# Formalizing Excusableness of Failures in Multi-Agent Systems

Eugen Staab and Thomas Engel

Faculty of Sciences, Technology and Communication, University of Luxembourg,
6, rue Richard Coudenhove Kalergi, L-1359 Luxembourg
Tel.: 00352-466644-5319
{eugen.staab,thomas.engel}@uni.lu

**Abstract.** To estimate how much an agent can be trusted, its *trustworthiness* needs to be assessed. Usually, poor performance of an agent leads to a decrease of trust in that agent. This is not always reasonable. If the environment *interferes* with the performance, the agent is possibly not to blame for the failure. We examine which failures can be called *excusable* and hence must not be seen as bad performances. Knowledge about these failures makes assessments of trustworthiness more accurate. In order to approach a formal definition of *excusableness*, we introduce a generic formalism for describing environments of Multi-Agent Systems. This formalism provides a basis for the definition of *environmental interference*. We identify the remaining criteria for *excusableness* and give a formal definition for it. Our analysis reveals that *environmental interference* and a strong commitment of the performing agent do not suffice to make a failure *excusable*.

**Keywords:** Multi-agent systems, trust, dynamic environments, service-oriented computing, mobile ad hoc networks.

"Both types of mistake - trusting too well and not well enough - can be costly." [1]

## 1 Introduction

In Multi-Agent Systems (MAS) the assessment of agents in terms of trustworthiness is an important topic in research [2, 3, 4, 5, 6]. Many trust-models use experiences made with interaction partners as basis for assessments. But – as Ramchurn et al. point out [2] – only few models account for the context in which the behavior of other agents was assessed.

In [7] Falcone and Castelfranchi present a comprehensive study of trust in Multi-Agent Systems. They mention that "when $x$ trusts the internal powers of $y$, it also trusts his abilities to create positive opportunities for success, to perceive and react to the external problems". A profound investigation on the causality in the other direction though could not be found in literature, i.e. which experiences should be used to diminish or augment the trust in $y$.

A. Ghose, G. Governatori, and R. Sadananda (Eds.): PRIMA 2007, LNAI 5044, pp. 122–133, 2009.

An example shows why this is important. Assume the trustworthiness of agent $a$ and agent $b$ is assessed at time $t_1$ and time $t_2$ respectively. Even if the performances of agents $a$ and $b$ are the same, if the environment has changed between time $t_1$ and $t_2$, agent $a$ may succeed while $b$ fails. The agent who was assessed during the disadvantageous setting is most likely trusted less than the other one. This is an incorrect perception of reality and a sophisticated trust-model should not have this characteristic.

We therefore introduce the property of *excusableness*. If an agent performs well but the environment interferes (such that the agent does not accomplish the task), the failure should not generally be classified as bad performance of the agent; in some cases the failure is *excusable*. *Excusable* failures are those which are wrongly used for diminishing the trust in the failing agent. So if the evaluation of experiences is seen as classification problem then *excusable* failures would belong to the set of false negatives[1]. In our work this set is identified by developing a theoretical understanding of *excusableness* and by providing a definition for it. This definition and all requisite definitions are given from an objective or external viewpoint. In this way, the paper contributes to the conceptual analysis of trust and its assessment. To make the results suitable for the use in formal trust-models, all definitions are presented both informally and formally – by the use of predicate logic.

By the definition of *excusableness*, knowledge about *excusable* failures can make assessments of trustworthiness more accurate. That is important both for the assessing agent and for the agent being assessed. The paper does not address how this knowledge can be obtained, i.e. how *excusableness* can be measured in practice. Further we do not investigate moral or legal aspects of the property.

For many definitions in this paper we provide examples for the purpose of illustration and to demonstrate the appropriateness of the formalisms. The setting for the examples are Mobile Ad Hoc Networks (MANETs) [8]. A MANET consists of mobile nodes communicating over wireless technologies. The network does not rely on a pre-existing infrastructure as it cannot be predicted how the nodes will move over time.

In the next section we introduce a generic formalism for describing environments of Multi-Agent Systems. Using this formalism, *environmental interference* and *riskiness* of performances are introduced in Sect. 3. In Sect. 4 the criteria for *excusableness* are derived and a formal definition is given. We have a look at the work that has been done on related topics in Sect. 5. Conclusions are drawn and an outlook on future work is given in Sect. 6 and 7 respectively.

