Explanation and Trust: What to Tell the User in Security and AI?

Wolter Pieters

September 10, 2010

Abstract

There is a common problem in artificial intelligence (AI) and information security. In AI, an expert system needs to be able to justify and explain a decision to the user. In information security, experts need to be able to explain to the public why a system is secure. In both cases, the goal of explanation is to acquire or maintain the users' trust. In this paper, we investigate the relation between explanation and trust in the context of computing science. This analysis draws on literature study and concept analysis, using elements from system theory as well as actor-network theory. We apply the conceptual framework to both AI and information security, and show the benefit of the framework for both fields by means of examples. The main focus is on expert systems (AI) and electronic voting systems (security). Finally, we discuss consequences of our analysis for ethics in terms of (un)informed consent and dissent, and the associated division of responsibilities.

Keywords: actor-network theory, confidence, expert systems, explanation, information security, informed consent, systems theory, trust

1 Introduction

In real life, we are tempted to trust persons if they can explain to us why they do what they do. And we are tempted to trust a car if the dealer can tell us why it is safe (which is harder if you just had to call back lots of cars because of safety issues). This is often how trust appears to work: it requires an *explanation* of the person or thing that we may or may not trust. Such explanations we may simply accept, or we may base our decisions upon them. If you have given me satisfactory explanations in the past, I may even refrain from requesting them in the future.

In this sense, explanation and trust seem to be common partners in everyday life. What we focus on in this paper, is the special case of interactions in the digital environment. Also in the digital world explanation and trust show up together quite often, and in very different domains. Artificial agents need to explain their decision to the user in order to gain trust, and the designers of

secure websites need to explain to the banking client why they can safely do their transactions online.

Trust in digital environments has been called 'e-trust', and the question whether this is possible at all has received considerable attention.[1] Issues that could influence one's opinion here are 1) whether trust is possible without face-to-face interaction and 2) whether artificial agents are capable of trusting and/or being trusted. In the present analysis, we assume that e-trust is possible based on the simplifying assumption that trust refers to "expectations which may lapse into disappointments".[2] We will elaborate on our notion of trust, based on Luhmann's analysis, in further sections. Similarly to the concept of e-trust, we speak about e-xplanation to refer to digital forms of explanations, or traditional forms of explanation that concern digital devices.

In this paper, we will investigate the relation between e-xplanations and e-trust from a philosophical perspective. After discussing the research background and definitions of the necessary concepts (section 2), we will analyse this relation based upon literature study and conceptual analysis (section 3). Following this, we will apply the analysis to both information security (section 4) and AI (section 5). Finally, we discuss the ethical consequences of the analysis (section 6), and draw conclusions on the benefits and limitations of our analysis (section 7).

The contributions of this research are 1) the notion of explanation program and its relation to explanation trees[3], 2) an account of the relation between explanation and trust based on system theory and actor-network theory, 3) the application of this analysis to AI and information security and 4) the ethical implications of the analysis in terms of informed consent.

2 Preliminaries

2.1 E-xplanation research

In artificial intelligence, research has been done into explanation in expert systems. Expert systems are systems that suggest solutions to problems that would normally require a human expert to solve. Such problems may include medical diagnosis, industrial process analysis, and financial decisions. A particular type of such systems are case-based reasoning systems, in which solutions to problems are proposed based on retrieval of similar problems from memory, and adapting their solutions. Explanation in such systems has been addressed by Sørmo et al. [2005] and Roth-Berghofer and Cassens [2005]. In a quite different setting, research has also been done into explanations for belief-desire-intention (BDI) agents in virtual training environments.[4]

Ye and Johnson [1995] give three possible types of explanations in expert systems: traces, justifications and strategies. With traces, a detailed record

[1] Taddeo [2009].
[2] Luhmann [1988].
[3] The latter in a philosophical rather than technical sense, cf. Freuder et al. [2000].
[4] Harbers et al. [2009].

of reasoning steps is given. Justifications focus more on the logical argument, whereas strategies are higher-level approaches that the expert system applies to the information it possesses.

Empirical research into user's trust in agents has revealed some interesting results that easily fit into our analysis. Glass et al. [2008] conclude that trust depends on *granularity* of explanations and on *transparency* of the system. Another study compares different explanation interfaces for recommender systems in terms of user trust.[5] The results suggest that what the authors call an 'organisation-based' explanation does a better job than a simple computational explanation of why a recommendation shows up in the list. In organisation-based approaches, recommendations are categorised according to common features. Benefits of explanations in intelligent systems are discussed by Gregor and Benbasat [1999]. This paper offers an account of why explanations in computer systems are a good idea in the first place, from a psychological perspective.

