The Mechanics of Trust: A Framework for Research and Design

Jens Riegelsberger, M. Angela Sasse & John D. McCarthy
{jriegels, a.sasse, j.mccarthy}@cs.ucl.ac.uk

Department of Computer Science,
University College London,
Gower Street,
London WC1E 6BT,
United Kingdom.

Corresponding Author:
Jens Riegelsberger (Tel: +44 207 679 3643, Fax: +44 207 387 1397)

ABSTRACT

With an increasing number of technologies supporting transactions over distance and replacing traditional forms of interaction, designing for trust has become a core concern for researchers in both HCI and CMC. While much research focuses on increasing trust in mediated interactions, this paper takes a systemic view to identify the factors that support trustworthy behavior. In a second step, we analyze how the presence of these factors can be signaled to allow the formation of well-placed trust. For our analysis we draw on relevant research from sociology, economics, and psychology, as well as empirical findings in HCI and CMC research.

The key factors that warrant trust in another actor are *contextual properties* (temporal, social, and institutional embeddedness) and the *trusted actor's intrinsic properties* (ability and motivation). In first interactions, trust is mainly warranted by contextual properties, as they provide external incentives and threat of punishment. As interactions are repeated over time and trust grows, intrinsic properties become more important. To increase the level of well-placed trust, researchers and designers need to identify signals for the presence of such trust-warranting properties that are reliable and easy to interpret. At the same time, they must be cheap to emit for actors whose actions are governed by them but costly to mimic for untrustworthy actors.

Our analysis provides a frame of reference for the design of studies on trust in technology-mediated exchanges, as well as a guide for identifying trust requirements in design processes. We demonstrate application of the model in three scenarios: e-commerce, voice-enabled online gaming, and ambient technologies.

KEYWORDS
Trust, social capital, dis-embedding, interpersonal cues, human computer interaction, computer mediated communication, computer supported collaborative work, decision-making, game theory, e-commerce

1. INTRODUCTION

In the past, interactions between individuals who never met face-to-face used to be rare. Today, an ever-increasing number of first-time encounters are technologically mediated: people find business partners in online discussion forums and dating partners on *Yahoo! Personals*. In many such encounters, actors do not expect to ever meet "in real life": People buy and sell goods from each other on eBay or spend hours playing against each other on Xbox-live without ever communicating face-to-face.

These interactions involve different types and levels of risk, and they are only possible if actors trust each other, and the systems they use to meet, communicate and transact. Yet, in many recent applications, this essential quality has proved difficult to attain: The widely reported 'lack of trust' in e-commerce (Consumer Web Watch, 2002) demonstrates that insufficient trust can lead to users "staying away" from e-commerce systems altogether. Since trust is a critical factor for user acceptance and long-term success, it is not surprising that we have seen an increasing number of publications on this topic in human-computer interaction (HCI) and computer-meditated communications (CMC) research. However, many of these are focused on *increasing users' trust perceptions*, rather than *enabling correct trust decisions*. In our view, only systems that support the exchange of reliable trust cues - and thus allow for correct trust attribution - will be viable in the long run. If users experience that they cannot rely on their trust perceptions when ordering goods or taking advice via videoconferencing, trust in the technologies and application domains may be lost, or result in a system burdened with costly regulation and control structures.

There is also a less topical – but more far-reaching – argument to make trust a core concern of systems design: Any technical system that is brought into an organisation can only work efficiently as part of the larger socio-technical system – i.e. the organisation and its human actors (Checkland, 1999). Organisations are more productive if they have *social capital*[1] (Putnam, 2000). Some authors claim that reported failures of systems to yield the expected productivity gains in organisations (Landauer, 1996) partially stem from a reduction in opportunities to build social capital (Resnick, 2002). Trust can be formed as a by-product of informal exchanges, but if new technologies make many such exchanges obsolete through automation, trust might not be available when it is needed. Many studies show the economic benefits of high-trust interactions: Trust enables exchanges that could otherwise not take place, reduces the need for costly control structures, and makes social systems more adaptable (Uslaner, 2002). We find similar considerations in the field of sociology and public policy: The drop in indicators of social capital seen in modern societies in recent years has been attributed – among other factors – to the transformations of social interactions brought about by advances in communication technologies (Putnam, 2000). Interactions that used to be based on long-established personal relationships and face-to-face interaction are now conducted over distance or with automated systems – a process known as *dis-embedding* (Giddens, 1990). According to this view, by conducting more interactions over distance or with computers rather than with humans, we deprive ourselves of opportunities for trust building.

If we are to realize the potential of new technologies for enabling new forms of interactions without these undesirable consequences, trust and the conditions that affect it must become a core concern of systems development. The role of systems designers and researchers is thus not one of solely increasing the functionality and usability of the

[1] Trust that is based on shared informal norms that promote cooperation (Fukuyama, 1999).

systems that are used to transact or communicate, but to design them in such a way that they support trustworthy action and – based on that – well-placed trust. The design goal should not be to increase users' levels of trust or to help individuals appear more trustworthy, but to encourage trustworthy action and – subsequently – trust. Designers must be aware of their role as social engineers when creating on-line markets, meeting places, and environments. The design of these systems will shape how people behave – and it will impact the level of trust and cooperation.

The recent surge of trust research in HCI and CMC resulted in many empirical studies on trust that are largely focused on establishing guidelines for increasing users' perception of trustworthiness of either systems (e.g. Cassell & Bickmore, 2000; Zimmerman & Kurapati, 2002) or organisations (e.g. Egger, 2001; Nielsen, Molich, Snyder, & Farrell, 2000; Riegelsberger & Sasse, 2003a). Since trust is a term in everyday language that applies in many different situations, and that is also discussed in many different disciplines it is not surprising that this in turn has resulted in a large number of operationalisations of trust (Corritore, Kracher, & Wiedenbeck, 2003; Luhmann, 1976; McKnight & Chervany, 2000; Riegelsberger, Sasse, & McCarthy, 2003b). Several researchers have recognized the need for a more unified approach and have presented models of trust in HCI:

1. Corritore et al. (2003) developed a high-level model of trust perceptions of informational and transactional websites.
2. Tan & Thoen (2000) presented a generic model of trust in e-commerce.
3. McKnight & Chervany (2000) developed a domain-free model of trust in technically mediated interactions.

The aim of these models is to unify existing terminology and to categorize the factors that contribute to the perception of trustworthiness. In this paper, we focus not on the perception of trustworthiness, but on incentives for trustworthy behavior. We aim to identify personal and contextual properties that support trustworthy action. We then ask how the presence of these properties can be signaled via computer interfaces to build well-founded trust. This approach ensures that researchers and designers focus on trustworthy behavior, rather than just increasing trust and perceived trustworthiness. Like McKnight & Chervany (McKnight et al., 2000), we do not restrict our analysis to specific domains, but begin with an abstract, hypothetical situation and iteratively add properties and examples. In taking this approach, we aim to identify the 'mechanics of trust' that underlie many of the current online (and offline) approaches to solving problems of trust. Our goal is not to classify these existing approaches, but to illustrate why and how they work. Our analysis offers guidance to both researchers and practitioners by exposing salient features of trust in current socio-technical systems, and thus provides a basis for extrapolation to new technologies and contexts of use.

In section 2.1 we first lay the terminological and structural foundations for our framework. Subsequently we consider how trustworthiness can be signaled (2.2). Finally we introduce and illustrate a range of trust-warranting properties (2.3, 2.4). To complete the discussion of the mechanics of trust, we link our analysis to existing categories of trust (2.5). Section 3 illustrates how our framework informs research (3.1) and how it can structure design approaches in use scenarios such as e-commerce, online-gaming and ambient technologies (3.2). Finally section 4 summarizes our findings and presents design heuristics that build on the mechanics of trust we identified.

2. MECHANICS OF TRUST

In this section we lay the foundation for a framework of trust in technology-mediated interactions. We start by introducing structural conditions that define risky exchanges. Our key concern is to identify (1) the factors that allow for the formation of trust, and (2) the incentives they provide for trustworthy behavior. This knowledge can be used to identify how new technologies, which enable new types of interactions or transform existing ones, can support trust.

2.1 The Basic Model

Trust is only required in situations that are characterized by *risk* and *uncertainty*. Only if something is at stake, and only if the outcome of a situation is uncertain, do we need to trust. Uncertainty, and thus the need for trust, stems from our lack of detailed knowledge about the other actor's *abilities* and *motivation* (Deutsch, 1958). If we had accurate insight into their reasoning, trust would not be an issue (Giddens, 1990). We develop our framework from the sequential interaction between two actors (not always people): *trustor* (the trusting actor) and *trustee* (the trusted actor). Figure 1 shows a model of a prototypical trust-requiring situation.

