Effective Notification Systems Depend on User Trust

Scott LeeTiernan, Edward Cutrell, Mary Czerwinski, and Hunter Hoffman

Microsoft Research, One Microsoft Way, Redmond, WA 98052 USA

Abstract: Intelligent messaging systems attempt to determine what information is important to a user's task and when and how to interrupt the user with that information. Such systems are probabilistic, and will be wrong some percentage of the time. We conducted an experiment to assess the impact of variable reliability in a notification system on users' trust and use of the system. We showed how an unreliable system that violates users' trust might lead to abandonment of the system. This disuse behavior pattern persisted despite subsequent improvements in the reliability of the underlying intelligent system. Our results provide guidance in the design of notification user interfaces.

Keywords: Instant messaging, empirical studies, interruptions, intelligent systems

1 Introduction

Intelligent systems (e.g., Horvitz, Jacobs and Hovel, 1999) have recently been developed to help manage the onslaught of incoming information –from email, help systems, scheduling programs, etc.- that pervade today's computing experience. These systems filter messages, making decisions about what information is important, the optimal time for a notification, and how to display the message. Such systems face difficult design problems associated with the user interface model, and there has been surprisingly little research to guide system designers (although see Maglio & Campbell, 2000; McFarland, 1999).

One psychological phenomenon central to designing a good user interface for notifications is the trust the user develops in the system. When an intelligent system makes mistakes, which are especially likely in early, learning stages, users may place less trust in the messaging system. In the extreme case, users may adopt a strategy of completely ignoring or disabling the system.

Maltz, and Meyer (2000) studied a demanding visual task in which potentially beneficial cues were provided. The cues varied in their validity from invalid, moderately valid, to highly valid cues, or there were no cues (control condition). By the second block of trials, only the participants receiving highly valid cues continued to utilize them.

The question of users' trust in the system therefore is important when designing a notification interface for a system known to be somewhat unreliable. If a first impression dominates subsequent interpretations, then the interface should strive to mitigate any negative first impressions.

In this paper, we present a study of behavioral reactions to a system with changing reliability. What happens when a system is initially unreliable (e.g., when learning), but becomes more reliable later on or vice versa? Once users' trust of a notification interface has been broken will they ever reassess system reliability and update their behavior to incorporate new information?

2 Empirical Study

2.1 Procedure

Sixteen participants, ranging in age from 19 to 51, each completed 84 word puzzles similar to the game Boggle. Participants were shown a 6 X 6 grid of letters and given a specified time frame to find the 5-letter solution word beginning with a letter in bold.

Periodically, the system sent notifications to participants that, if responded to, revealed the first three letters of the solution word. Notifications were either subtle or salient. Salient notifications consisted of a large spinning graphic shown near screen-center, accompanied by a loud sound. Subtle notifications were smaller graphics shown at the lower right of the screen, accompanied by a quiet sound. Participants were told that a salient notification meant the computer was confident the message was helpful to the current task, while a subtle notification meant the system was not confident of the helpfulness of the incoming message.

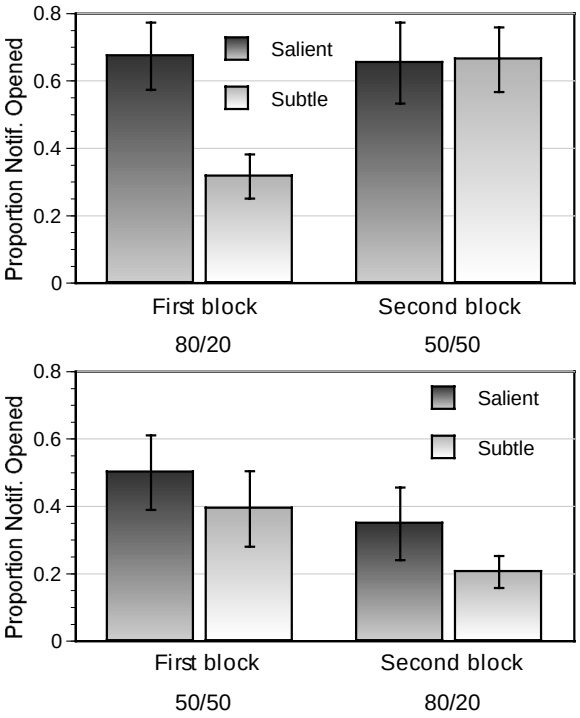

Figure 1: Proportion of notifications opened by congruency and notification type. Top, 80% congruency block first, and bottom, 50% congruency block first.

Despite the instructions, sometimes (incongruent trials) the computer made a mistake (e.g., a salient notification contained an unhelpful message). Each participant experienced 2 blocks of 42 trials each. In one block the computer was correct 80% of the time, and in the other block it was correct 50% of the time.

As a dependent measure we assessed the amount of system use under different levels of system reliability. Our measure of system usage was the percent of notifications opened.

2.2 Results

We performed an analysis of the proportion of notifications presented that participants actually opened. A 2 (block order) x 2 (congruency) x 2 (notification style) ANOVA showed a significant interaction between block order and congruency, $F(1,7)=9.56$, $p<0.01$ (Figure 1). The half of the participants who received the 80% block first clicked on 67% of the salient notifications and 31% of the subtle notifications in that block. Moreover, in the block that followed, participants clicked on more than 65% of *both* types of notifications.

In contrast, those participants who received the 50% congruency block of trials first only clicked on 35% of the salient and 21% of the subtle notifications in the 80% congruency block of trials that followed. In other words, the effect of the low reliability of the expert system was to significantly reduce the number of notifications opened, even when reliability of the system increased substantially in a later block of trials.

3 Discussion

This experiment demonstrated the tradeoffs associated with high versus low system recommendation reliability and user interface presentation style. Generally, the decision to open notifications carried over from the first to the second block, despite marked differences in system reliability.

When participants experienced high system reliability in the first block of trials, they were more likely to continue to trust the system and open notifications in the second block even when the system had become unreliable. Conversely, participants first exposed to an unreliable system lost trust, and opened 20% fewer notifications throughout the rest of the experimental session, despite a dramatic improvement to reliability, supporting Maltz and Meyer (2000).

In designing intelligent notification systems, ensuring high-quality filtering from the outset is of utmost importance because once users perceive a system to be unreliable, it is very hard to win them back. In future work, we hope to take the important next step of evaluating variable notification system reliability in real-world task domains such as word processing and email.

References

Horvitz, E., Jacobs, A. and Hovel, D. (1999). Attention-sensitive alerting. *15th Conf. on Uncertainty and AI (UAI '99)*, Stockholm, Sweden. Morgan Kaufmann Publishers: San Francisco, pp. 305-13.

Maglio, P. P. & Campbell, C. S. (2000). Tradeoffs in displaying peripheral information. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2000)*.

Maltz, M. and Meyer, J. (2000). Cue utilization in a visually demanding task. In Proceedings of the IEA 2000/HFES 2000 Congress. San Diego, CA: Human Factors and Ergonomics Society, pp. 283-284.

McFarlane, D. C. (1999). *Coordinating the Interruption of People in Human-Computer Interaction*, Human-Computer Interaction - INTERACT'99, Sasse, M. A. & Johnson, C. (Editors), 295-303.