From a computer security perspective, there is quite a substantial amount of research into trust.[6] Here, the question is how it is possible to communicate the analysis that experts have made of a security-sensitive system to the public. Why is it secure? Or, more appropriately: How is it secure? Thus, it is (implicitly) assumed that explanations are required for trust. Explanations are thought to bridge the gap between "actual security" and "perceived security", which, when taken beyond its common sense meaning, is a philosophical problem in itself.[7]

In this paper, we focus on the case of electronic voting (e-voting). When paper voting was increasingly replaced by electronic voting machines or even Internet voting, this issue has led to debates in various countries. In the USA, public pressure has led to the printing of paper copies of each vote cast on a machine.[8] In the Netherlands, electronic voting has been abolished altogether based on the research and perseverance of a pressure group.[9] Parallel to these developments, new electronic voting schemes were designed in computing science, but the security of such schemes is complicated, and users may not be easily convinced. In the testing trajectory of a Dutch Internet voting system, too complex vote verification procedures reduced trust in the system.[10]

Explanations of security are not the same as usability, which is also important in electronic voting. Of course, easy operation and good instructions on how to use the system are vital, but this problem has been dealt with elsewhere.[11] Here, we focus on responses to questions on how the votes are protected.

In artificial intelligence, explanations are usually provided by the system itself. In information security, explanations are provided by the designers.[12]

[5] Pu and Chen [2006].

[6] Shneiderman [2000]; Fahrenholtz and Bartelt [2001]; Nikander and Karvonen [2001]; Chopra and Wallace [2002]; Oostveen and Van den Besselaar [2004]; Randell and Ryan [2006].

[7] Pieters [2010].

[8] Mercuri [2002].

[9] Gonggrijp et al. [2006].

[10] Hubbers et al. [2005].

[11] See e.g. Bederson et al. [2003].

[12] Even when the system explains, the designer of course designs the method of explanation.

Nonetheless, in *both* artificial intelligence and information security, the role of explanations consists for a major part of acquiring and maintaining the trust of the user of the system. From the AI perspective as well as the information security perspective, there is a need for a better understanding of the relation between explanation and trust. In order to achieve this, we first need to look at definitions of central concepts.

2.2 Central concepts

2.2.1 Explanation

Dictionary definitions of the verb 'explain' acknowledge that explanations may have different *goals*: they may be about describing something in detail, about offering reasons, or about giving instructions on how to do something. We do *not* consider the latter category here. In computer science, this type amounts to explanations on how to use the system, which are instructions rather than explanations in a stricter sense. We focus on the meanings of justification (offering reasons) and transparency (describing in detail).

Roth-Berghofer and Cassens [2005] and Sørmo et al. [2005] distinguish five different explanation goals for case-based reasoning expert systems: justification (explain why the answer is a good answer), transparency (explain how the system reached the answer), relevance (explain why a question asked is relevant), conceptualisation (clarify the meaning of concepts) and learning (teach the user about the domain). Relevance can be seen as a special kind of justification. Conceptualisation and learning have goals similar to instruction, which we said we would not discuss. The remaining two goals, transparency and justification, are the central ones in our framework.

When an explanation is given with respect to a specific goal, certain aspects of it may require further explanation. These are called *subgoals*. In this paper, we make use of *explanation trees* to visualise the relation between explanation goals and subgoals. An *explanation tree* is a tree in which the goals and subgoals of an explanation are ordered systematically (see figure 1). Whereas Freuder et al. [2000] use the concept in a technical sense, we interpret it in the wider context of explaining the decisions or design of a system to the user.

In information security, such trees have a close relation to *attack and defence trees*.[13] An *attack tree* is a tree in the mathematical sense in which possible ways to compromise the security of an information system are systematically ordered. The nodes in the tree correspond to the different steps that an attacker would have to take to break into the system. It is possible to construct a similar tree with defence measures, a *defence tree*.

Similarly, we can construct a pair of a *question* and an *explanation* tree when the concern is not securing the system, but making it able to provide the user with explanations. If the system is not able to give the user sufficient information, the 'attack' has succeeded.

This will be dealt with further in the paper in terms of the concept of delegation.

[13] Schneier [1999]; Mauw and Oostdijk [2006].

[husband] Why did you take the bus?
[wife] Because it was raining and I didn't have an umbrella.

[husband] Why was it raining?

[husband] Why didn't you have an umbrella?

Figure 1: Example explanation tree

As in attack trees, nodes in explanation trees can be AND or OR nodes. An AND node indicates that all connected subgoals need to be realised in order to make the explanation successful; an OR node means that only one of the subgoals needs to be achieved. For reasons of concision, we include both questions and answers in the same tree, using indentation to represent subgoals (i.e. subquestions).

2.2.2 Trust

Trust is a form of self-assurance. It entails reliance upon something else, and the belief that this other will not fail in meeting certain expectations. However, the grounds on which self-assurance is based can be quite different.

In earlier work[14], we distinguished between confidence and trust in information systems based on the work of Niklas Luhmann [1979, 1988]. Confidence means self-assurance of the safety or security of a system without knowing the risks or considering alternatives. Trust means self-assurance by assessment of risks and alternatives. The essential difference is that in case of trust, a *decision* is made to rely or not to rely on the person or system. In daily life, we rely on many expectations without consciously considering the possible impact in case of failure. We have confidence in electricity supply, in people obeying traffic rules, etc. When there are different options possible, such as in choosing a bank for one's savings, a comparison needs to be made, and trust takes the place of confidence.