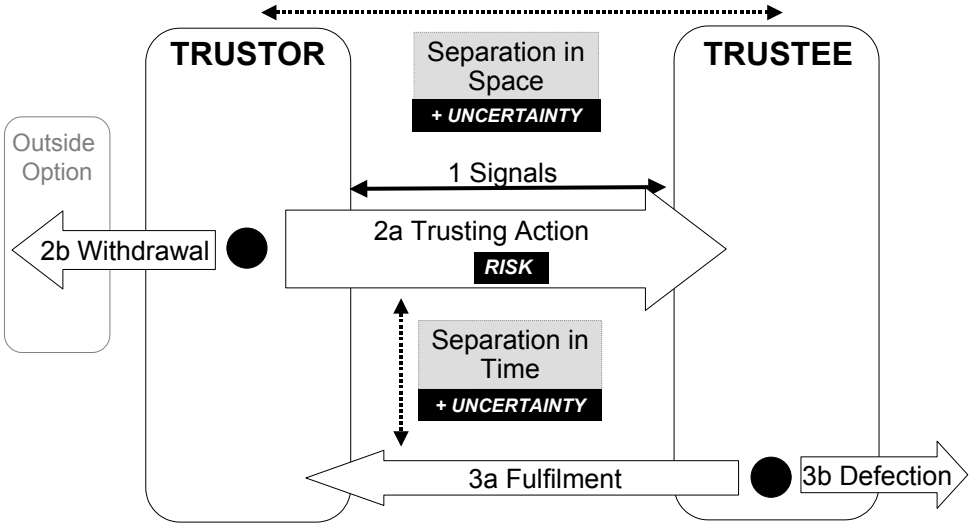

Figure 1: The basic interaction between trustor and trustee.

We have two actors about to engage in an exchange. Both can realize some gain by conducting the exchange. The exchange might involve money, but it also applies to information, time, or other goods that have value to the actors. Our exchange may depict a one-off encounter, or a 'snapshot' of an established relationship consisting of many subsequent exchanges. Prior to the exchange, trustor and trustee perceive information about each other (1). If trustor and trustee are separated in space (i.e. if their interaction is mediated by technology), less information might be available; a factor that can increase uncertainty (Giddens, 1990). The trustor has to make the first move; she can only achieve a benefit by first engaging in some form of trusting action (2a). Goods that are risked by trusting action (2a) are not only financial, but also anything with utility to the trustor: time, personal information, or psychological gratification. Even the act of trusting itself can be seen as an investment, because misplaced trust can not only lead to the loss of the invested good, but also the psychological cost of having acted naively (Lahno, 2002a).

However, the trustor has also the option to withdraw from the exchange with a given trustee (2b, e.g. by seeking another trustee or not engaging at all). The availability of outside options will thus also influence trusting action in a given situation. What makes the trustor's action risky is her dependence on the trustee's action. The trustee may fulfill (3a) his part of the exchange or not (3b, defection). In the case of fulfillment (3a), the trustor realizes a benefit. This can be financial, but her desired outcome might also be entertainment, sociability, time saved, reduced cognitive effort, or seamless collaboration (Corritore et al., 2003; McKnight et al., 2000; Rempel, Holmes, & Zanna, 1985). Defection, on the other hand, can take the form of not delivering a product, or one of lower than promised quality. In many situations, there will also be a considerable delay between trusting action (2a) and fulfillment (3a). This decreases the significance of the signals perceived prior to trusting action, as a trustee's motivation and ability can change over time. Thus, temporal separation of trusting action and fulfillment prolong the trustor's period of uncertainty and increase the need for trust (Giddens, 1990).

The commonly used definition of trust as *an attitude of positive expectation that one's vulnerabilities will not be exploited*[2] applies to the situation sketched in Figure 1.

This scenario is also captured in the game-theoretic *Trust Game* (Berg, Dickhaut, & McKabe, 2003; Dasgupta, 1988). In a Trust Game, the best outcome for the trustor is fulfillment (3a); the next-best outcome is withdrawal (2b), and then defection (3b). For the trustee the best outcome is to be trusted and defect (3b), followed by fulfillment (3a), followed by not being trusted (2b). The trustor knows that, by trusting she gives the trustee an incentive not to fulfill (e.g. keeping a customer's money but not delivering a good) and should therefore – in the absence of other motivational factors – not engage in trusting action. However, in most real-world situations with a Trust Game structure we observe trusting actions and fulfillment in spite of situational incentives to the contrary: Vendors deliver goods after receiving payment, banks return money, individuals do not sell their friends' phone numbers to direct marketers. This is because in many cases, trustees' actions will be governed by contextual and intrinsic properties whose effects outweigh the immediate gain from defection. Meyerson, Weick, & Kramer (1996) posit, that when I trust I *"...accept the possibility of ill, but usually do not expect it"* (p. 173). However, the outcome of a situation is also subject to *parametric risk* (Raub & Weesie, 2000), i.e. it depends on factors beyond the trustee's control. An important document might not arrive because it got lost in the mail, not because the trustee failed to fulfill. As we will show later, the degree to which outcomes are attributable to trustees' actions is an important determinant for trust and trustworthy behavior.

The Trust Game models an *asynchronous* and *asymmetric* exchange – the trustor and trustee act sequentially and under different situational incentives. This models many real-life trust situations: A sale, acting on advice, lending money to someone, etc. It also applies to relationships in which we speak of *mutual trust* (such as work teams or life partners). In such relationships we can identify specific exchanges, often overlapping or succeeding quickly, in which one actor is trustor and the other actor is the trustee.

However, there is one class of situations in which actors depend on each other and effectively act at the same time with the same incentives. In such a situation, every actor is simultaneously trustor and trustee and incentives are symmetrical. The classic game theory paradigm for these situations is the *Prisoner's Dilemma* (PD)[3]. Situations with a

[2] This definition closely mirrors those proposed by Corritore et al. (2003); Mayer et al. (1995); McAllister (1995); Rousseau et al. (1998).

[3] The Prisoners Dilemma was conceived by Merrill Flood and Melvin Dresher but framed in its now well-known form by Alfred Tucker (1950).

PD-structure carry an additional risk based on *strategic insecurity* (Lahno, 2002b). Strategic insecurity arises from the possibility of pre-emptive defection: My reason for not cooperating in a synchronous and symmetric situation might not be that I do not trust the other actor, but that I expect not to be trusted (Riegelsberger et al., 2003b). As much empirical research on trust and cooperation is based on the Prisoner's Dilemma rather than on the Trust Game, we will make reference to this game where its findings apply, e.g. when discussing the effects of repeated encounters (section 2.4.1). However, most real world transactions are more accurately described by the sequential model as defined in the Trust Game (c.f. Riegelsberger et al., 2003b).

The basic Trust Game already allows us to clarify a number of key concepts: A trustee who lacks the ability or motivation to fulfill is *untrustworthy*. The trustor should only make herself vulnerable if she has reason to believe that the trustee is trustworthy. Factors that influence the trustee in such a way that he favors fulfillment are called *trust-warranting properties* (Bacharach & Gambetta, 1997). Identifying and signaling them is the key design concern if we want to develop systems that foster trust and trustworthy behavior.

Actors in this framework may be individuals, organisations, or technological artifacts and for many exchanges, we have to consider the interaction on these different levels. Take the example of the interaction with a bank clerk. We can use the model to analyze the exchange with the bank clerk in the role of the trustee. However, if we 'zoom out', we can analyze the interaction of the trustor with the organisation as the trustee, or – alternatively – the trustee could be the organisations' online-banking system. As we change the focal point of the model, different trust-warranting properties come into play, and they will be signaled in different ways.

2.2 Signaling Trustworthiness: Symbols and Symptoms

When mediating exchanges, technology can perform several roles. It can (1) transmit signals prior to trusting action (e.g. on a corporate website), (2) be the channel for trusting action (e.g. by entering personal information in a web form), and (3) be used for fulfillment (e.g. by emailing ordered software).

We focus our discussion on the signaling function of technology. In the simplest form the signals from the trustee and the context allow the trustor to form expectations of behavior. As many mediated transactions are novel, the observed lack of trust can often partially be explained by a lack of experience and inappropriate assumptions about baseline probabilities (Riegelsberger & Sasse, 2001).

In first-time or one-off interactions, the signaling of trust-warranting properties is particularly important. In exchanges between actors who have interacted before, it becomes more important to signal identity, as this allows the trustor to extrapolate from historical knowledge about the trustor. To act on cues of trust-warranting properties or identity, we need to trust the channel and the channel provider to transmit them reliably and without bias. Signals are also subject to *mimicry*. Non-trustworthy actors may aim to *appear trustworthy* in order to reap the benefits. Burglars may wear couriers' uniforms, people may look you in the eyes when they lie to you, and an 'on-line bank' may be hosted from a teenager's bedroom. Mimicry will occur if the cost of emitting a signal for being trustworthy is smaller than the benefit one can expect from appearing to be trustworthy (Bacharach et al., 1997). In the view of many consumers who abstain from e-commerce, moving transactions on-line reduces the cost for mimicry (*'Anyone could make up a professionally-looking site'* (Riegelsberger et al., 2001)). Drawing on a distinction from semiotics, we can identify two broad categories of signals: symbols and symptoms (see Figure 2).

