two aircraft will close simultaneously, within 5 miles laterally and 1000 feet vertically. Figure 1 presents the schematic for the lateral dimension only. Such closure is known as a **loss of separation (LOS),** shown on the left of figure 1. Hence the CA predicts any LOS that is forecast to occur within a look-ahead time of 75-135 seconds. When the CA system predicts such an LOS, the data tags on the controllers' display start to flash. The algorithm underlying the CA generates a linear extrapolation both on the horizontal (map) plane and the vertical plane, of the current heading and vertical speed of both aircraft (FAA, 2003).

We were provided data from the FAA for 494 conflict alerts, extracted from the busiest 2-hour periods from a sample of 2 or 3 days in each of five en-route ATC centers. Such data (distributed across three different data bases) included for each CA: (1) properties of the pair of trajectories predicted by the CA (e.g., predicted point of closest passage, time of alert), (2) the actual radar tracks & altitude of the aircraft (sampled every 10 sec), and (3) a short analysis of the actual conflict as it was played out (See Wickens, Rice, et al., 2008, for details). The most important element of this third set was a metric (minmax ratio or MMR) describing the inverse severity of the conflict. A value of 0 corresponded to an actual collision and a value of 1 was the threshold for a loss of separation. Higher values indicated passage with greater lateral and vertical separation than the minima. Two key variables provided to us for each center were the "busyness" of the center (the number of encounters per hour (where "encounter" is the point at which the CA algorithm begins to examine track pairs), and the number of CA's during the equivalent period. Table 1 shows these two parameters across the five Centers (row 2 and 3) along with the ratio in row 4 of the total CA's to the total encounters within the center; an estimate of the CA rate. Importantly, Table 1 reveals that what might be defined as the "CA-rate" in the bottom row did not vary substantially across Centers, in spite of the 8-fold variation in "busyness".

Table 1. *Basic data from CA systems.*

ZLC	ZHU	ZLA	ZTL	ZID
1126	1,589	5,529	5679	8,813
22	36	72	435	124
22/4525=.005	124/26440=.004	36/4767=.007	235/38815=006	72/16589=.004

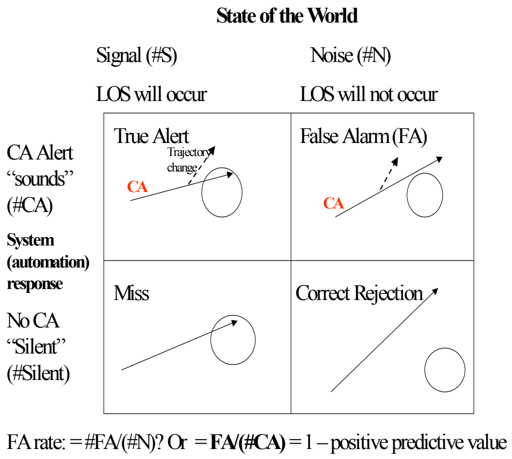
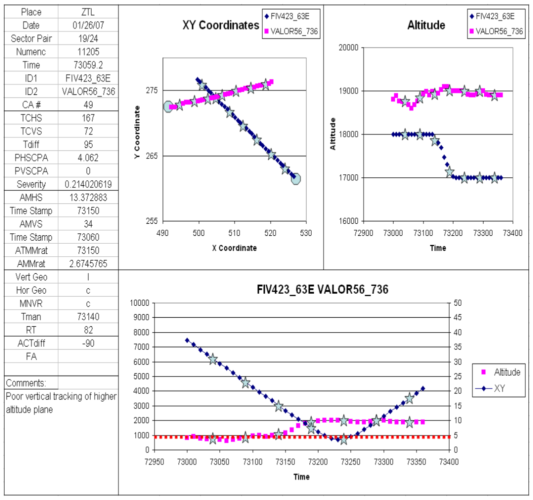


Figure 1: Geometry of CA system shown when an LOS is pending (left column) and not (right), and when the alert is triggered (top) or not (bottom). Within the triggered alerts, controllers may (dashed line) or may not (solid line) initiate a maneuver.