## 2 Formalizing the Environment

An environment is described through a set of *environmental variables* and the relations between them. The set of variables is constructed by taking the power set of all *atomic variables* where *atomic variables* are those which are not composed

[1] False positives cannot be defined symmetrically and therefore shall be investigated separately.

from other variables. Agents are variables, too. Each environmental variable has a certain *intensity* depending on time. *Intensity* is a degree of strength (over time) with which the variable influences its environment. The relations between the environmental variables are expressed as their *exposure* among each other. A variable may be exposed to another one heavily or not at all. The influence one variable has on another one could then be calculated on the basis of intensity of the influencing variable and the exposure of the influenced variable towards the influencing variable. However this will not be an issue of this paper.

An environment $\Psi$ can formally be described as a triple $\langle \Omega, I_\Omega, E_\Omega \rangle$, where:

- The set $\Omega$ contains all environmental variables $\omega$. The set $\Omega$ is constructed as power set of all *atomic variables*. Each environmental variable is then a set consisting of atomic variables and/or other composed variables. The set of agents (each agent is a variable), denoted by $\mathcal{A}$, is contained in $\Omega$, so $\mathcal{A} \subseteq \Omega$.
- The set $I_\Omega$ contains for each variable $\omega \in \Omega$ a function which represents $\omega$'s intensity over time. Every such function $\iota_\omega : \mathbb{R} \to [0, \infty)$ maps points in time $t$ to the intensity of the environmental variable. The intensity of a variable can range from its minimum 0 (no intensity at all) to $\infty$. We call a function $\iota_\omega$ the *intensity-function of variable* $\omega$.
- The set $E_\Omega$ contains for each variable $\omega_i \in \Omega$ a function which represents the exposure of $\omega_i$ to all other variables. For a point in time $t$ and an environmental variable $\omega_j$ such a function $\varepsilon_{\omega_i} \in E_\Omega$ returns the exposure of $\omega_i$ to $\omega_j$ at time $t$. The function is of the form $\varepsilon_\omega : (\Omega * \mathbb{R}) \to [0, 1]$. A variable can be totally exposed to another one (1) or not at all (0). We call a function $\varepsilon_{\omega_i}$ the *exposure-function of variable* $\omega_i$.

*Example 1 (Environment).* Let the nodes of a MANET be the set of agents $\mathcal{A}$. As agents are environmental variables each one has an intensity-function which represents the radiation power of an agent's network card. When an agent $a$ doesn't send at time $t$, its intensity-function is zero, i.e. $\iota_a(t) = 0$. The exposure between two agents $a$ and $b$ depends on their distance. If their distance changes over time the exposure-functions $\varepsilon_a(b, t)$ and $\varepsilon_b(a, t)$ change (in the same way). So the movement of variables can completely be described by the exposure-functions. Other variables are for instance walls; their intensity-function is constant and depends on the material and the thickness of the wall.

## 3 Interference and Riskiness

In this section we introduce the concepts of *interference* and *riskiness*. Both are important for the definition of *excusableness* given in Sect. 4.

### 3.1 Interference with Actions

Interference addresses the impact the environment has on the actions of an agent. More precisely it describes the case in which the environment is the *critical* cause

for a failure in an agent's performance. *Critical* means that if this cause had not been given, the agent would have succeeded. If there are $n$ environmental variables $\omega_1, ..., \omega_n$ only together causing a failure, then the variable $\omega = \bigcup_{i=1}^{n} \omega_i$ interferes with the action.

To express *interference* formally we first need to define two predicates, namely $perform$ and $success$.

**Definition 1 (Performance).** *Let $\mathcal{A}$ be the set of agents within the set of variables of $\Psi$ and $a \in \mathcal{A}$. The predicate $perform(a, \alpha, t, t', \Psi)$ marks the attempt of agent $a$ to perform an action $\alpha$ in an environment $\Psi$ within time-frame $[t, t']$, i.e. the performance starts at time $t$ and ends at time $t'$.*

The predicate $perform$ does not say *how much* it was attempted to perform $\alpha$; so this effort is left unparameterized for the sake of simplicity. It suffices to assume that if this predicate is used more than once in a formula then the effort is the same for each agent for which the predicate is used.

