Does higher security always result in better protection? An approach for mitigating the trade-off between usability and security

Caroline Merkel & Rebecca Wiczorek
Berlin Institute of Technology, Germany

Abstract

Security software protects computers against undesired programs, informing users about potential danger. Maximum protection should be achieved with high security levels, which, however, have the disadvantage of generating a high number of unnecessary warnings. A frequent interruption of workflow is perceived as annoying and may even decrease compliance with the security software. The objective of this study was to examine this trade-off between security and usability. While performing a computer game, participants were randomly attacked by a virus. Security software informed them about potential damage. The control group was forced to use a very high security level. The experimental group who was given the opportunity to select a security level chose for medium-high levels. Performance, behavioural and subjective data was collected. The analysis revealed significant differences between the two groups. The experimental group complied more with the system which led to a better performance due to less damage. Furthermore, their acceptance of and trust in the system were higher while their perceived workload was lower. These findings indicate that a reduction of security level might increase overall protection as users are more willing to follow advices given by the security software.

Introduction

The use of web-enabled computers and mobile devices entails the risk of receiving malicious programs, e.g. viruses, Trojans, or worms. To minimise the potential damage resulting from the intrusion of undesired software, security systems such as anti-virus software or firewalls have been developed. They are implemented to warn users against possible hazards. However, the use of security systems does not always result in an increase of safety. Several users do not know how to use protection mechanisms correctly or simply ignore the security system (Sasse & Flechais, 2005). This behaviour is problematical since security systems can only work optimally if users comply with protection mechanisms as intended (Ben-Asher, Meyer, Moeller, & Englert, 2009).

One possible reason for the disuse of such systems may be their lack of appropriateness regarding the trade-off between security and usability. The

In D. de Waard, K. Brookhuis, F. Dehais, C. Weikert, S. Röttger, D. Manzey, S. Biede, F. Reuzeau, and P. Terrier (Eds.) (2012). Human Factors: a view from an integrative perspective. Proceedings HFES Europe Chapter Conference Toulouse. ISBN 978-0-945289-44-9. Available from http://hfes-europe.org

protection mechanisms should be proportional to the severity of the possible threat enabling users to accept the additional time and effort of coping with the security system (Ben-Asher et al., 2009). Therefore it is important to take the context of use into consideration when designing security systems. When entering websites of unauthorized content or installing new software that claims access to other programs, the security system informs users about a potential threat. Usually a pop-up window appears asking whether to continue, to get more information or to stop the current process. Users are thus interrupted in their ongoing activity. When this happens too often it is perceived as poor usability. This impression amplifies, when users realise that a considerable high number of the warnings are false alarms.

However, generating a lot of unnecessary warnings is one characteristic of most warning and alarm systems. As security systems are not perfectly reliable, designers have to choose between two types of erroneous function: either missing a potential threat or the false indication of idem. In order to avoid missing virus attacks, usually a very liberal (i.e. low) threshold is set, especially in safety-critical environments such as control rooms or cockpits. As a result, a high number of unnecessary warnings are generated indicating danger in harmless situations. Unfortunately the high number of unnecessary warnings may cause undesired side effects, as it can lead to unsafe or unproductive user behaviour (Lee & See, 2004).

False alarms can lead to the perception of low system reliability, which may in turn reduce the users' compliance with warnings (Meyer, 2004). Users have been found to respond more slowly to warnings (Getty, Swets, Pickett, & Gonthier, 1995) or to completely ignore them (Bliss, Gilson, & Deaton, 1995). Since these findings indicate the disuse of warning systems, this phenomenon is referred to as the *cry wolf effect* (Breznitz, 1984). Once users ignore security system's warnings, the risk of virus attacks increases. The liberal threshold setting is also known as *engineering fail-safe approach* (Swets, 1992).