Place	ZTL
Date	01/26/07
Sector Pair	19/24
Numenc	11205
Time	73059.2
ID1	FIV423_63E
ID2	VALOR56_736
CA #	49
TCHS	167
TCVS	72
Tdiff	95
PHSCPA	4.062
PVSCPA	0
Severity	0.214020619
AMHS	13.372883
Time Stamp	73150
AMVS	34
Time Stamp	73060
ATMMrat	73150
AMMrat	2.6745765
Vert Geo	l
Hor Geo	c
MNVR	c
Tman	73140
RT	82
ACTdiff	-90
FA	
Comments:	

Poor vertical tracking of higher altitude plane

Figure 2: Example of radar tracks used to classify alert as true or false, and to identify controller response. A trajectory change (descent of blue) is clearly visible in the vertical track at the upper right. The lower plot depicts separation on both axes.

Results 1: CA system analysis.

For the CA system, we calculated the false alert rate as the proportion of alerts that were categorized as "false". As noted above, we also distinguished between alerts (both true and false) when controllers did and did not impose a trajectory change, as such change was inferred from the radar tracks. When a change was implemented, a visual analysis was used to extrapolate the pre-change course, to assess if an LOS **would have** simultaneously compromised lateral and vertical minima, had the alteration **not** taken place. This analysis was carried out on ground track data, an example of which is shown in figure 2. The analysis was carried out by two independent observers for two of the centers, and by one of these observers for the remaining three.

We examined the computed FA rate as a function of the CA rate (CA/encounter) across the five centers. Two features became evident from this examination. First, there was considerable variance in FA rate, from a low of 0.28 to a high of 0.58, and on average approximately half of the CA's were "false". This allowed us to examine the cry wolf effect. Second, there appeared to be no relationship between CA rate and FA rate. A separate analysis also revealed that FA rate did not co-vary with overall traffic density. We also analyzed and categorized the geometry of the trajectories of the pairs of aircraft entering into each CA, and of the controller responses (e.g., climb, turn); these analyse can be found in Rantanen & Wickens (2009); and Wickens Rice et al (2008).

Results 2: Controller performance Analysis

Categorical analyses. Before examining the influence of FA rate on manifestations of the cry wolf phenomenon, it is necessary to identify the overall prevalence of those manifestations in our sample of data. Thus, in addition to the dichotomization of true versus false alerts discussed above, - characteristics of the automation - we examined two other important dichotomies which are characteristic of the human (controller): the presence or absence of a response (as inferred from visual analysis of the track data), and the presence or absence of a loss of separation (LOS). As noted in the previous section, it was usually relatively easy to identify whether a distinct change in trajectory was implemented in the time period around a CA (see the descent of the blue aircraft in figure 2), hence allowing inference of the presence and delay of a controller response. However, for a small sample, this classification became quite difficult and so those trials were not included in the data base.

Our analysis revealed that on roughly 10% of the CA's there was no evidence for a controller response, at least as indicated by a trajectory change by either of the two aircraft involved in the CA. These non-responses were statistically more prevalent when the CA was false (18%) than when it was true (1.5% $\chi^2(1, N = 437) = 37.5, p < .0001$). Such a result might be anticipated to the extent that the trajectories triggering a false alarm are, by definition, more likely to yield a more distant "closest passage" or miss distance and hence more likely to be considered by the controllers not to require a trajectory change.

Our analysis also revealed that the LOS rate is, like the non-response rate, approximately 10% of the data base. Also, it appears that the two types of outcomes are unevenly distributed across the two types of alerts. Specifically, True alerts are more likely to precede a loss of separation (21%) than are false alerts. (3%; $\chi^2(1, N = 373) = 20.3$, $p < .0001$) Here too, this is a plausible outcome, given that the true alert will occur on a trajectory pair that is more dangerous, and hence slightly more likely to yield the ultimate loss of separation, even *with* a controller intervention.

We then examined the relationship between controller response and LOS, to establish the extent to which non-responses might be associated with a LOS. These observations are collapsed over true vs. false alerts. This analysis indicated that when the controller **did not** respond, this was very unlikely to produce an LOS (5%; and those two events were restricted to a single center), whereas such LOS events were somewhat more prevalent when the controller **did** respond (9%) although the difference in proportion was not significant. ($\chi^2(1, N = 380) = .778$, $p < .378$). We note here that this finding does not necessarily imply that controller responses were counter-productive, since the vast majority of LOS cases occur on true alerts, where there would definitely have been an LOS had the controller not intervened with a trajectory change.