Similar examples are found in relation to digital devices. If a voting system functions properly, people will have confidence in it without exactly knowing how it works or considering alternatives. When problems arise and e-voting and paper voting are compared as alternatives based on risks assessment, trust (or distrust) takes the place of confidence. The conclusion of our earlier analysis was that by *drawing a clear distinction between e-voting and paper voting*, a pressure group in the Netherlands succeeded in creating consensus on the necessity for voting systems to be trustworthy, rather than reliable only. This is because when two alternatives are compared, their properties need to be visible, which was not the case with the existing e-voting systems.

This analysis can be generalised to other technologies. Computer security experts generally aim at exchanging confidence for trust by explicating the risks of systems. We have seen this with building access cards, privacy in Facebook,

[14] Pieters [2006].

and many more. The question we ask in this paper is which role explanations play in the dynamics of confidence and trust.

2.2.3 Black boxes

In both expert systems and security-sensitive systems, the black box character of systems lacking explanations is often mentioned.[15] The concept of black box then denotes a *lack of visibility or observability.* As it is easily argued that black box systems are not trustworthy either, as we have seen in the previous discussion of confidence and trust, the concept of black box can form an important connection between explanation and trust. However, this concept can mean very different things depending on the language game in which it is used. We therefore need to distinguish these meanings clearly before we proceed.

At least two meanings of black box can be distinguished. In the common sense meaning, a black box is something that outputs something based on certain inputs, but that we do not know the inner workings of. This applies above all to technological artefacts. In a more philosophical sense, as advanced by Bruno Latour [2005] in his actor-network theory, a black box is something that has been 'blackboxed'; a theory or technology of which the supporting network of actants has become invisible. An actant, according to Latour, is anything that participates in actions in a network of relations, and becomes what it is by means of the network. In the latter sense, other phenomena such as scientific theories or political systems can be characterised as black boxes as well. As there is no opportunity to discuss actor-network theory in detail here, the important point to remember is that black boxes need not always be purely technological.

In the first sense, a black box consists only of non-human parts. This is what is usually meant when it is said that electronic voting machines are black boxes. In the second sense, both humans and non-humans can be part of a black box. In this sense, paper voting could be said to be a 'blacker box' than electronic voting, because the network around paper voting has been largely concealed over its relatively long history, hiding risks and security measures inside. It is the latter meaning in which we will use the concept of black box in the following.

Latour associated the process of blackboxing with three other phenomena: translation, composition and delegation. We will use these concepts in our analysis of explanation and trust, but first we will discuss the meaning given to these concepts by Latour.

Composition means that actants in a network form a composite actant to which actions can be attributed. In this way, the government and an electronic voting machine manufacturer can be 'composed' when they address the security of the machines, or an expert system and its designer can be composed when justifying the decisions of the system. Translation denotes that intentions and possibilities for action change when actants join forces. Latour calls these intentions and possibilities the 'action program'. Following a traditional example,

[15] Harris [2003]; Nugent and Cunningham [2005]; Gonggrijp et al. [2006]; Open Rights Group [2007].

a man plus a gun has different action possibilities than a man or a gun alone.[16] Lastly, part of an action program can be delegated to different actants. The responsibility of keeping an eye on the speed limit can thus be delegated to a ramp.

In the following, we will combine the actor-network terminology with the accounts of explanation and trust, in order to get a comprehensive understanding of their relation.

3 E-xplanation and e-trust

In this section, we combine the notions of explanation, trust and black box, as discussed above, in a conceptual analysis of their relation in information systems. The analysis thus combines Luhmann's definitions of trust and confidence with an actor-network view on social relations. This combination is pragmatic rather than aimed at authenticity to the original viewpoints of the sources.

3.1 Explanation programs

In the following, we 'translate' the actor-network concepts to the field of explanation and trust. First of all, the type of action that we are specifically interested in is explanation. Actants can thus be said to have an *explanation program*, i.e. their action program projected on the domain of explanation. When actants are asked to explain something about a theory or system, they have certain intentions and possibilities for explaining in a certain way. This explanation program is translated when actants join forces. For example, the government plus a commercial manufacturer has different explanation possibilities than the government alone when it comes to e-voting: because of commercial interests, it may no longer be able to reveal the source code of the program used.

Responsibilities for explanation can be delegated to other actants, but this also means that the explanation program changes, because the other actants will have different interests and a different understanding of the problem. This holds both for delegation to other humans or organisations, and for delegation to machines. In both cases, the new actant will not have the same capabilities for explanation as the actant that delegated the responsibility for explanation to it. If explanation of decisions is delegated to an expert system, it will have different explanation possibilities than its designer, if only because it has more limited knowledge of the world.

Delegation means exchanging one's own trust for confidence: in delegation, one no longer needs to understand what is to be explained fully oneself. Instead, one has confidence in the actant to which the responsibility of explanation is delegated.

An explanation program can be represented in an explanation tree, as a security policy can be represented in a defence tree. The formal composition

[16] Verbeek [2005].

of explanation programs and explanation trees of different actants, both for cooperation and for delegation, would be a topic for further study.

3.2 Explanation-for-{trust,confidence}

An explanation may have different goals, as we have seen. The most important goals we distinguished are *transparency* and *justification*. Depending on the goal, an explanation can either aim at acquiring confidence or at acquiring trust. *Explanation-for-trust* can thus be contrasted with *explanation-for-confidence*. When we remember that trust entails a decision and confidence does not, the former aims at enabling the user to compare different alternatives by describing them in detail. The latter aims at allowing the user to be confident in using a system, without having to consider different options.