One possible reason for the reduction of compliance is a lack of trust in the security system. As research has shown a high number of false alarms can lead to a decrease in users' trust (Madhavan, Wiegmann & Lacson, 2006). The amount of trust on its part has been found to predict the use of security systems (Lee & Lees, 2007). Most trust theories assume that trust is a multidimensional construct (cf. Lee & See, 2004). It develops based on system characteristics such as reliability, intention, utility or transparency (e.g. Lee & Moray, 1992; Muir, 1994; Wiczorek 2011). The perceived low reliability due to many unnecessary warnings may reduce trust. The same is true for the lack of system transparency when users do not understand why the system generates so many false alarms. As a consequence, users may doubt the utility of the system as well.

Even though the problem of reduced trust and decreased compliance as a consequence of many false alarms resulting from a low threshold setting is well known, it is difficult to find alternatives. This is especially true for high-risk environments. However, less safety-critical situations such as private computer use offer more possibilities for designers to find practicable solutions. Thus, designers of such systems should try to establish an acceptable level and type of protection mechanisms in order to mitigate the trade-off between usability and security with the

aim of raising up users' trust and compliance. One strategy may be to leave the decision of the warning system's threshold to the users and let them determine deliberately the ratio of misses and false alarms. Prior research by Moeller, Ben-Asher, Engelbrecht, Englert & Meyer (2011) investigated users' strategies in relation to threshold changes. However, it is not clear whether users agree with the designers' liberal threshold setting or whether they prefer more conservative thresholds despite the increase of misses. Furthermore, research has not addressed the relation of deliberately chosen alarm thresholds and peoples trust in and compliance with security systems. More importantly, it is unknown whether an individually chosen threshold may raise safety or even worsens it. The current study tries to answer these questions.

Current Study

The aim of the current study was to explore one possible solution for the trade-off between usability and security. Therefore participants had to play a computer game with an integrated security system which informed them about potential virus attacks. One group was instructed to use the most liberal of seven possible warning thresholds of the security system, whereas the experimental group could test different thresholds and then deliberately chose a security level. Due to the *cry-wolf phenomenon* subjects were expected to select more conservative (higher) thresholds in order to receive a tolerable amount of warnings. It was investigated how that influences subjective ratings of trust in and acceptance of the system, participants' perceived workload, their compliance, as well as their performance in the computer game.

First, it was hypothesised that the experimental group would experience higher trust in the security system. Being able to choose between different warning thresholds, the experimental group becomes more familiar with the trade-off between misses and false alarms which should result in a better transparency of the system's function. Furthermore, a reduction of unnecessary warnings was expected to lead to a higher perception of system reliability. Additionally, if the participants in the experimental condition are interrupted less often by the security system they are expected to rate system acceptance higher. With a more conservative threshold, the amount of overall warnings decreases. Thus, participants switch less frequently between the two tasks: playing the computer game and coping with the security system. As a consequence, they have to reallocate their attention less often. According to Wickens' model of multiple resources (1992) the general, task-independent mental resource that is responsible for attention allocation is demanded less, which is assumed to result in a lower perceived workload for the experimental group.

The second hypothesis relates to participants' compliance with the security system. Since research has shown that liberal thresholds result in a decrease of compliance with the warning system (Bliss et al., 1995; Getty et al., 1995) the experimental group is expected to follow the security system's advice to a greater extent than the control group.

The third hypothesis regards the performance in the computer game. Since the experimental group is expected to follow the warning system more efficiently whereas the control group tends to ignore the security advices, participants in the experimental condition should be attacked less often by a virus. This should result in less damage, i.e. performance, costs in the computer game. Thus, the experimental group is assumed to have a better overall performance.

Method

Participants

40 students (19 females, 21 males) were randomly assigned and equally distributed to one of the two conditions. Their age ranged from 21 to 30 years (M = 25.35; SD = 2.71). Participants were paid a basic of €8 and received a performance-related maximum bonus of €7. They received €11.74 on average.

Task Environment

The PC-based Tetris micro world developed by Ben-Asher et al. (2009) served as task environment (see Figure 1). Participants played a modified version of the Tetris game positioning the falling bricks in such a way that full rows were built at the bottom of the screen. The aim of the game was to collect as many complete rows as possible since participants earned ten points per row and were paid according to their performance.