2.2.1 Symbols of trustworthiness

Symbols have an arbitrarily assigned meaning. They have been specifically created to signify the presence of trust-warranting properties. Examples of symbols for such properties are e-commerce trust seals or uniforms. Symbols can be protected by either making them very costly to fake or by sanctioning their misuse. To keep the cost of emitting symbols low for trustworthy actors, they are often protected by sanctions. Symbols are a common way of signaling trustworthiness, but their usability is limited: As they are created for specific settings, the trustor has to know about their existence and how to decode them. At the same time, trustees need to invest in emitting them and in getting them known (Bacharach et al., 1997).

2.2.2 Symptoms of trustworthiness:

Symptoms are *"information manna"* (Bacharach et al., 1997) and are not specifically created to signal trust-warranting properties; rather, they are given off as a by-product of trustworthy actions. For example, the existence of a large number of customer reviews for a product sold on an e-commerce site is seen as a symptom of a large customer base. Symptoms come at no cost to trustworthy actors, but need to be mimicked at a cost by untrustworthy ones. Trust can also be based on the absence of symptoms of untrustworthiness (e.g. nervousness, scruffy looks, a poorly kept shop or a ineptly designed web interface). Interpersonal cues (e.g. eye-gaze) are often considered to be symptomatic of emotional states and thus thought to give insight into people's trustworthiness (Goffman, 1959). However, it has been shown that we overestimate our abilities to read interpersonal cues (Horn, Olson & Karasik, 2002). These cues can be used strategically in face-to-face situations, and even more so in mediated communication that allows the sender more control over the cues he gives off (e.g. in pre-recorded video or in an avatar-representation; Garau et al., 2003). Nonetheless, we seem to be predisposed to take them into account in our trust formation (Fogg, 2003a; Reeves & Nass, 1996).

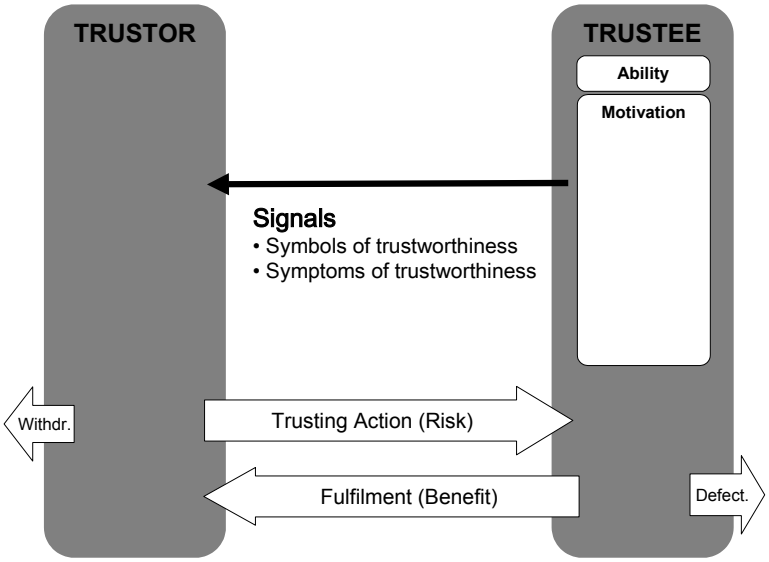

Figure 2: Ability and Motivation as foundation of trustworthiness. Trustworthiness can be signaled via symbols and symptoms.

2.3 Contextual Properties

As described in section 2.1, a trustor should trust if the trustee has the *ability* and *motivation* to act as promised. Information about ability and motivation can be inferred from signals of trust-warranting properties. Raub & Weesie (2000) identify three contextual properties that can create incentives for fulfillment (see Figure 3). They are *temporal, social,* and *institutional embeddedness*.

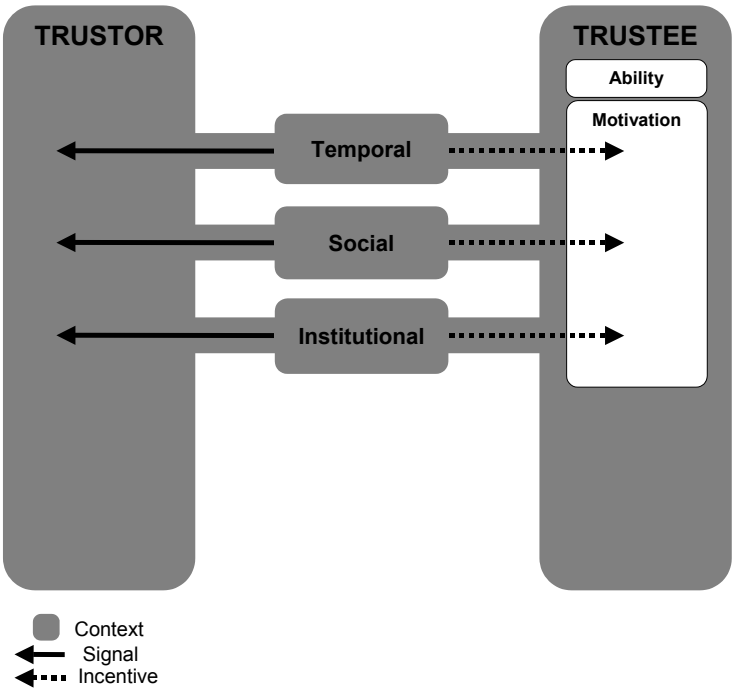

Figure 3: Contextual trust-warranting properties give the trustee incentives for fulfillment. They can also signal the presence of intrinsic trust-warranting properties (see Figure 4).

2.3.1 Temporal Embeddedness

Axelrod (1980) called the prospect of future interaction the *"shadow of the future"*, as it gives trustees an incentive to fulfill in a present encounter. This effect has been shown empirically in several studies with iterated Prisoner Dilemmas (Axelrod, 1980; Kollock, 1998; Sally, 1995) and Trust Games (e.g. Berg et al., 2003; Bohnet, Frey & Huck, 2001). If actors have stable identities and reason to believe that they will interact again, fulfillment becomes preferable for trustees. While a trustee could realize an immediate gain from defection, he also knows that in the case of defection, the trustor would not place trust in future encounters. Defection in the present encounter thus carries the cost of the gains that could be realized from future exchanges (Friedman, 1977)[4]. Additionally, if future encounters with reversed roles can be expected, defection carries the cost of

[4] In the known last round of an experimental Trust Game and a Prisoner's Dilemma the 'shadow of the future' is diminished and defection is commonly observed. By way of backward induction it could then also expect defection in the penultimate game up to the first game. However, empirically complete backward induction is not observed and cooperation commonly only erodes in the few last rounds (also called endgame effects (Poundstone, 1993)).

withdrawal from future exchanges but also the potential for retaliation. Outside the lab, a trustor's level of trust will thus – amongst other things – be determined by the likelihood of future encounters. When looking at business transactions, the trustee's demonstrated interest in future business and expected longevity in the market are thus indicators of trustworthiness. Barriers to entry and exit (e.g. in the form of high initial investments) and a shortage of other trustors are structural conditions that support trustworthiness in a given market place. In the case of interactions between work colleagues, the continuation of exchanges – often with reversed roles – is institutionally assured (see section 2.4.3). In other situations, shared group membership (e.g. in a sports club) or geographical proximity (e.g. neighborhood) is a good indicator for the likelihood of future encounters. If we share group membership with other across a range of social settings it becomes more likely that the actors will face future exchanges with reversed roles (Resnick, 2002). Repeated interactions with stable identities also allow the trustor to accumulate knowledge about the trustee and to make better predictions about his behavior. Thus, by extrapolating from past behavior trust in future encounters can grow.

System designers can capitalize on the 'shadow of the future': stable identities and the traceability of outcomes to actions are the key factors to implement for this contextual trust-warranting property. Aids for the trustor to record the outcomes of individual exchanges will further support its effect on trustworthy behavior. Signals for the applicability of this property are shared group membership, investment in initial transactions, and scarcity of trustors.

2.3.2 Social Embeddedness

Social embeddedness allows for the exchange of information about a trustee's performance among trustors. Trustees who know that trustors exchange information about their behavior have an incentive to fulfill, even if they don't expect future interaction with a given trustor. Interest in future interactions with anyone who might gain access to reputation information is an incentive for fulfillment in the present encounter (Raub et al., 2000). Reputation can thus act as a 'hostage' in the hands of socially well-embedded trustors (Raub et al., 2000), as they can threaten to tarnish the trustee's reputation if fulfillment is below expectations. A second function of reputation information is the signaling of intrinsic properties. On the assumption of the stability of intrinsic properties (e.g. benevolence, see section 2.4), reputation information can also be used to make assumptions about expected behavior in the current situation.

Reputation has been studied extensively in game theory (Dellarocas & Resnick, 2003). Beyond identifiability and traceability, reputation effects depend on (1) the social connectedness of the trustor, (2) the density of the network, (3) the cost of capturing and disseminating performance information and (4) the degree to which reputation information itself can be trusted to be truthful. The Internet or ambient technologies allow for the cheap dissemination of reputation information across a large but loosely knit network (. Reputation Systems are increasingly receiving attention in the HCI community (e.g. at the MIT Reputation Systems Symposium (Dellarocas et al., 2003)). However, trust based on reputation alone is vulnerable to strategic misuse, as inherently untrustworthy vendors can build up a good reputation to then 'cash in' and leave the market place. Finally, eliciting reputation information from trustors after transactions have been completed poses a public good dilemma in itself: trustors may have no personal benefit from sharing this information, but usually incur a cost (e.g. time spent entering feedback) from making it public. Approaches to solving this problem include showing the personal efficacy of feedback and collecting implicit feedback, by recording actor's behavior (for a detailed discussion see (McCarthy & Riegelsberger, 2004)).