*Example 2 (Performance).* In a MANET an agent attempts to send a packet if it sends the packet with the radiation power of its network card.

**Definition 2 (Success).** *Let $\mathcal{A}$ be the set of agents within the set of variables of $\Psi$ and $a \in \mathcal{A}$. The predicate $success(a, \alpha, t, \Psi)$ marks the successful accomplishment of an action $\alpha$ through agent $a$ in an environment $\Psi$ at time $t$. It always holds: $success(a, \alpha, t', \Psi) \Rightarrow \exists t < t'.perform(a, \alpha, t, t', \Psi)$.*

*Example 3 (Success).* An agent in a MANET succeeds to send a packet if it is received without errors.

We are now ready to formalize the predicate $interfere(\omega, a, \alpha, t, t', \Psi)$. Recall that $\iota_\omega(t)$ returns the intensity of variable $\omega$ at time $t$. We use the notation $I_\Omega[\iota_\omega(t) := y]$ to denote the substitution of the value of $\iota_\omega$ to be $y$ at time $t$.

**Definition 3 (Interference of Variables with Actions).** *An environmental variable $\omega$* interferes *with the attempt of an agent $a$ to perform an action $\alpha$ within time-frame $[t, t']$ in an environment $\Psi$, iff the agent does not succeed but would have succeeded if $\omega$ had had a lower intensity.*

*We write $interfere(\omega, a, \alpha, t, t', \Psi)$ and have*

$$\begin{aligned} interfere(\omega, a, \alpha, t, t', \Psi) \equiv\ & perform(a, \alpha, t, t', \Psi) \wedge \neg success(a, \alpha, t', \Psi) \\ & \wedge \exists y < \iota_\omega(t), \Psi' = \langle \Omega, I_\Omega[\iota_\omega(t) := y], E_\Omega \rangle. \\ & (perform(a, \alpha, t, t', \Psi') \Rightarrow success(a, \alpha, t', \Psi')) \,. \end{aligned}$$

The first line of the formula only guarantees that the action is not successfully performed – if it was, there would be no interference. The other two lines enforce the existence of a threshold for the intensity of the interfering variable; if $\omega$ had an intensity below this threshold the action would have been successfully accomplished. That excludes the possibility that the agent is not motivated or incapable. Note that the exposure towards the interfering variable needs not to be considered, as the intensity of it can be set to zero, causing the overall influence of the variable to become zero, too (regardless of how much other variables are exposed to it). Analogously a definition could be given using only exposure.

*Example 4 (Interference).* Imagine an agent in a MANET tries to send a packet to another agent. They are separated by a wall which interferes with the transmission. If the wall was not there or maybe only thinner (i.e. the intensity-function lower), the transmission would have been successful. Another example is the collision of transmissions between two agents using the IEEE 802.11 radio technology and another agent using the Bluetooth radio technology. As they both operate on the 2.4 GHz ISM band an interference is quite possible [9]. If one of the sending agents had not sent (thus the intensity-function would have had value 0) the other transmission would not have been interfered.

The definition for *interference* given above, referred to one specific environmental variable. This can be generalized to apply for an entire environment by quantification over the environmental variables. We use the symbol $*$ as wildcard character:

**Definition 4 (Interference with Actions).** *An environment* $\Psi = \langle \Omega, *, * \rangle$ interferes *with the attempt of an agent* $a$ *to perform an action* $\alpha$ *within time-frame* $[t, t']$ *in that environment* $\Psi$*, iff there is at least one variable* $\omega$ *which interferes with the performance.*

*We write* $interfere(a, \alpha, t, t', \Psi)$ *and have*

$$interfere(a, \alpha, t, t', \Psi) \equiv \exists \omega \in \Omega . interfere(\omega, a, \alpha, t, t', \Psi).$$

### 3.2 Interference with Tasks

In the previous section a definition for *interference* of an environment with an atomic action $\alpha$ was given. Now we can proceed to the problem of interference with the accomplishment of a task, e.g. one agent asks another agent to fulfill a task for it. Such a task can consist for instance in delivering products or solving problems. Generally speaking a task can be accomplished by the sequential execution of one or more atomic actions. The interesting thing here is that every agent may have its own way to carry out the task. The environment possibly interferes with only some ways to solve the task but not with others. Note that for a rational agent following the BDI-architecture [10, 11] such a task would be a *desire* and could be accomplished by the execution of nested *plans*.