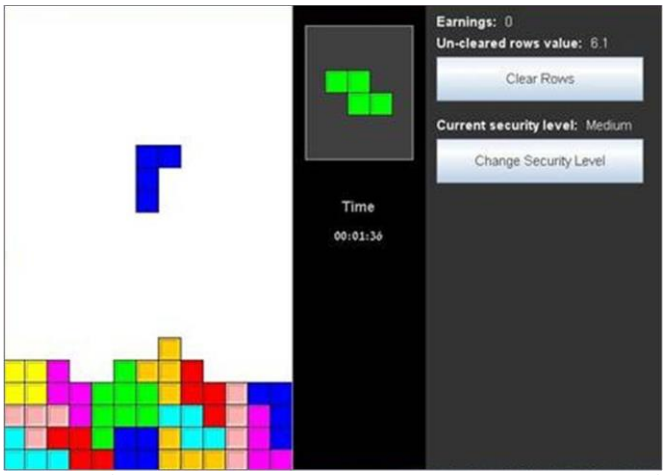

Figure 1. Screenshot of the Tetris microworld by Ben-Asher et al. (2009). On the left: area of current game, in the centre: display of next brick and time, on the right: display of earnings, number of unsaved rows and current security level; buttons for saving rows and changing security level.

In contrast to the original game, completed rows did not disappear automatically but had to be cleared by the participants using the "Clear Rows" button. Pressing this button started a saving process that took 20 seconds impeding the user from continuing to play. This time loss operationalises the poor usability of saving actions similar to those implemented in common anti-virus programs. If the panel filled

completely, it was cleared automatically causing a penalty loss of 50 points in order to prevent participants to use this as a strategy to get rid of incomplete rows.

Random virus attacks during the game caused the deletion of several bricks. Rows that had been completed up to this point were damaged and thus could not be cleared. An integrated security system informed participants about a potential upcoming attack via a pop-up window. In order to manage the security of the game, participants had to continuously decide when to clear rows due to a possible virus attack. Thus, participants were supposed to trade the collection of further rows against the time loss for saving already completed rows.

Design and dependent measures

Participants were divided in two groups. The control group was assigned to use a very liberal warning threshold whereas the experimental could deliberately select one of seven possible thresholds. The dependent measures were trust in and acceptance of the security system as well as compliance with the warnings and performance in the Tetris game.

Subjective data were collected via questionnaires. *Trust* in the security system was measured by means of a multidimensional trust questionnaire (FMV; Fragebogen zur mehrdimensionalen Erfassung von Vertrauen; Wiczorek, 2011). Participants rated trust on four subscales: (1) perceived reliability, (2) utility, (3) intention, and (4) transparency of the warning system. *System acceptance* was measured using items such as “If I played the game again, I would not use the security system” or “If I played the game again, I would choose another security level”. Subjects specified their agreement to the statements on a four-point Likert scale stating if they did not agree, did *rather* not agree, did *rather* agree or did agree. To measure perceived *workload* participants filled in the NASA-TLX (National Aeronautics and Space Administration Task Load Index; Hart & Staveland, 1988).

Behavioural and performance data were collected from the log file of the Tetris micro world software. *Compliance* was operationalised as the storing-rate, i.e. the percentage of storing actions after a warning when completed rows were available. The absolute number of warnings differed between participants in dependence of the security level. A storing action was considered as indicating compliance when it occurred within 21 seconds after a warning appeared which was the fixed period of time between a warning and a virus attack.

Three indicators were calculated to measure *performance*. First, the number of stored rows was regarded. Second, the amount of full panel events was considered. The achieved points were calculated as a third variable by summing up the points per collected row and subtracting the penalties due to full panels from this value. So poor performance which might be caused by a higher amount of virus attacks, was regarded as well.

Procedure

On arrival, participants were briefly instructed about the course of the experiment and filled in a demographic questionnaire. Participants read all instructions on the computer screen of their desks. Paper-and-pencil versions of questionnaires were used.

