Awareness, Training and Trust in Interaction with Adaptive Spam Filters.

Henriette S.M. Cramer, Vanessa Evers, Maarten W. van Someren, Bob J. Wielinga
Human-Computer Studies, University of Amsterdam
Science Park 107
1098 XG Amsterdam, The Netherlands
hcramer@science.uva.nl
+31 (0)20 525 6660

ABSTRACT

Even though adaptive (trainable) spam filters are a common example of systems that make (semi-)autonomous decisions on behalf of the user, trust in these filters has been underexplored. This paper reports a study of usage of spam filters in the daily workplace and user behaviour in training these filters (N=43). User observation, interview and survey techniques were applied to investigate attitudes towards two types of filters: a user-adaptive (trainable) and a rule-based filter. While many of our participants invested extensive effort in training their filters, training did not influence filter trust. Instead, the findings indicate that users' filter awareness and understanding seriously impacts attitudes and behaviour. Specific examples of difficulties related to awareness of filter activity and adaptivity are described showing concerns relevant to all adaptive and (semi-) autonomous systems that rely on explicit user feedback.

AUTHOR KEYWORDS
Adaptivity, autonomy, spam, filters, trainable systems

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Spam filters can help users deal with unwanted, unsolicited email. These filters are a widespread example of systems or agents [10] that make decisions on behalf of the user. In that capacity, spam filters offer a fine opportunity for studying user interaction with and trust in (semi-) autonomous and adaptive systems in real-life contexts. Users for instance need to trust a spam filter's competence, as they risk losing communication that is relevant to them. Trainable filters pose an additional challenge in achieving user trust as they rely on correction of filter mistakes, especially when a filter isn't pre-trained. Users need to spend time and effort to train their filter. They have to be convinced to keep on training their system and need to understand the way the system learns. Studies on spam filters can therefore provide interesting information about the usage of spam filters in specific, but also about interaction with autonomous and adaptive systems in general. Despite this opportunity, spam filter usage is a surprisingly underexplored area.

Ways to develop spam filters are widely available, but information on user interaction with spam filters is scarce. Research is available on interaction with e.g. user-adaptive recommender systems [e.g. 1,2,12], but these studies do not address user attitudes towards filtering systems that have been used in an everyday context over a longer period of time. Research into trust in adaptive filters has focused mostly on collaborative recommender systems (which base recommendations on preferences of other users who are similar to the current user), and e.g. concentrated on trustworthiness of other users of collaborative recommender systems [e.g. 15], instead of trust in the system that makes decisions for its user. Using spam filters might also be seen as more risky, as in contrast to recommenders, they do not recommend additional information items, but instead potentially delete information sent by others.

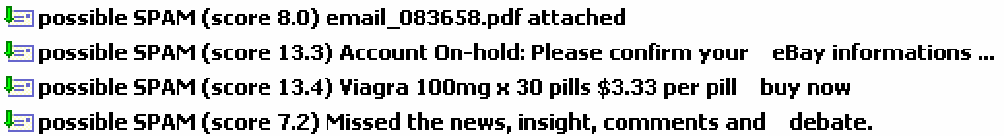
This paper presents an effort to gain more understanding in the ways that users trust systems that make (semi-) autonomous decisions on their behalf by evaluating how people interact with spam filters. It investigates user attitudes and trust toward adaptive, trainable, as well as toward non-adaptive non-trainable filters.

BACKGROUND

Trust and awareness

Whether tasks such as deleting spam email, are delegated to a system is guided by trust in a system [6]. Besides trusting a system's intentions or internal goals, users evaluate the trustworthiness of a system by assessing its competence [4]. Risk and users' own competencies also play a role. Jøsang and Lo Presti [9] point out that even though a user might trust a system, he or she may not choose to rely on it when

CHI 2009, April 4–9, 2009, Boston, MA, USA.
Copyright 2009 ACM 978-1-60558-246-7/09/04…$5.00



possible SPAM (score 8.0) email_083658.pdf attached

possible SPAM (score 13.3) Account On-hold: Please confirm your eBay informations ...

possible SPAM (score 13.4) Viagra 100mg x 30 pills $3.33 per pill buy now

possible SPAM (score 7.2) Missed the news, insight, comments and debate.