In summary, reputation mechanisms – like the expectation of future encounters – require stable identities and the traceability of actions. Additionally, performance information needs to be (1) accurate, (2) easy to disseminate, and (3) easy to elicit. Finally, while reputation provides an incentive for trustworthy action independent of intrinsic trust-warranting properties, it can also be a signal for these intrinsic properties.

2.3.3 Institutional Embeddedness

Examples of institutions are law and law enforcement, but also organisations, e.g. in the form of companies or industry associations. Institutions shape behavior because they can sanction defection or insufficient fulfillment. Both trustor and trustee know that defection of an actor who operates under institutional constraints can result in litigation, punishment, or the loss of a job, etc. This mutual knowledge allows the trustor to make herself vulnerable even if she knows very little about the intrinsic properties of the trustee (Raub et al., 2000). Institutions assure fulfillment by punishing defection. However, for punishment to be a credible threat clear definitions of defection and cooperation are needed and this approach is often only viable if the cost of defection is large compared to the cost of investigation and punishment. Furthermore, with regard to contract law and enforcement it has been noted that institutional enforcement also relies on shared norms and a sense of goodwill among the parties, as a contract cannot specify all eventualities that might arise and that could thus provide loopholes (Fukuyama, 1999). We will discuss the effect of norms in section 2.5.2.

A specific form of institutional assurance is provided by organisations in the form of job roles. Individuals who want to be in continuous employment have to act in accordance with rules set out by an organisation. This is the basis of trust in many of our everyday interactions with bank clerks, sales assistants and call centre employees. We rely on our experience with these roles and trust in the institutions that define them, rather than trying to elicit information about the personal qualities of the individual performing the role. Sociologists observe that cooperation in everyday interactions is increasingly based on *institutional trust* and roles, rather than on *personal trust* based on the personal qualities of a specific individual (Giddens, 1990; Lahno, 2002a). This process vastly increases the efficiency of everyday trust decisions, as we do not have to rely on observing an individual's behavior over time, but can aggregate our trust in a corporation or a brand (see section 2.6 for the discussion of the disadvantages of this approach).

We do not normally ponder risks when we hand money over to a bank clerk. The bank clerk is identified by his position in a building, by wearing a uniform and possibly by some standard interaction pattern. Risks do not come to mind as long as these signals conform to our template of 'bank clerk interaction'. McKnight & Chervany (2000) refer to this as a perception of *situational normality*. When new technologies transform the way in which we interact, our templates of situational normality may not apply any more. This can lead to a perception of increased risk until enough successful transactions allow us to establish a new perception of situational normality (e.g. Riegelsberger et al., 2001). Additionally, if we interact through technology with trustees that are distant and based in other societies or cultures, we are less familiar with the institutions that might govern their behavior. Finally, besides structuring incentives for trustees, institutions in the form of organisations can also act as signals for trustee's intrinsic properties. Organisations select their members carefully and membership can gives cues about a trustee's qualities.

To summarize, institutions support trustworthy action by dis-incentivating non-fulfillment. To have an effect on trustees' behavior, actions must be traceable to identifiable actors and the cost of investigation and punishment must be low compared to the assured risk. Brands can bundle institutional trust in organisations.

2.4 Intrinsic Properties

While contextual properties can motivate rational self-interested trustees to fulfill (see Figure 2), they do not fully explain how actors behave empirically (Riegelsberger et al., 2003b; Fukuyama, 1999). Cooperation based on contextual properties only is bound to break down in their absence. However, actors also do fulfill in such situations based on intrinsic properties such as benevolence or ethics.

Intrinsic trust-warranting properties such as benevolence are widely believed to manifest themselves in interpersonal cues that can be easily and instantly read when interacting with others (Bacharach et al., 1997; Goffman, 1959). Face-to-face interaction is considered the broadest channel for the exchange of such cues. Examples of interpersonal cues include facial expression and body language (Bacharach et al., 1997). Numerous studies, however, found that individuals over-estimate their abilities to read such cues (O'Doherty, Kringelback, Rolls, Hornak, & Andrews, 2001). For human actors several researchers have compiled detailed overviews on character traits and personality factors that are perceived as indicators of intrinsic trustworthiness (McKnight et al., 2000). Below we introduce the core dimensions and extend, where necessary, the discussion to organisational and technological actors (see Figure 4).

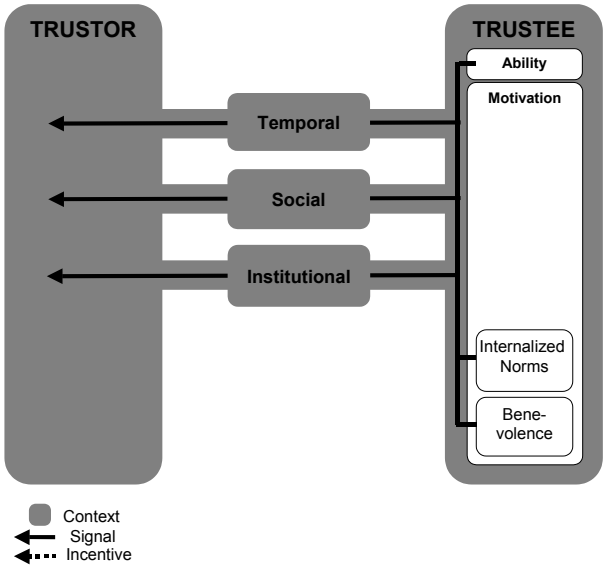

Figure 4: Intrinsic trust-warranting properties impact trustee's fulfillment independent from contextual incentives.

2.4.1 Ability

While the Trust Games or Prisoner's Dilemmas focus on willful defection, the more salient concern in everyday trust decisions will often be whether an actor is able to fulfill. Mayer et al (1995) define ability as a *"... group of skills, competencies, and characteristics that enable a party to have influence within some specific domain"*. Ability is domain-specific and the signals used to infer on ability will depend on what needs to be done to fulfill (Riegelsberger et al., 2001). A trustor can infer ability from contextual properties (previous encounters, reputation, or institutional certification), but also directly through interpersonal cues and by observing behavior in the situation. Similarly, a technical system can demonstrate its ability by allowing users to test it or by giving a record of past behavior.

2.4.2 Internalized Norms

We can observe actors fulfilling, even when they do not fear punishment, repeated interactions or tarnishing their reputation. Many people leave tips in restaurants even when they are on their own, and do not plan to ever visit again. Thus, fulfillment can be motivated by the desire to act in accordance with internalized norms or be a habitual response (Fukuyama, 1999). To account for such behavior, game theorists define different types of actors. Selfish actors are only motivated by situational pay-off, but they can be induced to trustworthy action based on contextual properties (see section 2.4). An actor of the 'intrinsically trustworthy' type, on the other hand, values cooperation higher than defection without the presence of these contextual properties. Acting in accordance with internalized norms has a utility irrespective of incentives provided by contextual properties. Human behavior is often driven by a combination of the effects of intrinsic and contextual properties (c.f. Fukuyama, 1999).

Fulfillment based on internalized norms is captured in the personality trait of *integrity: "… the trustee adheres to a set of principles that the trustor finds acceptable."* (Mayer et al., 1995). Inducing actors to internalize social norms is a difficult and lengthy process. The foundation is laid in our socialization, in which we are *'culturally embossed'* (Brosig, Ockenfels, & Weimann, 2002)[5] with the most basic norms. The socialization process also lays the foundation for habitual norm-compliance. However, social norms (such as generalized reciprocity) also differ across groups, they have to evolve over time, and triggering them may depend on the trustor's signaling of group membership (Bohnet et al., 2001; Fukuyama, 1999). Not all norms are desirable *per se*, as strong in-group reciprocity may come at the cost of hostility or aggression towards non-members (Fukuyama, 1999). Thus, when encouraging the evolution and internalization of norms researchers and designers have to be aware of the potential for such undesirable consequences.

A factor that calls into action culturally embossed norms is social presence. Visual identification (Bohnet & Frey, 1999), a photo of the interaction partner (Olson, Zheng, Bos, Olson, & Veinott, 2002), and even a synthetic voice (Davis, Farnham, & Jensen, 2002) have been shown in experimental studies to increase cooperation in Prisoner's Dilemma games. It is widely held that the reduction in social presence stemming from interaction via small bandwidth communication channels decreases norm compliance and encourages anti-social behavior, such as flaming.

The desire to act in accordance with internalized norms is an important intrinsic property that ensures trustworthy action in the absence of contextual assurance mechanisms. Designers can capitalize on it by increasing social presence, strengthening a sense of group identity, and allowing actors to exchange information about group membership.