The experiment was comprised of three sessions (see Figure 2).

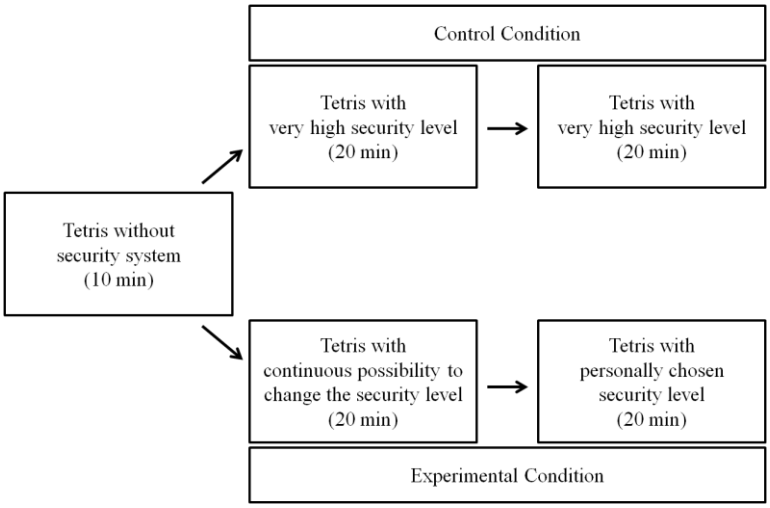

Figure 2. Procedure.

During the first session, participants were introduced to the modified version of the Tetris game without the security system. They played for 10 minutes. During the second session, participants became acquainted with the security system. While the control group was asked to use the highest security level and adverted to the possibility of receiving false warnings, the experimental group was introduced to the possible seven thresholds and their influence on the ratio of false alarms and misses. Participants in the experimental condition were then asked to start with the highest level and subsequently try different levels in order to find out which one they prefer. The control group used the most liberal threshold all along. They played for 20 minutes. During the third session, participants in the experimental group were asked to choose one security level of their preference and keep it throughout the whole 20-minute-period. The control group, on the other hand, was given the highest security level. Participants were informed to be paid according to their performance in this last game. Subsequently, they filled in the NASA-TLX, FMV and the questionnaire of system acceptance.

Results

In a first step the choice of security level and the resulting number of warnings were analyzed in a descriptive way. Afterwards the influence of different security levels on subjective ratings, behaviour and performance was compared between the experimental and the control group using one-tailed *t*-tests for independent groups.

Choice of security level

Participants in the control group could decide on systems configuration choosing one out of seven security levels with *one* being the most conservative and *seven* being the most liberal threshold. In contrast the control group had the directive to use security level *seven*, which generated the highest number of warnings. Choices of the experimental group ranged from security levels of 3 to 6 (M=4.4, SD=1.14) and can be seen in Figure 3. On average, users of the experimental group experienced 25.3 warnings, 16.2 of them were unnecessary because no virus attack occurred. The security system used by all the participants of the control group generated an average of 54.6 warnings, whereof 42.35 were unnecessary.

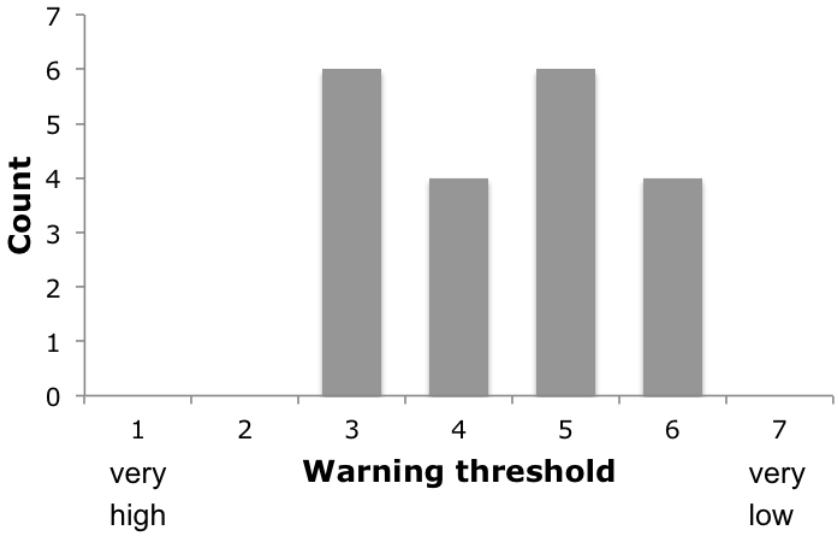

Figure 3. Range of warning thresholds chosen by the experimental group.