Explanation-for-trust is explanation of how a system works, by revealing details of its *internal* operations. Explanation-for-confidence is explanation that makes the user feel comfortable in using the system, by providing information on its *external* communications. In explanation-for-trust, the black box of the system is opened; in explanation-for-confidence, it is not.

In both meanings of the concept of black box, a black box cannot acquire trust, but only confidence. Black boxes can be explained to their environment, but only as an explanation-for-confidence: the explanation concerns the external communications of the system. Black boxes can be opened when trust is required instead of confidence; this opening produces an *explanation-for-trust* of how the system or network does what it is supposed to do; it reveals part of the inner workings, thereby reveals part of the risks, and thereby trades confidence for (possible) trust.[17]

A network has an explanation program that can reply to questions on transparency and justification. This explanation program is distributed over (delegated to) different actants in the network. If the network can only reply to questions of justification, it can be considered a black box. In such a case, the network can only acquire confidence of the environment. Once trust is required, the black box needs to be opened in order to supply explanations-for-trust, in response to questions of transparency. In the latter case, the system thus needs to be designed in such a way that this is actually possible; this amounts to design for transparency.

If the explanation program of the network *around* a technology is strong enough, the black box of the inner mechanisms of the technology itself may not need to be opened. This was the case with electronic voting in the Netherlands before the efforts of a pressure group forced explanations aiming for transparency.

[17] Following Vico [Berlin, 1976], we may argue that we can understand better something that we have created ourselves than something that is 'given'. In that sense, the human mind is more a black box than a computer system, and we can explain the decisions of a computer system better than those of a human mind. Apparently, this does not mean that we trust a computer more than a human being.

4 Explanation and trust in information security

In the domain of information security, explanation of the security of the system to the user is an important requirement. This is especially true because security is not instantly visible in using a system, as security of a system is not a *functional* requirement. One cannot argue that because the system produces acceptable results, it is therefore secure. Intruders may have broken in and changed results without anyone noticing. Instead, insight must be given in the measures that have been taken to protect the system against intruders.

Users also need to be instructed in how to operate the system securely, for example checking whether they are really communicating with the e-banking site by means of the certificate. This is not the type of explanation we focus on here, as it is another example of explanation meaning instruction. Here, we are interested in the role of explanations that allow the user to form an opinion about (the security of) the system.

In the case of information security, explaining is about describing something in detail, in this case the security measures that are implemented in the system in order to protect the user from harm. *Transparency* is usually seen as the main goal, especially in e-voting. Transparency is then seen as essential for allowing the users to understand what the designers have done to protect them. Whether transparency also contributes to the security of the system itself is heavily debated: some would argue that making the protection mechanisms public will enhance the capabilities of the attackers, whereas other would argue that protection mechanisms can be improved by public scrutiny. Keeping the security mechanisms inside the black box, disabling explanations for transparency, is often referred to as 'security by obscurity'.[18]

The security of a system thus needs to be explained to the user in order to allow her to make an informed decision on whether to use it. The explanation is an explanation-for-trust. This is, of course, only useful if alternatives are available. For example, in the Netherlands, citizens can decide for themselves whether they wish to be a donor, and the information provided is meant to enable them to make a reasonable decision on whether to accept the procedure. In case of an obligatory measure, like an electronic ID card or passport, it is more important to create confidence, as people do not have a choice.

The primary question in security is thus a 'how': the user may request an explanation of how the system is secured, in order to accept using it. However, even if the main goal is transparency, this may involve subgoals that can be of a different type. The explanation programs are usually associated with the designers rather than the system itself. Of course, part of the explanation program can be delegated to the system, e.g. in the form of a help function, as long as the help offered is not *only* instruction on how to use the program, but also information on how it *works* and how it is protected.

Once transparency is established (how?), questions may be asked regarding the reasons for design decisions, including security measures (why?). The ex-

[18] Mercuri and Neumann [2003].

[user] How do you protect your security (transparency)?
[system] Ask the designer!
[designer] By these measures.

> [user] Why these measures (justification)?
> [designer] Because they protect against these attacks.

Figure 2: An example explanation tree for information security.

planation goal then changes from transparency into justification. This can be represented in subgoals in the explanation tree (figure 2). In the tree, although not represented, different explanations are possible for the same question. These explanations may in turn trigger different follow-up questions. In design, such explanation requirements can be anticipated by including explanation trees in the design process, which would be a topic for further research.

As we have argued before, the explanation program in information security is typically delegated to the designers of the system. This means that explanation is not an explicit part of the design of the system, but rather a (business) strategy for dealing with questions about security.

Our case study in the information security field is e-voting. This is the same topic that was addressed in our earlier work.[19] We extend the analysis that was given there with the concepts of explanation and black box.

In electronic voting, two approaches can be distinguished: the Dutch and the British.[20] In the Dutch case, there was one channel available to the citizens to cast their votes, which can be electronic or paper. The local authorities decided which channel would be used (paper has been the only option since a change of law in 2009). In the UK e-voting pilots, multiple channels were offered to the voters, and they could decide themselves which one they wish to use. In the Dutch case, the government needed to create confidence in the systems used, since citizens did not have the choice to go for a different option. In the British case, explanations of the systems could have the role of allowing citizens to choose, enabling trust rather than confidence.