A task $\tau$ can be fulfilled through the successful execution of a sequence $\sigma_\tau$ of atomic actions. We represent such a sequence $\sigma_\tau$ as a set of triples $(t, t', \alpha)$, $0 \leqslant t \leqslant t'$, associating time-frames with atomic actions, e.g.

$$\sigma_\tau = \{(t_0, t'_0, \alpha_0), (t_1, t'_1, \alpha_1), ..., (t_n, t'_n, \alpha_n)\}.$$

The points of time given in that sequence are not absolute but relative to the point at which the execution of the sequence starts. So if one assumes the above sequence to be chronologically ordered, then $t_0 = 0$.

The successful execution of the atomic actions within their associated time-frames leads to the fulfillment of task $\tau$. As we already stated, there can be more than one way to solve a task. The set of all these sequences for one specific $\tau$ is written as $S_\tau$. Each agent $a$ has its own set of sequences $S^a_\tau$ for each task $\tau$.

The definition for *interference* of the environment with a certain sequence $\sigma_\tau$ through an agent $a$ is straightforward (based on definition 4). The only thing to take care of is that the relative dates in the sequence need to be added to the starting point of the execution.

**Definition 5 (Interference with Sequences).** *An environment $\Psi$* interferes *with the attempt of an agent $a$ to perform a sequence $\sigma_\tau$ in that environment $\Psi$ starting at $t_0$ iff it interferes with at least one atomic action of the sequence.*

*We write $interfere(a, \sigma_\tau, t_0, \Psi)$ and have*

$$interfere(a, \sigma_\tau, t_0, \Psi) \equiv \exists (t, t', \alpha) \in \sigma_\tau . interfere(a, \alpha, t_0 + t, t_0 + t', \Psi).$$

Accordingly the predicates *perform* and *success* of Def. 1 and 2 respectively can be adapted to sequences of actions. We then have $perform(a, \sigma_\tau, t, t', \Psi)$ instead of $perform(a, \alpha, t, t', \Psi)$ and $success(a, \sigma_\tau, t, \Psi)$ instead of $success(a, \alpha, t, \Psi)$.

Finally it is possible to state in which cases the environment interferes with the attempt of an agent to fulfill a task $\tau$. In order to perform a task $\tau$ within a given time-frame, an agent selects a certain sequence $\sigma_\tau$. This choice is based on observations on the environment, i.e. observations on the intensity- and exposure-functions of the respective environment $\Psi$. The choice is represented by the function $select_a(\tau, t_0, t_1, O_\Psi)$ with domain $(\mathcal{T} * \mathbb{R} * \mathbb{R} * \mathcal{O}_\Psi)$ and the set of according sequences $S^a_\tau$ as codomain; here $\mathcal{T}$ denotes the set of all tasks and $\mathcal{O}_\Psi$ the set of all possible sets of observations $O_\Psi$ on $\Psi$. Each agent $a \in \mathcal{A}$ has its own function $select_a$ because every agent may have a different strategy to select a sequence and a different set of sequences $S^a_\tau$.

**Definition 6 (Selection).** *The function $select_a(\tau, t_0, t_1, O_\Psi)$ returns the selection for a sequence of tasks $\sigma_\tau \in S^a_\tau$ an agents makes in order to accomplish a task $\tau$ in the time-frame $[t_0, t_1]$.*

Let $O^{a,t}_\Psi$ denote the set of observations on $\Psi$ to which $a$ had access to until time $t$. Then the environment *interferes* with the attempt of agent $a$ to fulfill a task $\tau$ in an environment $\Psi$ within time-frame $[t_0, t_1]$, *iff* the following holds: $\left(select_a(\tau, t_0, t_1, O^{a,t_0}_\Psi) = \sigma_\tau\right) \Rightarrow interfere(a, \sigma_\tau, t_0, \Psi)$.