The range of thresholds used in the experimental group was quite large. However, as expected users preferred more conservative thresholds and none of them chose the most liberal. These results provide the basis for a comparison of the two groups with regard to their subjective evaluation of the security system, their responses to the warnings and their performance in the task.

Subjective data

Trust. The experimental group trusted the security system more than the control group. The comparison of the two groups revealed significant results for overall trust ratings, $t(38)=3.93$, $p<.001$ (see Figure 4). The same pattern was found for the subscales transparency, $t(38)=3.95$, $p<.001$, reliability, $t(38)=2.69$, $p<.05$, and utility, $t(38)=2.38$, $p<.05$. Regarding the subscale intention the difference between groups was marginally significant in favour of the experimental group, $t(38)=1.74$, $p=.09$. When users were not forced to work with the most liberal threshold they experienced greater trust towards the security system as they perceived it to be more transparent, reliable, and useful as well as to have a more positive intention intention.

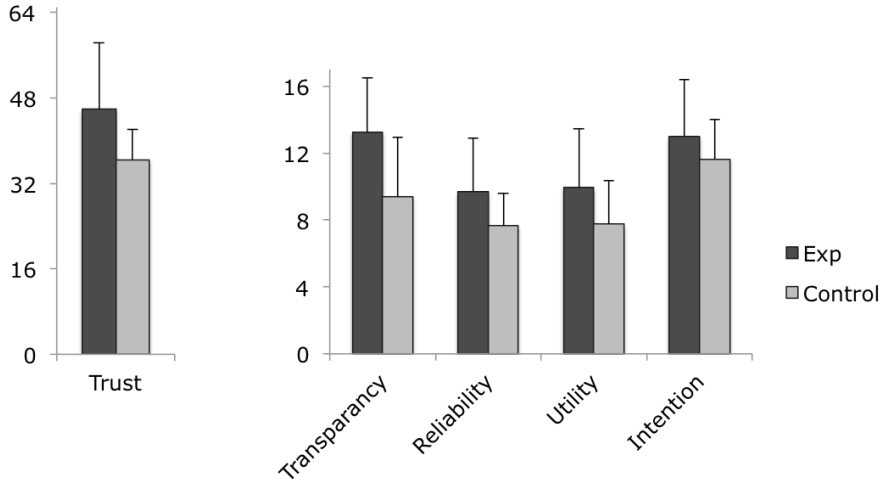

Figure 4. Group comparison of overall trust and trust subscales.

System acceptance. Participants of the experimental group accepted the system better compared to the control group, as their reluctance regarding anticipated future use of the security system was significantly lower, $t(38)=-2.85$, $p<.01$. Furthermore, they were significantly less dissatisfied with the security level they had used compared to the control group, $t(38)=-3.71$, $p<.01$. Users who were forced to work with a system with an extremely liberal threshold were less willing to use the security system again and if they had to, they would choose a different security level whereas the experimental group was more content with the one they have had (see Figure 5).

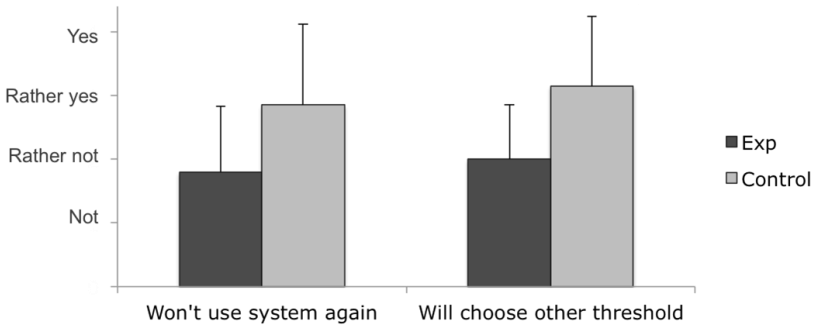

Figure 5. Group comparison of system acceptance.