Collectively, the above three analyses provide no evidence for the strongest form of cry wolf effect (non-response leading to a LOS) and indeed the number (2) of such joint events is even less than what the independent product of the two classes of events might predict (10% NR rate X 10% LOS rate = 1% of the CA events = 4). We next sought to determine if there was any causal relation between FA rate, *as it varied across centers*, and either non-responses or LOS events. Figure 3 shows the scatter plot of FA vs. non-response, and reveals a striking and pronounced trend: the greater the false alarm rate in the center, the less controllers tended to respond (r = 0.944; p<.05). However, when the LOS rate was examined as a function of FA rate across Center, there was no trend. This null effect suggests that the increase in non-responses in the more FA-prone Center were not associated with a reduction in safe separation.

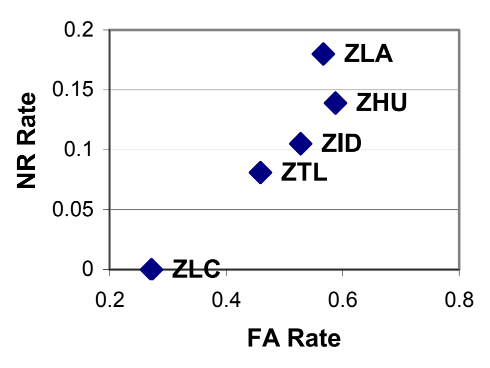


Figure 3: Non response rate as a function of false alert rate.

*Response Time***.** We then examined the second manifestation of the cry wolf phenomenon: the possible delay in controller response time associated with more FA's. Interpreting the delay between the CA and the trajectory change response required consideration of the total transmission lag (TTL).This is the time for the following processes to occur: (1) controller notices a dangerous convergence; (2) controller chooses a trajectory change and communicates this to the pilot; (3) pilot confirms and implements the change with flight controls; (4) the aircraft initiates a sufficient trajectory change to be evident in the radar track This TTL is estimated to be approximately 20-25 seconds (Allendoerfer & Friedman-Berg, 2007). Our analysis revealed that for about 45% of the CA's, controllers must have initiated the trajectory processing (noticing convergence and choosing a maneuver) **before** the CA occurred because the RT was less than 25 seconds. Indeed when we examined the distribution of response times, relative to the CA, we observed a distinct bimodality, with a minimum at around 25 seconds (See Wickens, Rice, et al., 2008). This bimodality, coupled with the estimate of a 25 second TTL, supported the notion that there were two categorically different types of responses: which we labeled **anticipatory** responses and **reactive** responses.

An ANOVA carried out on the ln-transformed RT data indicated that, for anticipatory responses, there was no difference in response time between true and false alerts ($p > .10$); however for reactive responses, true alerts were responded to approximately 73-59=14 seconds more rapidly than false responses ($t(193) = 2.4$, $p < .02$), reflecting the increased urgency of the former. There was no significant difference in RT between LOS and non-LOS encounters, so we can reject the hypotheses that the former resulted because of a delay in response time.

An analysis of three centers' data did reveal a main effect of center, $F(2, 154) = 6.78$, $p < 0.01$, with the highest density center (ZID) showing faster responses (30 s) than either the low (33 s) or mid (36 s) density Centers, an effect observed for both anticipatory and reactive responses. This effect is noteworthy because, whereas increasing density might have been anticipated to slow RT because of greater workload, the faster RT for ZID was observed **despite** its greater traffic density (See Table 1). Finally, within the non-LOS encounters (MMR > 1.0), we correlated RT with the value of MMR to test if later responses were responded with closer (but still above minima) passages. This correlation, examined for the three mid-level Centers was non significant ($p > 0.10$), suggesting that controllers did not compromise safety when their responses were delayed.

Finally, we examined the frequency of anticipatory vs. reactive responses for true vs. false alarms. Analyses of these data reveals that controllers were significantly more likely to anticipate on a true (0.58) than a false (0.37) alert ($\chi^2(1, N = 374) = 5.08$, $p = .024$). This is a plausible finding because the true alert trajectories should signal the impending conflict with greater salience in the raw data of the radar displays.

Discussion

The current data provide little or no evidence that the FA-induced cry wolf effect exists for the en-route CA system, as it is operationally defined by non-response to true alerts, and by later responses to all alerts. More particularly, false alerts do not appear to be responsible for safety-compromise in the ATC centers whose data were sampled. The generality and robustness of this conclusion is supported both by the wide range of center busyness, as well as the large sample size of the data, which provides for powerful statistical conclusions. (That is, the null hypothesis was not accepted simply because of a small N).