### 3.3 Riskiness

Before a sophisticated agent selects a sequence of actions $\sigma_\tau$ in order to fulfill a task $\tau$, it makes a risk assessment for the possible choices. The interest of the agent may be to minimize the risk of a failure (if he wants to appear trustworthy) or to minimize its costs or both. Its risk assessment is based on a set of observations on the environment:

**Definition 7 (Riskiness).** *The function $risk(a, \sigma_\tau, t_0, O^{a,t_0}_\Psi, \Psi)$ returns the probability which an agent $a$, given observations $O^{a,t_0}_\Psi$ on $\Psi$, assigns to the event that the sequence of actions $\sigma_\tau$ will interfere with the environment $\Psi$ when executed by agent $a$ at time $t_0$.*

*Example 5 (Riskiness).* Let $\tau$ denote the task to securely transfer a packet in a MANET. The probability that this will be accomplished successfully by using a simple substitution cipher for encryption is very low; it saves computation-costs for the sending agent but the risk that it will be compromised is very high. Let's write $\sigma_\tau$ for this approach. Public key cryptography is computationally more expensive but much more secure [12]. The risk is much lower. We write $\varsigma_\tau$ for the second approach. For a sufficiently sophisticated agent $a$ we get for a general environment: $risk(a, \varsigma_\tau, t_0, O_\Psi^{a,t_0}, \Psi) \ll risk(a, \sigma_\tau, t_0, O_\Psi^{a,t_0}, \Psi)$.

## 4 Excusableness

*Excusableness* identifies cases in which bad performances of agents are wrongly used to decrease the trust in these agents; therefore only *failures* can be *excusable*. It can be argued that assessments that account for *excusableness* are more credulous because certain bad experiences are ignored. In fact, the opposite is true: by ignoring only *excusable* failures, the difference between agents that are really trustworthy and those which are not is amplified. This has consequences on the social structures of a MAS and, since Castelfranchi et al. [13] identified trust as "relational capital", agents start to care more about their image as trustee.

In the following we analyze the factors that can cause an agent to fail and identify the criteria that make a failure *excusable*. Consequently the definition of *excusableness* is given.

### 4.1 Criteria

An agent can fail in performing a task as a consequence of one or more of the following four factors:

- The *ability* of the agent,
- the possibly *interfering* environment in which the agent performs,
- the *willingness* and
- the *commitment* of the agent.

*Ability* If an agent fails to perform a task due to general inability, the trust in that agent should be diminished. If its abilities do not change, the agent will never succeed in the future; other agents should be preferred to it (the agent should be trusted less).

When the environment interferes with the performance of one agent, that does not imply that the same environment would interfere with another agent's actions. One agent might be more robust than another agent and thus should be trusted more; therefore robustness belongs as well to the abilities of an agent. That demands for a definition of *excusableness* in respect to a reference set of agents. Only if no other agent of the reference set succeeded under the same conditions, a failure might be *excusable*. This reference set can consist for instance of all agents in a MAS or only the adjacent agents in a network.

*Interference.* As listed above, three causes for a failure come from the inside of an agent, namely the ability, the willingness and the commitment of an agent. If these are the only causes for a failure, the reason for failing must be one of reluctance, inability (in respect to the environment) or an insufficient level of commitment. These causes though demand for a decrease of trust in that agent as they can clearly be attributed to the agent itself. The environmental interference is the only external causation and thus must be a precondition for *excusableness*. As defined in Def. 3, *interference* implies that the agent is not generally unable to perform the task: the agent would succeed under different circumstances. The ability of the agent itself is thereby already guaranteed if *interference* is demanded.

In a *dynamic setting*, the trustworthiness of an agent should be independent of the environment in which the agent is assessed. Otherwise, if the environment of the agent changes, the trust in that agent developed earlier ceases to be consistent with the agent's recent trustworthiness. In a *static setting* however, the environment of an agent stays always the same. In that case, the agent's environment should be attributed to the agent itself and wouldn't make failures *excusable*. Therefore our definition addresses only *dynamic settings*.

*Willingness.* A capable agent that is unwilling to perform a task cannot be trusted as it will certainly fail due to reluctance. This directly implies that a failure as a result of unwillingness is not *excusable*. *Willingness* is already implicit in the predicate of *interference* (see Sec. 3.1). With *willingness* we address the willingness of an agent to perform the task. It does not address how much the performing agent is ready to invest – the next issue:

*Commitment.* An agent can select a way, out of a set of alternative approaches, to accomplish a task $\tau$. Assume the selected approach $\sigma_\tau$ is less costly for the agent, but, to its knowledge, riskier than another approach $\varsigma_\tau$; further assume the agent fails because the approach was risky. Then, of course, the failure is not *excusable*. As Falcone and Castelfranchi [7] put it: "When x trusts y, x is just assuming that other motivations will prevail over his [y's] economic interests or other selfish goals". Therefore, given a measure for riskiness as defined in Def. 7, the definition for *excusableness* has to require that the performing agent knew no approach that was thought to be less risky and would have succeeded.