Workload. The experimental group experienced less workload than the control group, which was reflected in a significant lower overall score of the NASA-TLX, $t(38)=-2.28$, $p<.05$ (see Figure 6).

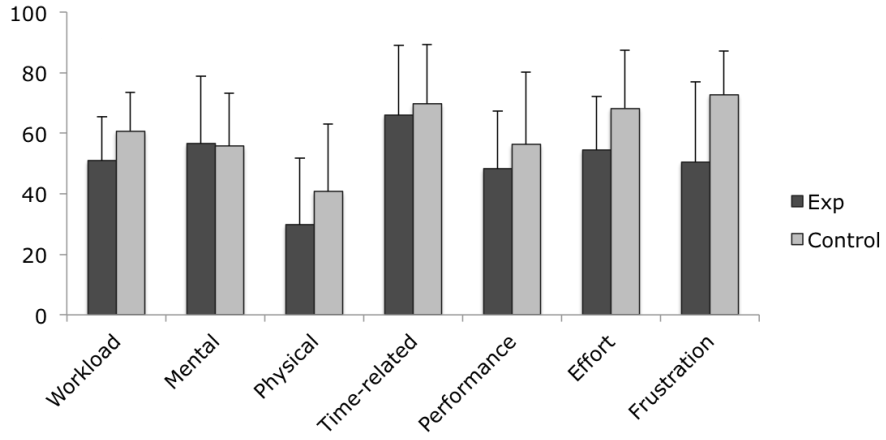

Figure 6. Group comparison of workload.

The difference between groups can be explained by the diverse perceptions of effort, $t(38)=-2,36$, $p<.05$, and frustration, $t(38)=-3.29$, $p<.01$, which were lower for the experimental group. Significant differences could neither be found regarding the mental, $t(38)=.12$, n.s., physical, $t(38)=-1.57$, n.s., and time related, $t(38)=-.56$, n.s., components of workload nor for the performance estimation, $t(38)=-1,17$, n.s. Users who had worked with more conservative thresholds experienced less frustration and less effort when dealing with the task of attention allocation between the Tetris game and the use of the security system.

Compliance

Participants of the experimental group complied more with the security system than participants of the control group. The comparison revealed a marginal significance,

t(38)=-1,87, *p*=.07. Participants, who had chosen their security level, responded to 80% of the given warnings by clearing their collected rows, whereas the control group only responded to 70% of the warnings (see Figure 7).

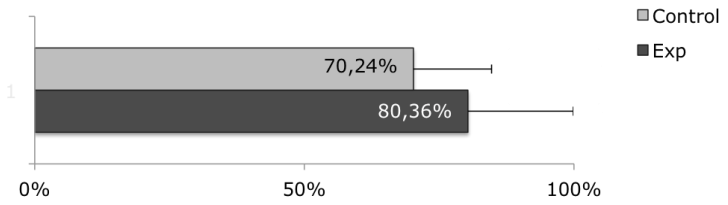

Figure 7. Group comparison of compliance.

Performance

The two groups also differed in regard to performance (see Figure 8). The experimental group had a marginally significant higher number of stored rows, *t*(38)=1.77, *p*=.085, and a significantly higher number of points, i.e. overall performance, *t*(38)=2.11, *p*<.05, than the control group. Differences regarding the number of "panel full" events did not become significant, *t*(38)=-.533; n.s.