In electronic voting, an explanation-for-confidence of the use of electronic voting machines is that they produce faster results. Or, alternatively, that they are more reliable and accurate than paper voting. Or, alternatively, that they have been tested by an accreditation organisation. In such explanations, the black box of the system is not being opened. The primary goal is justification.

An explanation-for-trust would be an account of the measures that have been implemented to guarantee security. At the highest level of detail, the source code could be made available. The latter, of course, would not be an explanation for the general public, and may therefore not be sufficient to establish public trust in the system. The primary goal in such explanations is transparency.

[19] Pieters [2006].
[20] Pieters and van Haren [2007].

Following this distinction, we can argue that the Dutch government should have had an explanation program that aimed for confidence, whereas the British government should have aimed for trust. Indeed, in the Dutch case, the government for a long time clung to the explanation that there was nothing wrong with the electronic voting machines, even when their security was challenged by the pressure group. From the analysis of explanation in relation to confidence and trust, this was a sensible way to handle the issue: as citizens did not have a choice, confidence in the existing system needed to be upheld.

In the British case, the government could be much more pragmatic: if the security of any of the systems would be challenged, this could be investigated thoroughly, and if the system was found not to be trustworthy, it could be excluded from further pilots.

The situation in the Netherlands can also be explained in terms of black boxes. Following Latour's analysis of technology, an e-voting system is composed of a network of actants, humans and non-humans. Part of the network may be black-boxed; the inner workings are not being observed from the outside.

The e-voting systems that were introduced in the Netherlands in the early nineties were able to hide in the existing black box of the voting system. One may argue that the paper voting system has increasingly become a black box over its relatively long history. The electronic voting machines were put inside without opening it. However, even for paper voting it has not always been like that: major debates have happened on the replacement of oral voting with paper voting.[21].

In any case, the black box was not opened further when electronic voting machines were introduced. An explanation-for-confidence was enough: e-voting would be faster and more accurate. Many e-voting systems of the same generation were black boxes in the common sense meaning. From a Latourian perspective, however, they are part of a network that helps to maintain the black box status of *the whole network*: the inner workings – not only of the technology but of its socio-technical surroundings as well – are kept invisible to the environment, for example by keeping evaluation reports secret.

At the same time, black box voting has become subject to increasing scrutiny, by pressure groups as well as the scientific information security community. These developments require the black boxes to be opened; they have led to a requirement for explanations-for-trust, related to transparency. Now that most countries have been studying their existing e-voting solutions following public pressure, a new generation of voting systems seems to be needed that can actually provide explanations-for-trust (or at least their designers should be able to provide these). This, however, is not trivial, as a bad explanation-for-trust may fail to create trust, and even lead to distrust.

What can happen to e-voting once the trust issues have been solved? If it will be a successful project at all, adjusting the explanation program to the requirements of the environment is necessary. To achieve this, new actors may need to be pulled into the network, which are able to complete the explanation

[21] Park [1931].

tree of the system. Such actors may include pressure groups. Getting the actors in the e-voting network requires making them trust the project. If the supporting network is stabilised in this way, confidence of the environment may be established. Only then can e-voting become a black box in the Latourian sense, by making the explanation program hide the details of the inner workings (again).

5 Explanation and trust in AI

In the case of AI, the most important explanation goal is justification, or *offering reasons for an action*. The reason for a decision, diagnosis or advice needs to be *justifiable* to the user. The primary question is a 'why'; the main goal of explanation in expert systems is justification.

Interestingly, in the history of AI, reasoning traces, which can be characterised as 'how'-explanations, preceded the 'why'-type.[22] The easiest way of telling the user what is going on is just dumping what has been going on in the system. In this sense, the 'why'-explanations are technologically more advanced, as they require a more subtle judgement on what should and what should not be shown to the user. Still, this also holds for the 'how' explanations in security, as we have seen in the previous section.

Even though the primary goal in AI is justification, the other explanation goals for CBR systems can occur as subgoals in an explanation tree with justification as the root goal. For example, in order to justify a decision, it may be necessary to explain certain concepts, or to provide more detail about how the system reached the decision. Thus, whereas the main goal in AI can be characterised as justification, other goals play a role as well.

Subgoals may thus include transparency of system design; from this point on, trust is the issue instead of confidence. For example, if the user does not have confidence in the explanation, she may wish to find out how the system constructed that explanation. She may suspect an error in the system, and will now proceed to request transparency. The explanation goal then changes from justification into transparency. This can be represented in subgoals in the explanation tree (figure 3).

Note that the latter question cannot be answered from the explanation program of the machine itself. Usually, answering this question should be done by the designer, except when it has been delegated to the machine via a help function. Note also that there is an analogy between explanations in AI and a common distinction in philosophy of science: the distinction between the context of discovery and the context of justification. Explanations-for-confidence then correspond to the context of justification (of a decision), whereas explanations-for-trust correspond to the context of discovery (of a decision).