2.4.3 Benevolence

Human behavior in romantic relationships is an example of trustworthy action motivated by strong feelings of benevolence. In such relationships the wellbeing of the other forms part of one's own gratification. Benevolence – albeit to a lesser degree – also applies to relationships between work colleagues or friends. However, strong feelings of benevolence only evolve over time and over repeated episodes of trusting and fulfilling (McAllister, 1995). In relationships of benevolence, actors do not expect immediate or equal returns.

Benevolence is also used to describe the behavior of organisational actors: Many companies aim to be good corporate citizens, they donate money to cultural events or

[5] See Fehr & Fischbacher (2003) for a discussion of the evolutionary explanations of norms of reciprocity and altruism.

humanitarian efforts; they aim to exceed customers' expectations, or to enrich their lives. As we are tuned to read signals of benevolence, such behavior and statements can increase trust, because it signals an actor's willingness to forego situational temptations and to derive gratification from the good of others. That said, actors can clearly aim to appear benevolent for strategic reasons, and much apparently benevolent behavior can be explained by the desire to be perceived as trustworthy as a means of attracting business, improving government relations, etc. Such strategic use of signals of benevolence carries the danger of destroying trust when it is perceived as manipulative (Riegelsberger & Sasse, 2002).

Benevolence can be encouraged by creating opportunities for (1) repeated interactions, and (2) to express vulnerability (e.g. through self-disclosure), liking (e.g. interpersonal cues), or good intentions (e.g. free offers, presents) prior to the trusting action.

2.5 Forms of Trust

In section 2.1 we stated that there is an abundance of trust constructs, categories and concepts. In this section we will tie some of the existing concepts of trust to the trust-warranting properties we identified. Our fundamental distinction between contextual and intrinsic properties is reflected in the discussion of other researchers. Trust based on contextual properties (e.g. institutional sanctioning) is also called *reliance* or *assurance-based trust* (Lahno, 2002a; Yamagishi & Yamagishi, 1994). Describing a similar concept, Rousseau et al. (1998) use the term *calculus-based trust*. Other terms that have been coined are *guarded trust* or *deterrence-based trust* (Lewicki & Bunker, 1996; Lewis & Weigert, 1985). *Relational trust* (Rousseau et al., 1998), on the other hand, evolves over time and is mainly based on intrinsic properties of the trustee and a history of successful exchanges. It has a higher *bandwidth* (Corritore et al., 2003; Rousseau et al., 1998) and ensures risk-taking across a wider range of situations.

Table 1: Different types of trust linked to stages in a relationship.

STAGE			AUTHOR
Early	Medium	Mature	
Deterrence-based	Knowledge-based	Identification-based	Lewicki & Bunker 1996
Calculus-based		Relational	Rousseau et al.1998
Basic/Guarded		Extended	Corritore et al. 2003
Swift			Meyerson et al. 1996

It is a common misunderstanding (Rousseau et al., 1998) that the ideal form of trust is that of identification-based or relational trust. Most exchanges are conducted on the level of calculus- or knowledge-based trust. Aiming for relational levels of trust in most exchanges would be too costly, and would tie the trust to an individual rather than to a role. It would be very time-consuming to aim for close relationships with sales assistants, bank clerks, or waiters. Similarly, it might be unnecessary or even harmful to aim for a close *parasocial relationship* (Horton & Wohl, 1956) between customers and e.g. virtual shopping agents on an e-commerce site (Fogg, 2003a).

Many researchers differentiate *cognitive* and *affective trust* (Corritore, Kracher, & Wiedenbeck, 2001; Lewis et al., 1985; McAllister, 1995; Rocco, Finholt, Hofer, &

Herbsleb, 2000) [6]. Cognitive trust is based on an inference of the trustee's incentive structures and ability. It reflects the economic understanding of trust as a rational choice. Affective trust, on the other hand, is based on immediate affective reactions, on attractiveness, aesthetics, and signs of intrinsic motivation. The amount invested in advertising, in its current form largely concerned with building affective trust, indicates the power of affective elements, at least for everyday consumer decision-making (Aaker, 1996). However, approaches that purely aim to build trust based on affective reactions without the guarantee of contextual trust-warranting properties can easily backfire: trustors might find themselves misplacing trust, and as a consequence withdraw trust from a domain in the long run. As a result, what we commonly observe is trustees signaling trust-warranting properties *and* aiming to elicit positive affective reactions (Giddens, 1990)

3. APPLICATION

In this section, we discuss how a systemic view of the mechanics underlying trust and trustworthy action (see Figure 5) reshapes the research agenda, and how it can inform the design of studies and generate hypotheses. We then apply the framework to three scenarios of technically mediated interactions to show how it can structure practitioners' approaches to designs that support trust and trustworthy action.

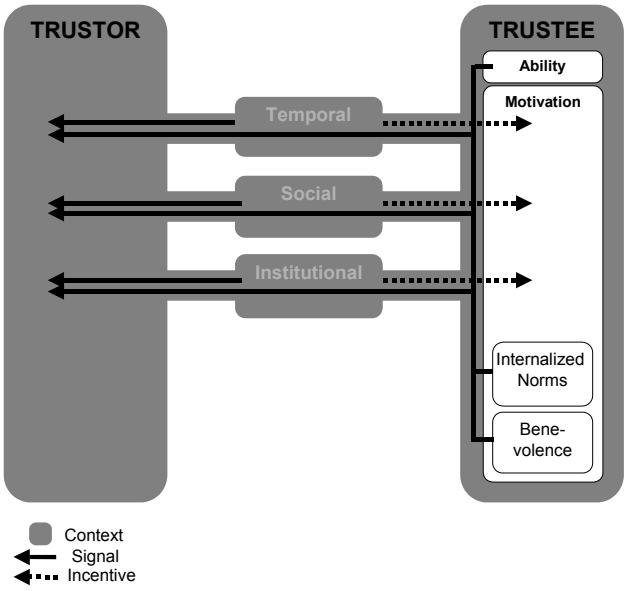

Figure 5: The complete framework

3.1 Research

The systemic framework identifies the key variables researchers have to take into account when planning studies on trust in mediated interactions. By making the trustee's incentive structure the core of our framework, we advocate the measurement of the correct trust decisions, rather than trustors' levels of trust. The framework furthermore

[6] While the distinction between affective and cognitive trust is widely used and helpful in clarifying the dynamics of trust, it is important to note that cognition and emotion are not distinct systems of decision-making. Findings in neuroscience indicate that the distinction between 'rational' and 'emotional' reasoning cannot be upheld (Damasio, 1994; O'Doherty et al., 2001; Zajonc, 1980).

allows formulating hypotheses about the impact of such incentives on trust. As an example, a study could manipulate the identifiability of a trustee and investigate how this mediates the effect of reputation information on trust and trustworthy action. Another example would be to vary parametric risk (i.e. traceability of actions) and identify whether this leads individuals to rely more on actors they have had previous experience with. Many model-based and empirical studies in economics have investigated such questions (for an overview see e.g. Dellarocas et al. (2003) or Baurmann, Matzat, & Lahno (2004)). However, these studies commonly look at the effect of one contextual property (e.g. reputation) in isolation. Furthermore, they have not been integrated with current HCI and CMC research. Thus, the framework can be seen as a step in the process of formulating a theory of trust in technically mediated transactions.

HCI research on the other hand has largely focused on trustors' perceptions, and findings have not been tied to systemic effects of trustee incentives. Research that is just focused on identifying factors that increase perceived trustworthiness will find current signals that can loose their significance once they are identified and manipulated by untrustworthy actors. Our framework takes a different approach by focusing on properties that induce trustees to act in a trustworthy manner. In the course of this analysis, we showed how known signifiers of trustworthiness relate to these contextual and intrinsic properties. By focusing research on underlying properties rather than on ephemeral signifiers, we are more likely to learn how to create environments that enable stable and well-placed trust. This approach also avoids porting signals of trustworthiness from one domain to another where they might loose their significance. While a smile in a face-to-face encounter can signal benevolence, a smiling avatar might just signal the desire to appear trustworthy.

By categorizing incentives for trustworthy action into *contextual* and *intrinsic properties*, we pinpointed a terminological ambiguity that has not been discussed widely in HCI and CMC research: The distinction between assurance and trust. It is widely claimed in the security literature (Schunter, Waidner, & Whinett, 1999), but also in HCI research (e.g. Sapient & Cheskin (1999)), that increased security, control and enforcement structures, and institutional embedding increase trust. Based on our analysis above, we argue that such approaches can increase observable trusting action by increasing one component of trust – reliance. However, reliance is largely independent from the intrinsic properties of trustee, and cooperation based on it is likely to cease when contextual properties are not in place. This insight suggests we also need to focus on intrinsic properties for trust in mediated interactions. This is particularly important, as contextual properties lend themselves more to automation and technical solutions, and thus can often seem more appropriate to HCI and CMC researchers and practitioners. The power of reputation information can be amplified through electronic networks (Dellarocas et al., 2003), digital agents can check a multitude of institutional certificates in seconds, and email allows us to trace and record past behavior. However, intrinsic properties are at least equally important (Fukuyama, 1999), even if they don't 'scale' as well as approaches based on contextual properties. Real trust relies on both properties, and the internalized norms as well as subjective trust assessments based on personal cues are essential ingredients of real trust. Thus, research needs to address how the design of digital environments can support the evolution and internalization of cooperative norms. If researchers disregard intrinsic properties, and try to ensure cooperation through external incentive mechanisms only, they run danger of replacing real trust by reliance on incentive schemes (Lahno, 2002a; O'Neill, 2002; Yamagishi et al., 1994). Such reliance can be costly and inflexible as contextual properties rely on exact a priori definitions of fulfillment and defection or on continued interactions.