### 4.2 Definition

At this point we are equipped with all the necessary formalisms to give a formal definition of the *excusableness*-property as derived in the previous section. Note that the selection of a sequence of actions made by the failing agent is based on the observations he had access to until time $t_0$. To stay comparable, the agents from the reference set need to be equipped with the same set of observations as the failing agent. Recall that $\varepsilon_\omega$ is the function that describes the exposure of the variable $\omega$ towards other variables. Also remember that $O_\Psi^{a,t}$ is the set of observations on $\Psi$ to which $a$ had access to until time $t$.

**Definition 8 (Excusableness).** *The failure of an agent $a$ in performing a task $\tau$ in time-frame $[t_0, t_1]$ and environment $\Psi = \langle \Omega, I_\Omega, E_\Omega \rangle$ is* excusable *in respect to a group of agents $B$ iff both*

1. *no other agent in $B$ would succeed in performing $\tau$ under the same conditions and*
2. *for the sequence of actions $\sigma_\tau$ selected by agent $a$ it both holds:*
   (a) *$\sigma_\tau$ is interfering with the environment and*
   (b) *there is no alternative sequence of actions $\varsigma_\tau$ with which agent $a$ would have succeeded and which is assessed by $a$ to be less risky than $\sigma_\tau$.*

*We write $excusable(a, B, \tau, t_0, t_1, \Psi)$ with $\Psi = \langle \Omega, I_\Omega, E_\Omega \rangle$ and have*

$$
\begin{aligned}
&excusable(a, B, \tau, t_0, t_1, \Psi) \equiv \\
&\quad \forall b \in B, \Psi' = \langle \Omega, I_\Omega, E_\Omega[\varepsilon_b := \varepsilon_a] \rangle. && (1.) \\
&\qquad select_b(\tau, t_0, t_1, O_\Psi^{a,t_0}) = \sigma_\tau^b \Rightarrow \big(perform(b, \sigma_\tau^b, t_0, t_1, \Psi') \\
&\qquad\qquad \Rightarrow \neg success(b, \sigma_\tau^b, t_1, \Psi')\big) \\
&\quad \wedge select_a(\tau, t_0, t_1, O_\Psi^{a,t_0}) = \sigma_\tau^a \Rightarrow \big(interfere(a, \sigma_\tau^a, t_0, \Psi) && (2.(a)) \\
&\qquad \wedge \forall \varsigma_\tau \in S_\tau^a . (risk(a, \varsigma_\tau, t_0, O_\Psi^{a,t_0}, \Psi) < risk(a, \sigma_\tau^a, t_0, O_\Psi^{a,t_0}, \Psi) && (2.(b)) \\
&\qquad\quad \Rightarrow (perform(a, \varsigma_\tau, t_0, t_1, \Psi) \Rightarrow \neg success(a, \varsigma_\tau, t_1, \Psi)))\big).
\end{aligned}
$$

*Example 6 (Excusableness).* Let the four agents $a, b, c$ and $d$ be nodes in a MANET (as shown in Fig. 1). All of them are equipped with the same network cards and send with the same transmitting power. For the purpose of the example they are arranged in a line-topology *a-b-c-d*, i.e. agent $a$ can only communicate with $b$, $b$ also with $c$, $d$ only with $c$. Now imagine $a$ requests $b$ to deliver a packet to $c$ before time $t$. Agent $b$ sends the packet repeatedly until time $t$ but $c$ doesn't receive it as $d$ is sending at the same time: the messages sent by $b$ and $d$ collide and $c$ cannot receive any of them (this is also known as the *Hidden Terminal Problem* [14]). So $b$ fails in delivering the packet to $c$. The question is whether this failure should be used by $a$ to diminish its trust in $b$. That shouldn't be the case as it could not have been passed off better under these circumstances; this failure should be *excusable*. Let's check whether each condition required by Def. 8 is true:

1. No one else would have succeeded: Every packet sent with the same transmitting power would have collided with the packet coming from $d$ (and no other agent could send with a higher transmitting power).
2. (a) The environment did interfere: If the intensity with which the agent $d$ sent would have been 0, then $b$ would have succeeded.
   (b) There was only one possibility for $b$ to send the packet to $c$: No less risky alternatives were available to $b$ which would have succeeded.