In AI systems, the black box character is not necessarily a problem. As long as the users have confidence in the decisions of the system, they may not be interested in how it works. Therefore, the explanations of expert systems are

[22] Ye and Johnson [1995].

[user] Why did you make this decision (justification)?
[system] Because this and this is the case.

[user] How did you make this decision (transparency)?
[system] By these and these steps.

[user] Why these and these steps (justification)?
[system] Ask the designer!
[designer] Because I used the approach from this paper.

Figure 3: An example explanation tree for an expert system.

mainly explanations-for-confidence. Only when the user suspect that something is wrong, transparency will be required by means of explanations-for-trust.

The explanation trees in artificial intelligence are in a way mirrored with respect to information security. In security, justification emerges as a subgoal when an answer to a transparency question is not sufficient to the user. In AI, transparency emerges as a subgoal when an answer to a justification question is not sufficient. This mirror effect is one of the interesting results of our analysis. To understand the consequences of this result, a further dialogue between security and AI on the topic of explanation would be beneficial.

If expert systems can reach a level of explanation that creates as much confidence in these systems as we have in people, they may become increasingly blackboxed phenomena in our society. The need for knowing precisely how they work may become less pronounced, even if we know more about how they work than we know about how people work, for we designed expert systems ourselves.

6 Ethical consequences

The analysis of explanation and trust has ethical consequences when we connect it to the notion of informed consent, which can be defined as "an autonomous authorisation by a patient or subject".[23] Although often seen in a medical or research context, the notion is important to understand the meaning of explanation and trust for responsibilities. In the present case, this amounts to the explanation of the system to the user, and the object of consent is the use of the system (or its outputs). The main question here is what can be said to be *informed* consent given the characteristics of the explanation of an IT system, and what needs to be denoted rather as uninformed consent, informed dissent, or uninformed dissent. This has consequences for responsibility, as we will see.

Our point of view here is that 'informed' does not merely indicate that *sufficient* information has been given, but also that the *type* of explanation is justifiable and that *not too much* information is given. This is directly related to the concepts of explanation-for-trust and explanation-for-confidence, as the goals of these types of explanations are different. One cannot speak

[23] See e.g. Faden et al. [1986].

about informed consent if one gives too little information, but one cannot speak about informed consent either if one gives too much. Indeed, giving too much information might lead to uninformed dissent, as distrust is invited by superfluous information. When the user has a choice between different alternatives, explanations-for-trust needs to contribute to the *understanding* of the issues by the user. When there is only one sensible option, explanations-for-confidence can help in *justifying* it to the user. If an explanation-for-confidence does not suffice, and the user wishes to consider alternatives anyway, the system should be able to switch to an explanation-for-trust.

The characteristics of the explanations given by IT systems may have consequences for responsibility. If an acceptable kind of explanation is given, and the user trusts the application based on the explanation (informed consent), the user can be said to share the responsibility for the consequences of using the system.

The question of responsibility holds both for security and for AI. If the designers of a secure system can explain security measures and remaining risks to the user (explanation-for-trust), the user can be said to have a reasonable choice in deciding to use the system or not. Given the explanation, the user will not be able to hold the system (or its designers) responsible for security failures, because she has been given proper information about security measures and remaining risks. In such a case, responsibility for the risks could be said to rest with the user (even though legislation may judge otherwise).

In AI, a user of an expert system can be held responsible for a decision made with use of the system, as long as the user has a reasonable way of knowing whether the decision proposed by the system is sensible (explanation-for-confidence). A decision or diagnosis proposed by the system, when accompanied by a satisfying explanation, will keep the user responsible for accepting or rejecting the proposed solution, and thereby avoid users shirking their own responsibility.

These concepts will become increasingly important with the advent of ambient intelligence[24], which exhibits both the features of AI and security-sensitive systems. When everything in our environment is collecting information about us and making decisions for us, we will need a way of consenting to what is happening, or we will not be responsible for anything. This makes a remaining question quite urgent, which is how the socio-technical system around information systems can be designed such that the required explanations can be provided. It is important to avoid the pitfalls of explanation there.

There are two ways in which explanations can miss their goal. Too little detail does not explain-for-*trust*: it fails to open the black box, by only providing superficial reasons.[25] These reasons are usually 'why'-explanations instead of

[24]Cf. Brey [2005]: "Using smart objects requires a basic trust in their judgments, and if these judgments conflict with the users own judgments or intuitions, then the user has to choose whether to rely on herself or on a piece of technology that may or may not know her better than she does herself."

[25]Tavani [2004] provides an interesting discussion of the relation between informed consent and 'opacity', which is comparable to 'blackboxness'.

level of detail	result
too low	explanation fails
low (why?)	explanation-for-confidence, justification
high (how?)	explanation-for-trust, transparency
too high	explanation fails

Table 1: Different levels of detail in explanations

'how'. For example, the government may say that the e-voting systems are secure because they have been accredited. Such explanations may contribute to confidence (and were helpful in the Dutch case), but fail when trust is required, because the black box is not being opened. Too much detail, on the contrary, does not *explain*-for-trust. It fails to make the system comprehensible, because the user is not capable of processing the information at this level of detail.