3.2 Design

To illustrate the use of the framework in analyzing trust in technologically mediated transactions, we apply it to four scenarios. First we take a brief look at how technology can transform the effect of trust-warranting properties by comparing local bank branch transactions to call centre transactions. We then use the framework to discuss and question research on trust in e-commerce, which constitutes a large body of literature in current HCI. Finally, we apply it to two forward-looking scenarios: Voice-enabled online gaming and ambient technologies.

3.2.1 Branch to Call Centre: The transformation of trust through technology

An example of how the mechanics of trust have been changed for everyday transactions is the shift from interacting with a known local bank branch employee to talking to a call centre staff member. In a very small town the local bank branch manager was probably a member of the same local community as the customer, sharing the same network of friends and neighbors. Both could be expect to be members of these communities for the foreseeable future. Furthermore the bank branch manager, based on previous encounters had a good knowledge of the customer's integrity, ethics and ability to work. Based on the effect of these contextual and intrinsic properties, she (in the role of the trustor) would have been very likely to override the bank's lending rules and give the customer an unusually high short-term credit in the case of an unforeseen emergency. Conversely, the customer (in the role of the trustor) would have been very likely to give access to personal data when interacting with the branch manager.

Now imagine the local branch has been closed, but customers are given to a 24h hour call centre that processes customers' calls abroad. The only way in which the call centre employee can establish the customer's creditworthiness is by transaction history and applying baseline probabilities established by the bank's rating system. Conversely, for the customer to establish the employee's trustworthiness, he has to rely on the institutional assurance provided by the bank's job roles. It is unlikely that he will interact with a given call centre employee in the future, and he does not share any social network, if the call centre is placed in another country he does not even now under which levels of law and law enforcement the employee is operating. The bank's brand is now the main contextual trust-warranting properties. Thus, it becomes paramount to send signals for institutional embeddedness: The waiting loop plays the bank's auditory logo, and the employee gives his organisational affiliation beyond his name. In terms of intrinsic properties, again the customer has to rely on the bank's selection process for employees and on the paraverbal cues he can pick up in the voice of the employee (confidence, doubt, etc.).

This example illustrates how new technologies can transform the trust relationships among actors. Often such technologies are introduced to increase the efficiency of transactions, their trust transforming power might, however, can result in unexpected changes to trust, with often undesirable consequences. The current move of banks to give personal contact telephone numbers to customers and to re-open branches (Economist, 2004) are an indication for this. In the following application scenarios, we will illustrate how the framework can be used to identify chances for supporting trust and trustworthy action in the design phase.

3.2.2 E-Commerce

The application of the framework to B2C e-commerce helps designers to differentiate between risks that are related to the transmission technology as trustee (e.g. reliability of data transmission, risk of interception), and those that are related to the vendor's actions.

In this section we focus on the vendor as the trustee. Looking at temporal embeddedness (see Figure 6), vendors can signal trustworthiness in different ways: The first is to indicate that they are in the market for continued business and that they are interested in continued business relationship with the client. Indicators for the former can be given by visible investment in the business and the site. Many researchers on the topic identified '*professionalism'* of the website (Egger, 2001; Fogg, 2003b; Nielsen et al., 2000; Riegelsberger et al., 2003a; Shneiderman, 2000) as a core indicator of trustworthiness. This includes absence of technical failures, absence of mistakes, breadth of product palette, aesthetic design, usability, and information about physical assets. By way of extrapolation these indicators of professionalism also allow potential customers to infer the vendor's intrinsic properties, such as competence in post-order fulfillment. Vendors can show a continued interest in a specific client by investing in the first transaction. This can take the form of investment in market communications, trial-offers (e.g. Amazons first time visitor's voucher) or trial-purchases (e.g. trial bookings). Similarly, if a vendor has a high street presence or global brand, the likelihood of repeat purchases increases and so does the vendor's interest in good fulfillment. This aspect favors big, well-known players in markets that are perceived as risky.

Based on our initial interviews in the earlier days of e-commerce (Riegelsberger et al., 2001), social embeddedness, i.e. reputation, is a major factor for purchase decisions. Potential customers paid much attention to their friends' and families' recommendations and experiences with vendors. Similarly, media coverage or consumer reports were taken in account. From a customer's perspective this information was not considered an incentive for trustworthiness, but information about the vendor's intrinsic properties, such as competence or integrity. Taking a systemic view, reputation effects provide strong incentives for actors to act trustworthily and improve their quality of service. The Internet itself can be used to facilitate the formation and dissemination of reputation information about vendors: Services such as Epinions or Bizrate aim to do just that. However, they suffer from the additional cost users face when entering feedback on a different site after they received fulfillment. A more usable system would be integrated e.g. in the user's browser and could record repeat purchases implicitly (McCarthy et al., 2004), thus reducing the cost of contributing reputation information.

Indicators for physical location, such as branch or office addresses or photos are also frequently named by researchers as signals for trust (Egger, 2001; Shneiderman, 2000; Riegelsberger et al., 2001). They can demonstrate an interest in the continued presence in the market and thus a susceptibility to the effects of temporal and social embeddedness. However, the physical location can also indicate that the vendor is outside the influence of institutions (e.g. their own countries consumer protection laws) potential trustor's have trust in. An example would be information about impressive headquarters in the Cayman Islands. This example shows how a signal that is perceived as an indicator of trust can become a signifier of distrust depending on the configuration of other contextual properties. Taking a systemic view helps practitioners to anticipate such effects for their specific application scenario.

In the absence of a clear legal framework, when dealing with vendors in obscure locations, or when conditions of fulfillment are hard to specify, the possibility of legal recourse is a poor incentive for trustworthy action and thus a poor signal for trust. In such situations it should be possible to assure trust through Trust seal programs (Sapient et al., 1999). Such programs work by establishing rules of conduct (e.g. with regard to security technology or privacy policies) and checking their members' performance against these rules. Complying members are awarded 'trust seals': small icons they can display on their site. These seals are commonly linked to the certifying body's site to enable checking their veracity. The disadvantage of many seal programs is that the certifying

organisations are not well known and thus have no trust they could transfer (Riegelsberger et al., 2003a). Thus this approach by itself cannot build trust but just moves the problem of trust one step further. Trust seals given by well-known organisations that 'sublet' their trust by endorsing unknown vendors are more promising, as their own reputation is at stake. Amazon's *zShops* go beyond giving seal-based endorsements by giving space to independent vendors on their site and promising to enforce rules of conduct.

Vendor's intrinsic properties, as said above, can be inferred by site elements that indicate contextual incentives for trustworthy action. Competence or ability to fulfill can be extrapolated from professionalism in site design. Additional elements that are specifically referring to intrinsic properties include mission statements, privacy policies, upfront disclosure of terms & conditions and additional costs (e.g. shipping) to indicate a vendors' integrity; or at least to show that one is willing to set out the norms against one is willing to be judged later. Strong benevolence as it is identified in long-standing relationships between humans does not apply to e-commerce. However, with a continued business relationship a form of benevolence between vendor and customer can grow. This can take the form of strong brand identification on the side of the customer and loyalty schemes, more lenient payment conditions, or access to internal information on the side of the vendor. Policies such as 'returns with no questions asked' increase the vendor's vulnerability and can indicate that they want a benevolent relationship, in which neither partner is interested in reaping short-term benefits. As said before, these efforts can also backfire: Aiming to appear benevolent without sufficient actions to found these claims can be perceived as manipulative and thus decrease trust (Riegelsberger et al., 2002). For the extensively researched area of trust in e-commerce the framework showed how and why identified signifiers of trust do work – but also how their impact can change depending on situational factors, and thus where their limitations are. Figure 6 summarizes our discussion.

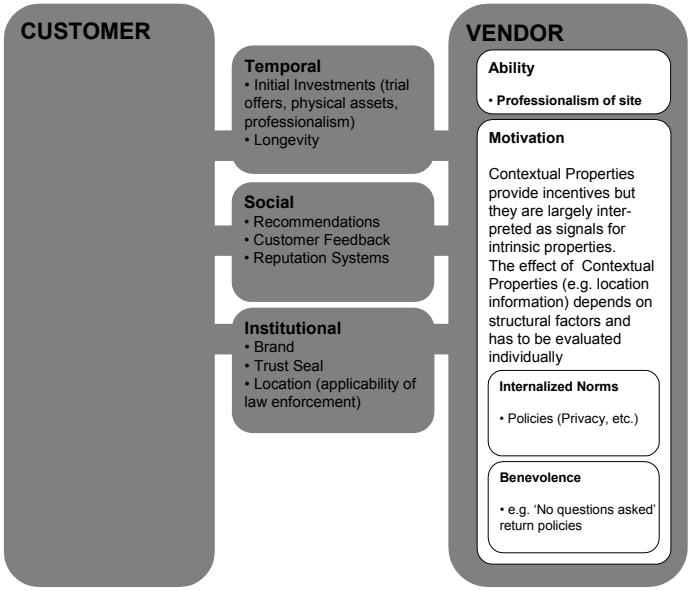

Figure 6: The framework applied to trust in consumer e-commerce.