Of course ours was not a true experiment with control exerted across all other aspects of the centers. As in any correlational study, confounding variables could have contributed to our results. One such potentially confounding variable is that traffic-induced workload differences between centers could have accounted for effects. Indeed while this is possible, two factors mitigate concern for this confounding interpretation of the result. First, the busiest center (ZID) was only in the middle of the range in terms of both false alerts and non-responses (Figure 3). If we assume busyness is a proxy for workload, then this result would not have been obtained had workload been a responsible factor. Second, the possible confound with workload would have been more problematic had we found that a higher FA rate was associated with more non-responses to true alerts, and/or late responses. In that case we would need to reason as to why workload was not responsible for the effect. But as noted, neither of these associations were observed.

In terms of why FA-induced cry wolf behavior did not appear to be observed here, we note that, while false, most of the alerts in the CA system were not wildly off the mark, and thereby signaled a system whose threshold was set just a little lower than it needed to be, in the conservative interests of preserving safety and avoiding misses or late alerts. Recently Lees and Lee (2007) have found that such alerts can actually be beneficial to performance, in confirming that the system is generally functioning well. In the current case, for the large number of anticipatory responses, one can think of the alerting systems reinforcing the conflict predictions (and trajectory alterations) that the controllers actually made in advance of the alerting system warning.

Acknowledgments.

This research was supported by grant # from the FAA to New Mexico State University. The authors wish to thank Kenneth Allendoerfer, Ken Leiden, Krisstal Clayton, and Jill Kamienski for their contributions.

References

Ahlstrom, V., Longo, M. , & Truitt, T. (2005). *Human factors design guide (DOT/FAA/CT-02-11).*Atlantic City, NJ: Federal Aviation Administration.

Allendoerfer, K. & Friedman-Berg, F. J. (2007). *Human Factors analysis of safety alerts in air traffic control*. Final Report DOT/FAA 07/22. Washington D.C.: Federan Aviation Agency.

Barnes, L. R., Gunfest, E., Hayden, M. H., Schultz, D. M., & Benight, C. (2006). False alarms and close calls: A conceptual model of warning accuracy. *Weather and Forecasting, 22,*1140-1147.

Bliss, J (2003). An investigation of alarm related accidents & incidents in aviation. *International Journal of Aviation Psychology, 13,* 249-268.

Breznitz, S. (1983). *Cry-wolf: The psychology of false alarms*. Hillsdale, NJ: Lawrence Erlbaum.

Dixon, S. R. & Wickens, C.D. (2006). Automation reliability in unmanned aerial vehicle flight control: A reliance-compliance model of automation dependence in high workload. *Human Factors, 48*, 474-486.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse the misses? *Human* FAA (2006).

FAA (2003). Common ARTS computer program functional specification for conflict alerts. NAS-MD-632. Wash DC.: Federal Aviation Administration.

FAA (2007) Human Factors Study of Air Traffic Control Safety Alerts White paper and progress report. Washington DC: Federal Aviation Administration (2007).

Kuchar, 2001, March 27-29). *Managing uncertainty in decision-aiding and alerting system design*. Paper presented at the 6th CNS/ATM Conference, Taipei, Taiwan

Lees, N. & Lee, J. D. (2007). The influence of distraction and driving context on driver response to imperfect collision warning systems. *Ergonomics, 30,*1264-1286.

NTSB, (2006) National Transportation Safety Board (NTSB) Safety Recommendation: A-06-44 through –47. Washington D.C.: NTSB.

Rantenen, E. & Wickens, C.D. (2009) Effects of conflict geometry on controller maneuver selection for conflict alerts. *International Symposium on Aviation Psychology*. Dayton: Wright State University

Sorkin, R.D. (1989). Why are people turning off our alarms? *Human Factors Bulletin*, *32*(4), 3-4.

Wickens, C. D., Rice, S., & Levinthal, B. (2009). Imperfect reliability in unmanned air vehicle supervision and control. In A. W. Evans (Ed.), *Human-robotics interaction in future military systems.*Brookfield, VT: Ashgate.

Wickens, C.D., Rice, S. et al (2008). Addressing the alert problem in ATC Facilities: Final report. Report 2008-10-2. Las Cruces NM.: New Mexico State University.

Xiao, Y, Seagull, F.J., Nieves-Khouw, F. Barczak, N., & Perkins, S. (2004). Organizational-historical analysis of the “failure to respond to alarm” problems. *IEEE Transactions on systems, man & cybernetics – part A: systems & humans. 34,* November; pp 772-776.