A too detailed explanation-for-confidence may fail to reach its goal, because it does not explain-for-*confidence*. It aims for trust instead of confidence, by opening the black box of the system. For example, a system may provide a complete reasoning trace when only some indications are required by the user in order to provide her with confidence. In that case, it may even *decrease* confidence. On the other hand, too little detail will not *explain*-for-confidence.

Explanations, therefore, should 1) aim for the right goal (why or how) and 2) carry the right amount of information, in order to provide informed consent to the user, and thereby keep (human) responsibilities clear. Thus, the level of abstraction on which the explanation is given needs to be right in order to speak about informed consent of the user. We can map levels of detail to different results of explanations (table 1).

All of this, obviously, does not mean that designers will no longer be responsible for what their systems do, as long as they have consent from the user. On the contrary, the designers are responsible for designing their systems in such a way that responsible behaviour by their users is encouraged. But users can only act responsibly if they have access to the right explanations.

7 Conclusions

In this paper, we analysed the relation between explanations and trust in information systems, in particular security-sensitive applications and expert systems. From the literature, we took the distinction between confidence and trust, different explanation goals and Latour's concepts of action program, translation, composition, blackboxing and delegation. Combining these in a conceptual analysis, we introduced the notions of *explanation program*, *explanation-for-confidence* and *explanation-for-trust*.

The framework helps us to make clear what we mean when we say that a system has to be able to explain things to the user, or that the system itself

needs to be explainable. Our analysis illuminates the difference between the use of explanations in AI and the use of explanations in information security.

In information security, explanation is mostly aimed at transparency with respect to security measures; this requires opening the black box of the system. In AI, explanation is mostly used to give the user confidence in the decisions of the system. This does not require opening the black box. The user is generally not interested in how the system reached the decision, but primarily in why it is judged to be a good decision.

We discussed that a bad explanation-for-trust may fail to create trust: too little detail does not explain-for-*trust*; too much detail does not *explain*-for-trust. A too detailed explanation-for-confidence may fail to reach its goal, because it does not explain-for-*confidence*; it aims for trust instead of confidence (such as a complete reasoning trace when only some indications are required); in that case, it may even destroy confidence. Too little detail does not *explain*-for-confidence. Only if the right kind of information is given can informed consent on using the system and its outputs be established, and can responsibility be clearly allocated.

The relation between explanation and trust is especially critical in the case of e-trust, since in that case, other mechanisms that relate to embodied presence are unavailable. Therefore, explanations may be an important prerequisite for the building of e-trust. In that case, the properties of the explanation programs, and the associated modes of trust, are vital for assigning responsibilities.

In this paper, we focused on trust of the user in the system. When explanations need to be given not only to humans but also to computer agents, explanations will probably take a different form. How the difference between confidence and trust can be applied in such a setting, and whether mutual trust between artificial agents can be addressed from the perspective of explanations, are interesting questions for future research.

We hope that the concepts we introduced are able to generate lively discussions on implementations of technology and the associated explanation obligations in general. Do not hesitate to contact us for further explanation on how and why we devised this framework.

Acknowledgements

The author wishes to thank Maaike Harbers for useful comments on drafts of this paper, and Jörg Cassens for initial discussions on the topic. This research is supported by the research program Sentinels (www.sentinels.nl). Sentinels is being financed by Technology Foundation STW, the Netherlands Organization for Scientific Research (NWO), and the Dutch Ministry of Economic Affairs. Part of this research was done while the author was employed at Radboud University Nijmegen and supported by a Pionier grant from NWO, the Netherlands Organisation for Scientific Research.

References

B.B. Bederson, B. Lee, R.M. Sherman, P.S. Herrnson, and R.G. Niemi. Electronic voting system usability issues. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 145–152. ACM New York, NY, USA, 2003.

I. Berlin. *Vico and Herder: Two Studies in the History of Ideas*. Hogarth, London, 1976.

P. Brey. Freedom and privacy in ambient intelligence. *Ethics and Information Technology*, 7(3):157–166, 2005.

K. Chopra and W.A. Wallace. Trust in electronic environments. In *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*, 2002.

R.R. Faden, T.L. Beauchamp, and N.M.P. King. *A history and theory of informed consent*. Oxford University Press, USA, 1986.

D. Fahrenholtz and A. Bartelt. Towards a sociological view of trust in computer science. In M. Schoop and R. Walczuch, editors, *Proceedings of the eighth research symposium on emerging electronic markets (RSEEM 01)*, 2001.

E.C. Freuder, C. Likitvivatanavong, and R.J. Wallace. A case study in explanation and implication. In *CP2000 Workshop on Analysis and Visualization of Constraint Programs and Solvers*, 2000.

A. Glass, D.L. McGuinness, and M. Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 227–236. ACM, 2008.

R. Gonggrijp, W.-J. Hengeveld, A. Bogk, D. Engling, H. Mehnert, F. Rieger, P. Scheffers, and B. Wels. Nedap/Groenendaal ES3B voting computer: a security analysis, October 6 2006. Availabe online: `http://www.wijvertrouwenstemcomputersniet.nl/images/9/91/Es3b-en.pdf`, consulted June 25, 2010.