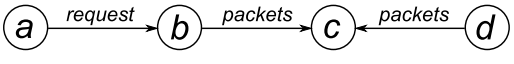


**Fig. 1.** The *Hidden Terminal Problem* in a MANET

Hence, according to the definition of *excusableness* it is indeed an *excusable* failure. In addition, all three conditions are necessary because if any of them was dropped the trust should be diminished and thus the failure should not be classified as *excusable*:

1. If another agent had succeeded, then $a$ should trust more in that agent than in $b$ concerning the fulfillment of that task.
2. (a) If the environment did not interfere, a failure would imply either $b$'s inability or its unwillingness: the trust in $b$ should be diminished.
   (b) If $b$ had sent with a higher transmitting power and would have succeeded in that case, $a$ could expect $b$ to have done so. Agent $b$'s motivation not to do so would be to save energy – but that is no excuse: agent $a$ should diminish its trust in $b$.

## 5 Related Work

The question about *excusability* is strongly related to the research field of *causality* (see e.g. Pearl [15]). In the definition of interference (Def. 3) for example, we use *counterfactuals*: if the environment had been different, the agent would have succeeded – to express this we use two different instances of an environment $\Psi$ of which one is not happening. As the definition of *excusableness* demands, one precondition is the relativity towards the abilities and commitments of other agents. This draws *excusableness* out of the field of pure *causality*.

Ramchurn et al. [2] give an overview on the research done in the field of trust in MAS. They point out that many trust models do not take into account in which context assessments of trustworthiness are made. They refer to Molm et al. [16] which state that information captured under certain conditions can become worthless under different conditions. This shows the importance of defining *excusableness* in dependence of a time-frame $[t_0, t_1]$ (at another time the failure might not be *excusable*).

In Şensoy and Yolum [5], experiences about past interactions are always associated with the environment within which they are made. This enables agents to evaluate their experiences context-sensitively. Similarly Hussain et al. [17] let their trusting peers assimilate recommendations of others based on the context and the time of the according interaction. Both approaches do not account for *excusable* failures but present frameworks that could be extended accordingly.

Falcone and Castelfranchi [7] present a comprehensive study on trust in Multi-Agent Systems. They discuss the importance of trust for MAS and analyze the elements forming a mental state of trust. They make a distinction of *external* and *internal* factors for trust. With *internal* factors they mean *ability*, *willingness* and *commitment*. With *external* they mean the environmental influence. They mention the reciprocal influence between external and internal factors. In a similar direction Boella and van der Torre [18] investigate the motivations of agents when they violate norms. They state that a sophisticated system of trust dynamics would not "decrease the trust if the other agent did its best to fulfill the promise, but failed due to circumstances beyond its control". The last two

papers [7] and [18] agree with our understanding of trust; the authors examine the impact of environmental factors for trusting agents and more general for the phenomenon trust. They do not address the problem of how experiences should be used in the process of developing trust. Furthermore they do not account for the relative aspect of an agent's *ability*.

## 6 Conclusion

In this paper we introduced the notion of *excusableness* and developed a formal definition for it. A failure shall be called *excusable* iff it should not be used to diminish the trust in the failing agent. When breaking this proposition down into a set of factors we found out that three conditions must be given such that a failure can be called *excusable*:

1. No other agent from a reference set of agents would have been able to succeed under the same conditions.
2. The environment has interfered with the performance of the failing agent.
3. No other way to accomplish the task was assessed by the failing agent to be less risky and would have succeeded.

It is concluded that environmental interference and a strong commitment of the performing agent do not yet imply *excusableness*. The relation of the failing agent towards other agents is as well essential in order to make a failure *excusable*. Further we pointed out that in dynamic settings *excusableness* has a different meaning than in static settings. In the latter case, the environment of an agent can be interpreted as part of the agent itself. Then environmental interference becomes part of the agent's inability and doesn't make a failure *excusable*.

An important question is how knowledge about *excusable* failures can be obtained in practice. As the definition of *excusableness* reveals, it is a quite complex property. Thus, in practice it will be hard to gather all the needed information to clearly state whether a failure is *excusable* or not. This fact promotes either an approach using statistical data analysis and/or a trade-off is made which however still improves the accuracy of trust-assessments.