Figure 1 Subject lines adapted by the rule-based filter to show an email had been marked as spam

its perceived benefits do not outweigh the risks of using a system. Users will also not depend on a system when they expect to outperform it themselves [3]. Trust especially guides reliance when a complete understanding of a system is difficult [6]. Understanding however also affects acceptance and trust [2,5,12]. While Tullio et al. [13] show that users of intelligent systems can over time counter initial misconceptions about a system's inner workings, they also found the overarching structure of users' mental models are relatively stable. When studying user trust in adaptive systems, awareness and understanding should be taken into account. Users first of all need to be aware that the system exists. This may seem trivial, but in situations outside a lab setting, users are often not aware of adaptive or trainable functionalities of the applications they use (e.g. trainable spamfilter icons that are ignored, menu bar options that suddenly 'disappear'). For adaptive, trainable systems, users may not be aware that they can provide feedback, or may not fully understand what effects training has on the filter. We argue that awareness of the system's existence and training facilities is a relevant pre-condition to studying trust in adaptive and autonomous systems.

Training adaptive systems

To learn which messages are spam and which are not, an adaptive spam filter needs user feedback. The explicit feedback often used for trainable spam filter (e.g. by direct marking of messages as spam or non-spam) offers the advantage of providing a sense of user control, which can positively impact trust [1,7]. Ideally, training would also increase trust through improvement of performance. However, understanding of systems and providing useful feedback to improve their performance can be challenging [16]. Trainable filters also do not immediately offer a high level of performance, but improve over time. The nature of trainable spam filters is that these systems rely on correction of errors, further impeding trust. Unfortunately, trust rapidly decreases when users notice errors and only slowly increases as a system performs without errors [11]. The study reported in this paper aims to investigate user attitudes and behaviours when using trainable and non-trainable spam filters. We are especially interested in understanding the role of trust and the factors that motivate users to spend time and effort training a spam filter system.

METHOD

Data was collected through 30-min. to 1.5-hour sessions combining observation, in-depth interviews and a survey. Participants were observed while using their regular email clients at their own workplace. Participants were asked to refrain from using email on the day of the session so that their usual routine of checking email and dealing with spam messages could be observed. The number of email messages, number of spam messages and the filter settings were recorded. Afterwards, a semi-structured interview was conducted. Twelve open-ended questions concerned users' experiences in interacting with (adaptive) spam filter, their attitudes towards spam and spam filter use. In addition, we asked participants to explain the way they thought their email filter worked and to explain why they did (not) train their spam filters. A survey that measured acceptance of and trust in information filters and in the training of information filters concluded each session. The questionnaire consisted of 7 items relating to the participants' background and 22 items based on [8] and [14], concerning perceived filter usefulness, perceived ease of use, attitude toward the spam filter and dependability of the filter (Table 1). The questionnaire addressed both using and training the spam filter using two separate sections.

Filters

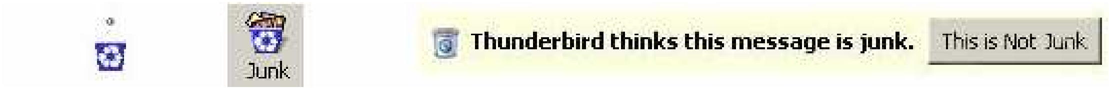
Participants' use of their own spam filters was evaluated. Used were the Mozilla Thunderbird email client's built-in adaptive filter (N=12), a rule-based, non-adaptive filter installed on a central mail server (N=19), or both (N=12). Both filters can label, move or delete spam messages.

The Thunderbird client offered an adaptive, Bayesian spam filter, which can be trained by correcting its mistakes. The filter labels messages as spam both in the list of emails using an icon (Fig. 2, left) and using a message when an email is opened (right). The 'Junk' icon and button can be used to actively (de-)label messages as junk (Fig. 2) (Note: more recent Thunderbird versions feature other icons, participants mainly used 2005 and 2006 versions).

The server-side filter was not adaptive or personalised. The server-side filter assigned scores to emails, based on which and how many of the indicators a message is spam (the server's rules) are fulfilled. If the score is high enough, it is then added to the spam email's subject line (see Fig. 1). Users could not correct the spam filter if it made a mistake, but could add their own server-side rules.

Participants

Forty-three participants took part in the study at their place



Thunderbird thinks this message is junk.

Figure 2 Mozilla Thunderbird a. Junk icon, b. Junk button and c. warning text and correction button.

of employment at two research organisations. Thirty were male. The mean age was 38 (range: 24-59, SD=11.7). 28 worked in an area related to computer science, the others did not. The average number of legitimate emails per day was 18 (range:2-58, SD=15.5), the average of spam emails was a relatively modest 26 (range:0-270, SD =31.8).

Table 1 Example items final scales, participant mean scores and standard deviations. Cronbach's α as reliability measure. All 7-point (0-6) Likert-type scales.