S. Gregor and I. Benbasat. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, 23(4):497–530, 1999.

M. Harbers, K. van den Bosch, and J.J. Meyer. A study into preferred explanations of virtual agent behavior. In Z. Ruttkay, M. Kipp, A. Nijholt, and H.Vilhjhalmsson, editors, *Intelligent Virtual Agents 9th International Conference*, volume 5773 of *LNCS*, pages 132–145. Springer, 2009.

B. Harris. *Black Box Voting: Vote Tampering in the 21st Century*. Elon House/Plan Nine, 2003.

E. Hubbers, B. Jacobs, and W. Pieters. RIES – Internet voting in action. In R. Bilof, editor, *Proc. 29th Annual International Computer Software and Applications Conference, COMPSAC'05*, pages 417–424. IEEE Computer Society, July 2005. ISBN 0-7695-2413-3.

B. Latour. *Reassembling the Social: An Introduction to Actor-Network-Theory.* Oxford University Press, Oxford, 2005.

N. Luhmann. *Trust and power: two works by Niklas Luhmann.* Wiley, Chichester, 1979.

N. Luhmann. Familiarity, confidence, trust: problems and alternatives. In D. Gambetta, editor, *Trust: Making and breaking of cooperative relations.* Basil Blackwell, Oxford, 1988.

Sjouke Mauw and Martijn Oostdijk. Foundations of attack trees. In D. Won and S. Kim, editors, *Proc. 8th Annual International Conference on Information Security and Cryptology, ICISC'05*, volume 3935 of *LNCS*, pages 186–198. Springer, 2006. URL `http://www.icisc.org/`.

R.T. Mercuri. A better ballot box? *IEEE Spectrum*, 39(10):26–50, 2002.

R.T. Mercuri and P.G. Neumann. Security by obscurity. *Communications of the ACM*, 46(11):160, 2003.

P. Nikander and K. Karvonen. Users and trust in cyberspace. In B. Christianson, B. Crispo, J.A. Malcolm, and M. Roe, editors, *Security Protocols: 8th International Workshop, Cambridge, UK, April 3-5, 2000, Revised Papers*, volume 2133 of *LNCS*, pages 24–35. Springer, 2001.

C. Nugent and P. Cunningham. A case-based explanation system for black-box systems. *Artificial Intelligence Review*, 24(2):163–178, October 2005.

A.M. Oostveen and P. Van den Besselaar. Security as belief: user's perceptions on the security of electronic voting systems. In A. Prosser and R. Krimmer, editors, *Electronic Voting in Europe: Technology, Law, Politics and Society*, volume P-47 of *Lecture Notes in Informatics*, pages 73–82. Gesellschaft für Informatik, Bonn, 2004.

Open Rights Group. May 2007 election report: Findings of the open rights group election observation mission in scotland and england, June 2007. Available online: `http://www.openrightsgroup.org/wp-content/uploads/org_election_report.pdf`, consulted June 25, 2007.

J.H. Park. England's controversy over the secret ballot. *Political Science Quarterly*, 46(1):51–86, March 1931.

W. Pieters. Reve{a,i}ling the risks: a phenomenology of information security. *Techné*, 2010. Forthcoming.

W. Pieters. Acceptance of voting technology: between confidence and trust. In K. Stølen, W.H. Winsborough, F. Martinelli, and F. Massacci, editors, *Trust Management: 4th International Conference (iTrust 2006), Proceedings*, volume 3986 of *LNCS*, pages 283–297. Springer, 2006.

W. Pieters and R. van Haren. Temptations of turnout and modernisation: E-voting discourses in the UK and The Netherlands. *Journal of Information, Communication and Ethics in Society*, 5(4):276–292, 2007.

P. Pu and L. Chen. Trust building with explanation interfaces. In *Proceedings of the 11th international conference on Intelligent user interfaces*, page 100. ACM, 2006.

B. Randell and P.Y.A. Ryan. Voting technologies and trust. *IEEE Security & Privacy*, 4(5):50–56, 2006.

T.R. Roth-Berghofer and J. Cassens. Mapping goals and kinds of explanations to the knowledge containers of case-based reasoning systems. In H. Muños Avila and F. Ricci, editors, *ICCBR 2005*, volume 3620 of *LNCS*, pages 451–464. Springer, 2005.

B. Schneier. Attack trees: Modeling security threats. *Dr. Dobb's journal*, December 1999.

B. Shneiderman. Designing trust into online experiences. *Communications of the ACM*, 43(12):57–59, 2000.

F. Sørmo, J. Cassens, and A. Aamodt. Explanation in case-based reasoning: perspectives and goals. *Artificial Intelligence Review*, 24(2):109–143, 2005.

M. Taddeo. Defining trust and e-trust: Old theories and new problems. *International Journal of Technology and Human Interaction*, 5(2):23–35, 2009.

H.T. Tavani. Genomic research and data-mining technology: Implications for personal privacy and informed consent. *Ethics and information technology*, 6 (1):15–28, 2004.

P.P.C.C. Verbeek. *What things do: Philosophical Reflections on Technology, Agency, and Design*. Pennsylvania State University Press, 2005.

L.R. Ye and P.E. Johnson. The impact of explanation facilities on user acceptance of expert systems advice. *Mis Quarterly*, pages 157–172, 1995.