3.2.3 Voice Enabled Gaming Consoles

An interesting application area that is currently suffering from low trust is the domain of voice-enabled online games (Hails, 2003). In many of these games, players are randomly matched to play and talk with others, about whom they know nothing. Such platforms allow encounters with a relatively high social presence, while they offer no possibility to pre-select interaction partners. High levels of reported unpleasant encounters and rude behavior are the result (Hails, 2003).

A first pass at overcoming this situation is to increase accountability in such environments by creating stable identities. This allows users to take note of who behaved in an unfair or unpleasant manner, and they can avoid future encounters with these actors (temporal embeddedness). The design of the technology can assist by providing block lists, so that players do not have to remember others' IDs. However, in a very large population of actors, loosing the chance for future encounters with a single actor is not a strong incentive for trustworthy action – there will always be someone who has not been annoyed yet. Designers of such systems could build on the effects of social embeddedness and allow users to make their block lists public. This would give every actor in the system a basic public reputation score. The effects of reputation can be improved by including positive or richer feedback that allows inferring players' intrinsic properties (e.g. playing abilities).

Institutional embeddedness can be harnessed in several ways. Users could report misbehavior to a law enforcement body that would investigate it (e.g. through recorded conversations or logs of in-game behavior) or penalize misbehaving actors (e.g. by banning for a given period of time). Obviously, such an approach is costly and open to dispute. Its applicability is limited to cases of gross misconduct. Another way of harnessing institutional embeddedness would be to allow for the formation of institutions within the gaming community. Such organisations would be allowed to formulate their own rules of conduct and to admit and dispel members based on their behavior. Such evolving organisations could then gain reputations and incentivise appropriate behavior of their members. To support this, systems must allow users to form organisations, discuss and define rules, to regulate membership and to allow them to create reliable signals for membership. A very different approach would be to focus on exposing intrinsic properties prior to in-game encounters, and to allow players to select partners based on these intrinsic properties. Profiles containing textual personal information (e.g. age, gender, education), photos, or voice samples could be the basis of such choices. This additional knowledge about other actors can also be expected to call into action norm-compliant behavior, as it shows other players as individuals that are otherwise only perceived as dis-embodied voices. Finally, by allowing users to not only keep track of people that annoyed them, but also of good gaming encounters, e.g. in the form of friends lists, benevolent relationships among players can emerge. Additional tools to increase the likelihood of re-encounters are presence indicators (e.g. in the form of 'buddy lists'), the possibility to email, or chat or to invite each other to games across platforms.

We demonstrated several ways of encouraging trustworthy behavior and trust based on the application of the framework in a scenario that is currently suffering from low-trust interactions. Current approaches focus on contextual properties in the form of e.g. institutional control or block lists (temporal embeddedness). The framework (see Figure 7) reveals that a multi-facetted approach is required that ultimately leads to benevolence or internalized norms of cooperation. In our view giving the players the tools for self-organisation and pre-selection, while providing assurance against the grossest incidents of misconduct is the most promising strategy.

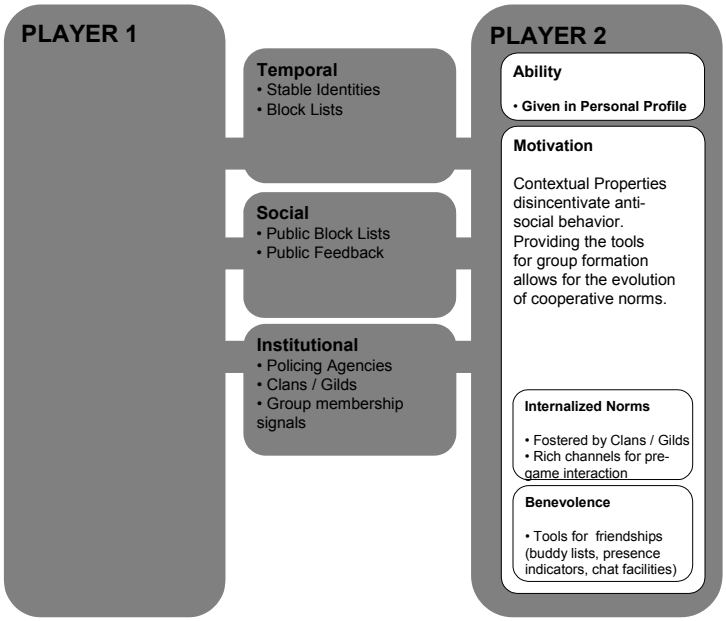

Figure 7: The framework applied to trust in online games.

3.2.4 Ambient Technologies

In this section, we apply the framework to the fictitious scenario of a PDA interacting with ambient technologies (see Figure 8). The 'ambient PDA' carries much personal information and it can share this information with the systems of other individuals and organisations. Benefits described for ambient technologies include the adaptation of physical environments to own preferences, introduction services to co-present others who share a friend or who might be a suitable dating partners, as well as customised offers in shops. Below we illustrate how the framework can be used to stimulate ideas for novel settings with a specific use scenario.

Imagine an individual in a hotel room in a foreign city working to finish an important report when the client rings and tells her that she is invited to his CEO's 50th birthday party. This is a great honor and opportunity to build a strong relationship with her client but she realizes that she does not have anything appropriate to wear. Fortunately her PDA can check local retailers for garments in her size that fit the occasion and that would go well with her wardrobe at home, taking into account preferred brands and materials. The PDA checks her schedule, organizes a bidding process among potential suppliers and books an appointment with a local fashion consultant who will bring a selection of outfits. While ambient technologies offer several benefits in such a scenario, they also pose various risks. An imposter consultant could steal from the customer in the hotel room or even harm her, he might have no knowledge of local fashion, he could use information on her price sensitivity to extract maximum prices (Shenk, 1998), or he bombard her with targeted advertising. For this example we focus on ensuring appropriate use of personal information. Fundamentally, the vision of ambient technologies relies on trading access to personal information for convenience. To be accepted ambient technologies need to allow users to retain control over how their personal information is used. However, the sensitivity of personal information cannot be classified *a priori* as it depends on the information receiver, context, costs and benefits.

These factors and others have to be assessed for every transaction (Adams & Sasse, 2001). The constant interactions among ambient systems would require users to make a myriad of such micro-trust decisions, rendering such an approach unusable. Thus, the full potential of ambient technologies can only be realized if these trade-off decisions can be delegated to an intelligent trust management system. Agents in ambient societies need to be trusted to reach correct trust assessments on the users' behalf.

Automating trust decisions based on contextual properties appears relatively straightforward. The ambient technology can record location information of actors (Terry, Mynatt, Ryall, & Leigh, 2002) and therefore establish baseline probabilities for future interactions. However, in the given scenario temporal embeddedness would not provide a strong incentive for fulfillment – the interaction between fashion consultant and customer is likely to be a one-off event. Social embeddedness in the form of reputation can also be implemented and automated trust management systems. Here it becomes important to share reputation information not in the customer's personal network, but e.g. in a locally relevant network of individual's who have stayed at the same hotel. However, if personal information is at risk, it is hard to trace violations of trust to specific actions. The customer might realize at a much later stage that her contact details had been sold on, and then it will be hard to track this defection down to the fashion consultant. Thus, in order for this approach to work, audit-trails are needed.

While institutional embeddedness in the form of law and law enforcement are likely to deter an imposter from harming the customer in the hotel room, their effect on preventing the misuse of personal information by the real consultant is less strong. It is hard to quantify the cost of small infringements of privacy. Not only will it be hard to pre-define appropriate uses of such information (what is helpful information and what is spam?), but also the costs of individual infringements are likely be too small compared to investigation and punishment. A more promising approach in the given scenario would be to ensure trust through organisational affiliation. The customer's PDA could ensure that only fashion consultants that are associated with her trusted vendor at home would be considered. The problem of trust is hereby transferred from the individual level to the organisation. This also brings temporal embeddedness into play, as future interactions with global organisations can be expected. Ambient technologies thus require unequivocal signs for organisational memberships and favor the emergence of trusted networks of service providers. These networks would guarantee a standard of conduct and information use and give the possibility for recourse.

Finally, an assessment of intrinsic personal properties (e.g. benevolence) is largely based on observation and the perception of interpersonal cues. It is hard to imagine a system that automatically interprets such cues. In our example, ambient technologies could combine contextual assurance and subjective assessment by only short-listing fashion consultants that have established credentials, and then allowing the user to make her final choice based on a short pre-recorded video or even a quick video conference chat. In such rich interactions, the customer will be able to pick up interpersonal cues and decide on whom to trust. Similarly, the very reason why the customer was invited to the CEO's party on that evening was that her client could assess her intrinsic trust-warranting properties: By exchanging small talk, by observing her behavior among others, her host can make many inferences about her integrity or benevolence. Ambient technologies can support human actors in their assessment of intrinsic properties of others by keeping track of interaction partners and by recording and managing the subjective impressions of them.