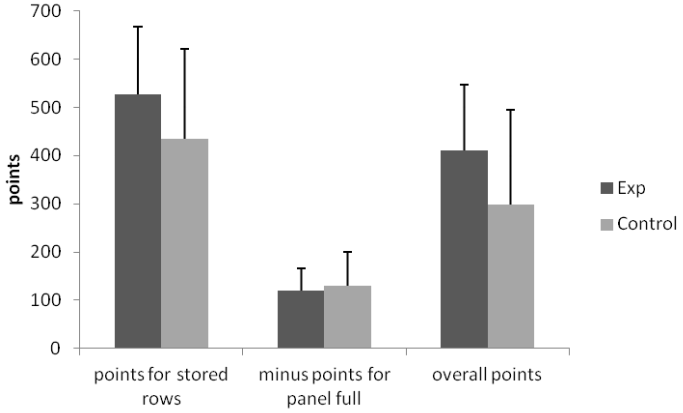

Figure 8. Group comparison of performance in the Tetris game.

However, the fewer number of events of the experimental group led to an advantage with regard to the number of achieved points which contributed to the overall performance. The use of the most liberal warning threshold led to a decrease in performance as users cleared fewer rows before damage. There was an even larger difference in favour of the experimental group in overall performance as for this measure the penalties for filling up the game panel were taken into account as well.

Discussion

The current study was conducted in order to investigate a potential solution for the trade-off between security and usability. If the negative consequences of misses are

less harmful - as it is the case in the context of private computers being exposed to virus attacks - there might be alternative solutions to the *engineer fail-safe approach*. It was assumed that users, given the possibility to choose, would prefer more conservative thresholds leading to fewer unnecessary warnings.

Furthermore, a different threshold, i.e. a reduced number of warnings, was expected to influence users' subjective evaluation of the system as well as their behaviour and their performance. It was hypothesised that fewer unnecessary warnings would lead to a higher amount of trust as users perceive the system to be more reliable. Their acceptance of the system should be greater and their perceived workload should be reduced due to less frequent interruptions of their ongoing task. Furthermore, users of the experimental group were expected to show a greater compliance with the security system, whereas the control group should ignore the warnings to a greater extent. As a consequence of the ignorance of warnings, the control group was assumed to suffer a greater loss in performance due to missed virus attacks.

In order to explore these assumptions, two groups of participants had to perform a computer game which could be attacked by a virus causing performance losses. Both groups were assisted by a security system generating warnings in order to give the users the opportunity to take precaution. The security system used by the control group had a classical liberal threshold setting leading to a high number of warnings, including a high number of unnecessary warnings. The experimental group in contrast was free to choose their preferred threshold after having experienced different threshold settings.

Participants of the experimental group chose medium to high security levels. On average, they experienced 25.3 warnings, whereas, the security system of the control group produced more than double, i.e. 54.6 warnings over a time span of 20 minutes. This experimental manipulation led to differences in regard to subjective, behavioural, and performance data. As hypothesised, the experimental group trusted the system significantly more. This can be ascribed to their higher perceptions of system transparency, system reliability, utility and intention. Higher trust might be the reason why participants of the experimental group ignored fewer warnings than participants of the control group. When users had one or more completed rows and received a warning, the experimental group cleared their rows in 80% of the cases accepting the interruption of 20 seconds. The control group was less compliant as they ignored 30% of the warnings regardless the possible damage of bricks.

The frequent experience of false warnings did not only reduce participants' willingness to follow the systems' advices but also led to a higher perception of workload. Due to the numerous interruptions, participants of the control group were more frustrated and experienced switching between the Tetris game and the security system as more effortful than the experimental group. Even though the control group experienced more workload they did not achieve better performance. Their reduced compliance put them at greater risk to lose bricks due to a virus attack. As a consequence, the control group achieved fewer points than participants in the experimental condition. Since their completed rows were destroyed more often, they

failed to collect as many rows as the experimental group. Even though the difference in penalties for filling up the game panel did not reveal significance, the fewer loss of the experimental group contributed to the overall performance difference as well. Considering their lower performance and their higher workload, it is not surprising that participants of the control group accepted the security system less than the experimental group. Whereas participants of the latter group reported the intention to use the security system in the future, control group participants would rather not. Furthermore, the control group participants indicated that they would change the security level of the system, whereas users of the experimental group were contented with the one they chose before.