Perceived usefulness of the filter 2 items, α = .869, M=5.31, SD=1.10 e.g. Use of a spamfilter enables me to deal with my email more quickly.
Perceived ease of use filter 2 items, α = .822, M=4.52, SD=1.37 e.g. I find the spamfilter easy to use.
Attitude toward filter M=5.69, SD=.72 Using a spamfilter is a good idea.
Trust in filter 4 items, α = .765, M=3.15, SD=1.27 e.g. I think the spam filter correctly assesses email as spam or non-spam.
Perceived usefulness of training 2 items, α = .738, M=4.85, SD=1.31 e.g. I find it useful to train the spamfilter.
Perceived ease of use of training 3 items, α = .860, M=4.53, SD=1.51 e.g. Training the spam filter is clear and understandable to me.
Attitude toward training M=5.09, SD=1.38 Training the spam filter is a good idea.
Trust in training 3 items, α = .852, M=4.26, SD=1.26 e.g. I trust the outcome of the training process of the spam filter.

RESULTS AND DISCUSSION

Using spam filters

To investigate whether the level of trust in a filter was related to the level of delegation to a filter, Kruskal-Wallis tests were used to compare three groups: participants that allowed their filter to label emails (N=15), to move emails (N=22) or to delete emails (N=4). Significant differences were found for usefulness (H(2)=13.896, p=.001), ease of use (F=3.655, p=.036) and attitude towards the filter (H(2)=11.844, p=.003). Jonckheere's test revealed a significant trend in the data: participants who would allow for more autonomy of the system (from labelling, to moving, to deleting spam) were more positive on usefulness, ease of use of the filter and attitude towards it. Trust however, was not found to be related to the choices for delegation. From the interviews, the social context, in which it was either acceptable or unacceptable to lose email messages from others appeared to play a large role, e.g. participants receiving email from new clients could not afford filter mistakes. 'Critical errors' also appeared influential (e.g. a friend's email marked as spam). Scale might also play a role; most participants received a modest amount of spam and e.g. participants who let their filter(s) delete messages also appeared to receive the most spam.

Comparing adaptive and non-adaptive spam filters

To check whether there was a difference between the filters in reported attitudes and perceptions, three groups were compared: users of the Thunderbird adaptive filter (N=12), a non-adaptive server-side filter (N=14), or both (N=16). Kruskal-Wallis tests did not yield any significant differences between the three groups for perceived usefulness, ease of use, attitude or perceived dependability.

The effects of user feedback

To see whether training of the filter affected trust, we compared users who actively trained or corrected their adaptive filter, or added rules to their server side filter (N=25) with those who did not (N=22). A significant difference was found for trust in the system's training process (U=118.5, p(1-tailed)=.036). Scores for participants who trained their filters were significantly higher (Mdn=5) than for participants who did not do so (Mdn=3.5). These results indicate that a user's decision to train or correct an adaptive spam filter has an important relation with trust. It's not trust in the system that plays a decisive role; instead it is trust in the training process in specific that is associated with training behaviour. During interviews, participants did report that training the spam filter increased their trust in the system as they could notice that it improved. However, even though they had spend considerable effort in training their filters, participants who corrected their filters were not found to delegate 'higher risk' tasks such as automatically moving or deleting spam.

Awareness and user feedback

The observation and interviews yielded interesting insights into the importance of awareness. A number of participants expressed uncertainty about the settings of filters and some worried that there might be more filters active than they knew about. They feared 'invisible filters' might be deleting emails before they even had the chance to exert any control over filter settings. Furthermore, participants often reported other filter settings than actually observed. In this study, all of the participants who had a server-side filter were aware of the filter's activities, while a considerable portion (29%) of the participants who had an active Thunderbird filter were not. Even though the Thunderbird filter showed spam icons, a 'mark as spam' button and a warning message in the content of mails that are labelled as spam (Fig. 2), not all users noticed and recognised these. Results from the observation studies indicated that showing a unambiguous and hard to miss 'possible spam' addition to an email's subject line, as the server-side filter did (Fig 1), worked better to make users aware a spam filter was active than showing an icon in the mailbox list of emails.

Such lack of awareness of participants of both filter activity and interface items led to less-than-optimal training behaviour in a number of ways. First of all, even recognising the filter was active did not guarantee correct usage of interface items related to training. In an extreme case a user did know about the filter and its ability to learn, but did not understand its ability to learn from example spam messages. Instead this user had found a menu-option to add explicit rules and effectively had manually written an own rule-based filter (adding rules on Viagra, V-I-A-G-R-A, etc.), and reported wondering why the system was so user-unfriendly. Sometimes, the spam button and icons were misinterpreted as delete buttons. This led to inadvertent training of the filter when users used the button to delete no-longer needed, but legitimate email. If the interface was understood and participants did train their filter, they still occasionally consciously decided to not mark specific spam messages as such. This decision concerned spam messages that in their opinion were very similar to non-spam messages, and was made to 'not confuse the filter'. Ironically, these messages would be most informative for the filter to improve and not make the subtle mistakes these users were worried about. This clearly indicates a gap in awareness relevant to all systems that rely on explicit feedback. More user support for training has to be provided, in which overall understanding and such boundary cases have to be taken into account.