## 7 Future Work

Currently we are working on a statistical mechanism that subtracts environmental interference in a reputation system iff it occurred for all assessed agents in the same reference groups. The approach ignores the riskiness-precondition as trade-off; this is acceptable when the reasonable assumption is made that agents do not agree on their risk-strategies over time. The assessing agents collaborate in order to collect as much information as possible. It is planned to verify the improvement within the Agent Reputation and Trust (ART) Testbed Architecture [19].

## References


1. Friedman, B., Kahn Jr., P.H., Howe, D.C.: Trust online. Commun. of the ACM, 34–40 (2000)
2. Ramchurn, S.D., Huynh, T.D., Jennings, N.R.: Trust in multi-agent systems. Knowledge Engineering Review, 1–25 (2004)
3. Huynh, T.D., Jennings, N.R., Shadbolt, N.R.: An integrated trust and reputation model for open Multi-Agent systems. Autonomous Agents and Multi-Agent Systems, 119–154 (2006)
4. Castelfranchi, C., Falcone, R.: Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In: Proceedings of ICMAS 1998, pp. 72–79 (1998)
5. Şensoy, M., Yolum, P.: A context-aware approach for service selection using ontologies. In: Proceedings of AAMAS 2006, pp. 931–938 (2006)
6. Teacy, W.T.L., Patel, J., Jennings, N.R., Luck, M.: TRAVOS: Trust and reputation in the context of inaccurate information sources. Autonomous Agents and Multi-Agent Systems, 183–198 (2006)
7. Falcone, R., Castelfranchi, C.: Social trust: a cognitive approach. Trust and deception in virtual societies, 55–90 (2001)
8. Haas, Z.J., Deng, J., Liang, B., Papadimitratos, P., Sajama, S.: Wireless ad hoc networks. In: Perkins, C.E. (ed.) Ad Hoc Networking, pp. 221–225. Addison-Wesley, Reading (2001)
9. Vilovic, I., Zovko-Cihlar, B.: Performance analysis of wireless network using bluetooth and IEEE 802.11 devices. In: Proceedings of Elmar 2004, pp. 235–240 (2004)
10. Bratman, M.: Intentions, Plans, and Practical Reason. Harvard University Press (1987)
11. Georgeff, M., Ingrand, F.: Decision-making in an embedded reasoning system. In: Proceedings of IJCAI 1989, pp. 972–978 (1989)
12. Schneier, B.: Applied cryptography, 2nd edn. Protocols, algorithms, and source code. C. John Wiley & Sons, Inc., Chichester (1995)
13. Castelfranchi, C., Falcone, R., Marzo, F.: Being trusted in a social network: Trust as relational capital. In: Stølen, K., Winsborough, W.H., Martinelli, F., Massacci, F. (eds.) iTrust 2006. LNCS, vol. 3986, pp. 19–32. Springer, Heidelberg (2006)
14. Yoo, J., Kim, C.: On the hidden terminal problem in multi-rate ad hoc wireless networks. In: Kim, C. (ed.) ICOIN 2005. LNCS, vol. 3391, pp. 479–488. Springer, Heidelberg (2005)
15. Pearl, J.: Causality: models, reasoning, and inference. Cambridge University Press, Cambridge (2000)
16. Molm, L.D., Takahashi, N., Peterson, G.: Risk and trust in social exchange: An experimental test of a classical proposition. American Journal of Sociology (2000)
17. Hussain, O.K., Chang, E., Hussain, F.K., Dillon, T.S., Soh, B.: Context and time based riskiness assessment for decision making. In: Proceedings of AICT/ICIW 2006, p. 104. IEEE Computer Society, Los Alamitos (2006)
18. Boella, G., van der Torre, L.: Normative multiagent systems and trust dynamics. In: Trusting Agents for Trusting Electronic Societies at AAMAS 2004, pp. 1–17 (2004)
19. Fullam, K., Klos, T., Muller, G., Sabater, J., Topol, Z., Barber, K.S., Rosenschein, J., Vercouter, L.: The agent reputation and trust (ART) testbed architecture. In: Workshop on Trust in Agent Societies at AAMAS 2005, pp. 50–62 (2005)