Several conclusions can be drawn from this experiment although it raises some questions. First, users of security systems for private use have been found to prefer conservative thresholds as they reduce the number of unnecessary warnings. Even though this result seems less surprising, it is essential for the design of security systems to realize that users' weighting of the different types of errors differs from the classical design perspective. Second, it has been shown that the more conservative threshold raises compliance with the security system indeed, reducing the number of ignored warnings. The third and most important result of this study is the fact that an alternative threshold setting does not necessarily lead to a decrease in safety. On the very contrary, the current study showed that higher compliance can minimize the potential damage as users take better precaution. These results can be used for a more appropriate design of future security systems.

However, the two groups of this experiment did not only differ in regard to the threshold itself. The system used by the experimental group can be seen as an adaptive system offering the possibility to be adjusted according to users' individual requirements. That might have had an impact on users' subjective evaluation and even on their interaction with the system. In fact it is not clear if differences in compliance and subjective ratings are due to the security level itself or rather a consequence of the possibility to choose the threshold individually.

Furthermore, there were found great differences in regard to the preferred security levels within the experimental group. It is not clear whether the choice of the threshold was due to individual differences with regard to the acceptable amount of risk users would take. For an optimal system design, it is essential to investigate the reasons for users' decision, the potential effects of the decision itself as well as the most suitable threshold setting with regard to users' compliance and the resulting safety more thoroughly.

References

Ben-Asher, N., Meyer, J., Moeller, S., & Englert, R. (2009). An experimental system for studying the tradeoff between usability and security (pp. 882-887). *International Conference on Availability, Reliability and Security*, 2009. ARES '09.

Bliss, J.P., Gilson, R., & Deaton, J. (1995). Human probability matching behavior in response to alarms of varying reliability. *Ergonomics*, 38 (11), 2300-2313.

Breznitz, S. (1983). *Cry wolf: The psychology of false alarms*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Getty, D.J., Swets, A., Pickett, R.M., & Gonthier, D. (1995). System operator response to warning of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied, 1*, 19-33.

Hart, S.G. & Staveland, L.D. (1988). Development of NASA-TLX (Task Load Index). Results of empirical and theoretical research. *Advances in Psychology, 52*, 139–183.

Lees, M.N., & Lee, J.D. (2007). The influence of distraction and driving context on driver response to imperfect collision warning systems. *Ergonomics*, *50*, 1264–1286.

Lee, J.D. & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*, 1243–1270.

Lee, J.D. & See, K.A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*, 50-80.

Madhavan, P., Wiegmann, D.A., & Lacson, F.C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, *48*, 241-256.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, *46*, 196-204.

Moeller, S., Ben-Asher, N., Engelbrecht, K.-P., Englert, R., & Meyer, J. (2011). Modeling the behavior of users who are confronted with security mechanisms. *Computers & Security*, *30*, 242-256.

Muir, B.M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*, 1905–1922.

Sasse, M.A. & Flechais, I. (2005). Usable security. Why do we need it? How do we get it? In L.F. Cranor & S. Garfinkel (Eds.), *Security and usability: Designing secure systems that people can use* (pp. 13-30). Sebastopol, CA: O'Reilley Media Inc.

Swets, J.A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, *47*, 522-532.

Wickens, C.D. (1992). *Engineering psychology and human performance* (2nd ed.). New York: HarperCollins.

Wiczorek, R. (2011). Entwicklung und Evaluation eines mehrdimensionalen Fragebogens zur Messung von Vertrauen in technische Systeme. In S. Schmid, M. Elepfandt, J. Adenauer, & A. Lichtenstein (Eds.), *Reflexionen und Visionen der Mensch-Maschine-Interaktion – Aus der Vergangenheit lernen, Zukunft gestalten* (pp. 198–199). 9. Berliner Werkstatt Mensch-Maschine-Systeme, Berlin: VDI.