CONCLUSION

The findings above appear straightforward, but become more interesting when generalised to other adaptive autonomous systems. They show system designers need to pay special attention to ensuring awareness about system activity and adaptivity. Trust in the effectiveness of training was found to play an important role in the user's willingness to invest effort in a system that can be taught to improve over time. Systems that depend on explicit user feedback need to be designed in such a way that this trust is optimised. An overview of filtering activity should be available to the user. How the system learns and can be taught should be made clear, but only on a level necessary to use the right interface items, avoiding problems with complete control [16]. Interface items should be recognisable as specifically dedicated to training of the system. Training a system on 'borderline cases' has to be encouraged when necessary. Risks can perhaps be decreased by providing an opportunity for users to tell the system a case is special, e.g. here by explicitly showing the system which similar messages they are worried about might be inadvertently labelled as spam on the basis of their feedback. Even while some of the found problems may seem mundane, this study shows they are still open challenges, impeding adaptive system success.

The findings reported in this paper indicate that a more positive attitude toward a system and a more positive assessment of a system's ease of use and usefulness increases the likelihood that a user delegates higher risk tasks to a system; in this case automatic deletion of messages marked as spam. Adaptivity did not appear to play a decisive role in reliance on the system. Instead, risks associated with the social context of filtering email appeared more influential. Users' choice to actively offer feedback to (train) an adaptive system relates to trust in the trainability of the filter. It was not directly affected by ease of use, usefulness of or trust in the filter itself. Interestingly while many of our participants invested extensive effort in training their filters, training did not appear to increase reliance on a system. This raises the question whether this investment is having other positive effects on user attitudes. Finally, qualitative findings indicate that facilitating awareness about system activity and adaptivity is extremely important in ensuring trust and useful training behaviour.

REFERENCES

1. Alpert S.R., Karat J., Karat C., Brodie C. and Vergo J.G. User attitudes regarding a user-adaptive e-Commerce web site. *UMUAI 13*, 4 (2003), 373–396.
2. Cramer H., Evers V., Van Someren, M., Ramlal, S., Rutledge, L., Stash, N., Aroyo, L. and Wielinga, B. The effects of transparency on trust and acceptance in interaction with a content-based art recommender. *UMUAI 18*, 5 (2008), 455-496.
3. Dzindolet M., Peterson S.A., Pomranky R.A., Pierce L.G. and Beck H.P. The role of trust in automation reliance. *Int. J. Hum-Comp Stud 58*, 6 (2003), 697–718.
4. Fogg, B.J. and Tseng, H. The Elements of Computer Credibility. In *Proc. CHI'99*, ACM Press (1999), 80–87.
5. Herlocker, J.L., Konstan, J.A., Terveen, L.G. and Riedl, J.T. Evaluating collaborative filtering recommender systems. ACM Trans. *Inform. Syst. 22*, 1 (2004), 5–53.
6. Lee J.D. and See K.A. Trust in automation: designing for appropriate reliance. *Hum. Fact 42*, 1 (2004), 50–80.
7. Jameson, A. and Schwarzkopf, E. Pros and Cons of Controllability. In *Proc. AH'02* (2002), 193-202.
8. Jian, J.Y., Bisantz, A.M., and Drury, C.G. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *Int. J. Cog. Erg. 4*, 1 (2000), 53-71.
9. Jøsang, A. and Lo Presti, S. Analysing the relationship between risk and trust. *In Proc. Trust Man.* (2004).
10. Maes, P. Agents that reduce work and information overload. *CACM 37*, 7 (1994), 30-40.
11. Muir B.M. and Moray N. Trust in automation. Part II. *Ergonomics 39*, (1996), 429-460.
12. Sinha, R. and Swearingen, K. The role of transparency in recommender systems. In *Proc CHI'02*, ACM Press (2002), 830–831.
13. Tullio, J., Dey, A. K., Chalecki, J., and Fogarty, J. How it works: a field study of non-technical users interacting with an intelligent system. In *Proc. CHI'07*, ACM Press (2007), 31-40.
14. Venkatesh, V., Morris, M.G., Davis, G.B. and Davis, F.D. User Acceptance of Information Technology. *MIS Quarterly 27*, 3 (2003), 425-478.
15. Victor, P., Cornelis, C., De Cock, M., da Silva, P.P. Gradual Trust and Distrust in Recommender Systems. *Fuzzy Sets and Systems* (In Press).
16. Waern, A. User Involvement in Automatic Filtering: an Experimental Study. *UMUAI 14*, 2-3(2004), 201-237.