The vision of the ambient technologies relies on the constant exchange of personal information among multiple devices, individuals and organisations. The high number of micro-trust decisions required in such a scenario calls for automated trust management. Trust decisions based on contextual properties can be delegated to ambient systems, e.g. in the form of reputation tracking, certification and organisational affiliation (see Figure 8). These contextual properties allow intelligent agents to check appropriate levels of incentivation for fulfillment and give signals for intrinsic properties. However, human trust is also vested in intrinsic properties such as integrity or benevolence that we can quickly assess in the form of interpersonal cues. Ambient technologies cannot replace this subjective trust-assessment, but should help support us in performing it.

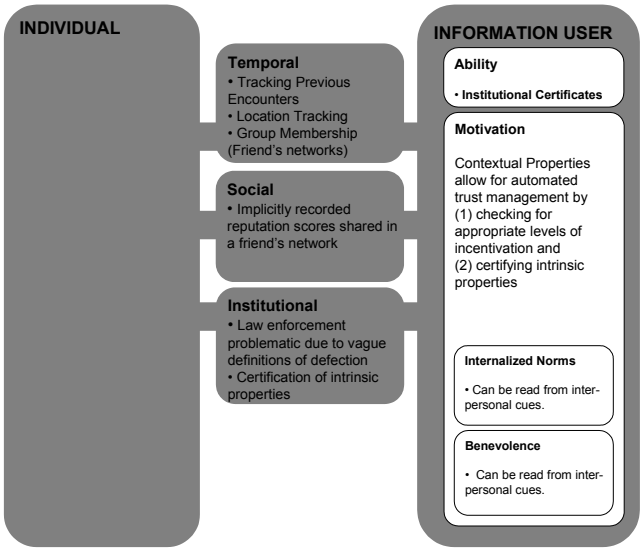

Figure 8: The framework applied to ambient technologies.

4. SUMMARY AND CONCLUSIONS

Trust is an integral part of human interactions. It allows us to engage in exchanges that leave both parties better off; it reduces the cost of these transactions; and on a societal level, trust correlates positively with productivity, low crime and health (Resnick, 2002). New technologies allow interactions between individuals who know little about each other prior to the encounter. Exchanges that have traditionally been conducted face-to-face are now mediated by technology or even executed with technology as a transaction partner. These changes put the responsibility for supporting trust and trustworthy action on the designers of the technical systems involved.

When transactions are mediated, risks, benefits and trust-warranting properties are transformed. Most importantly, however, signals of trust-warranting properties and identity can lose some of their significance once they are mediated. Conversely, trustors might not know about the signals and their significance in this new environment. Not surprisingly, this situation has led to a surge of research on trust in HCI and CMC. However, most of this research has focused on signals of trustworthiness, such as interface cues in e-commerce, and not considered risky exchanges on a systemic level. Drawing on relevant knowledge from other disciplines, we have put forward a systemic view of risky interactions. We wanted to expose the 'mechanics of trust', i.e. show how contextual and intrinsic properties influence the incentives of the trusted party – and how their presence can be signaled. This analysis advocates focusing research on the

correctness of trust-decisions, rather than on raising levels of trust. It further aids researchers in interpreting results from studies within a larger frame of reference. For designers, the benefit of the model is that it provides domain-independent design guidelines for incentives for trustworthy actions and their signals.

The model illustrates trust in one-off and continued interactions. For continued interactions, it allows for an analysis of trust at different stages of the relationship. It describes trust-warranting properties – contextual and intrinsic – that ensure trustworthy behavior. Contextual properties (temporal, social, and institutional embeddedness) will be of higher importance in first interactions and one-off encounters. Intrinsic properties of the trustee (ability, norm-compliance, and benevolence), on the other hand, are more important in continued exchanges and become increasingly relevant as trust matures. Signals of trust-warranting properties are the basis for trust. They are subject to mimicry and their significance can be lost when they are mediated. We identified two types of signals: Symbols and symptoms. Symptoms are signals of trustworthiness that are given as by-product of behavior. They are preferable to symbols, which may be costly to emit, are less reliable, and subject to mimicry.

To summarize the insight gained from our analysis of the mechanics underlying trust in risky exchanges, we present design heuristics that can be expected to support trustworthy action and personal and assurance-based trust (Table 2).

Table 2: Design heuristics for trust-supporting systems.

Heuristic	Description	Relevant Property
Stable Identity	Stable identities are a key requirement for contextual properties to take an effect (threat of punishment, threat of avoidance in repeated encounters, threat of avoidance based on poor reputation). They are also important for trust to grow based on intrinsic properties in repeated encounters	Contextual and intrinsic properties
Traceability Accountability	Contextual properties can only have an incentivating function if parametric risk is low, i.e. if outcomes can be traced to actions	Temporal, Social, Institutional Embeddedness
Group Membership Group Identity	Shared group membership and group size give indication for the likelihood of re-encounters and thus determine the strength of the contextual property temporal embeddedness. Furthermore, shared group membership and a strong group identity support the emergence of group-specific norms of cooperation or generalized reciprocity. Group membership can also be an indicator for competence and abilities and for the rules by which a trustee will be bound	Future interaction Norm compliance Ability Institutional Embeddedness
Social Presence	Several studies have shown that social presence calls norm compliant behavior into action. (Furthermore, through rich channels liking and thus benevolence can be signaled)	Internalized Norms Benevolence
Recording outcomes	Only if trustors have a possibility to record the outcome of interactions, temporal and social embeddedness can become incentives for trustworthy action.	Temporal, Social Embeddedness

While this list will be helpful in ensuring the presence of the basic requirements for trust, it should not lead researchers and designers to assume that trust and trustworthy action can be 'designed into a system'. We are convinced that a system also needs to give room for the evolution of groups and informal institutions. These can develop codes of conduct, signals for membership, and eventually can lead to the formation of intrinsic preferences and internalized norms. A well known example of such an approach are role

playing games that allow the formation of gilds with rules and membership signs. Thus, by ensuring the presence of the fundamental cornerstones of trust and by leaving room for evolution, a culture of trust can emerge over time.

We are convinced that what we sometimes see as a 'lack of trust' is not an unavoidable consequence of technology-mediated exchanges. Rather, these examples are symptoms of difficulties in adapting traditional ways of trust signaling and formation to new structural conditions. This should, however, not be interpreted as "in time, the trust problem will go away because people will adapt" – negative trust experiences can cause long-term damage to the technologies and/or application domains involved. The damage will not only be to commercial companies – technology providers and organisations offering innovative services – but may also deprive individuals and society of the benefits the technology or service could offer. Such developments would disenfranchise those who can least afford to take financial risk, and/or lack in-depth knowledge and technical savvy to distinguish trustworthy actors from untrustworthy ones. Researchers and designers must support the transformation process by providing empirical evidence on the formation of trust via technical channels and by creating well-informed designs. Thus, rather than seeing them as a threat to cooperation and social capital, we welcome the potential of these technologies to enable exchanges with others we would otherwise never have interacted with.

5. ACKNOWLEDGEMENTS

We would like to thank Richard Boardman, Ivan Flechais, and Hendrik Knoche for their comments on earlier versions of this paper.

6. REFERENCES

Aaker,D.A. (1996). Building Strong Brands. New York: The Free Press.

Adams,A.,& Sasse,M.A. (2001). Privacy in Multimedia Communications: Protecting Users, Not Just Data. Proceedings of HCI2001, Lille.

Axelrod,R. (1980). More Effective Choice in the Prisoner's Dilemma. Journal of Conflict Relosution, 24(3), 379-403.

Bacharach,M., & Gambetta,D. Trust in Signs. (1997). University of Oxford.

Baurmann,M., Matzat,U., & Lahno,B. Trust and Community on the Internet: Opportunities and Restrictions for Online Cooperation, Bielefeld.

Berg,J., Dickhaut,J., & McKabe,K. (2003). Trust, Reciprocity, and Social History. Games and Economic Behaviour, 10, 122-142.

Bohnet,I., & Frey,B.S. (1999). The sound of silence in prisoner's dilemma and dictator games. Journal of Personality and Social Psychology, 38, 43-57.

Bohnet,I., Frey,B.S., & Huck,S. (2001). More order with less law: On contract enforcement, trust and crowding. American Political Science Review, 95, 131-144.

Brehm,S.S., & Kassin,S.M. (1996). Social Psychology. (3rd Edition ed.). Boston: Houghton Mifflin.

Brosig,J., Ockenfels,A., & Weimann,J. (2002). The effects of communication media on cooperation. German Economic Review, 3(3).

Cassell,J., & Bickmore,T. (2000). External Manifestations of Trustworthiness in the Interface. Communications of the ACM, 43(12), 50-56.

Checkland,P. (1999). Soft systems methodology. a 30-year retrospective. Chichester: John Wiley.

Consumer Web Watch. A Matter of Trust: What User Want From Web Sites. http://www.consumerwebwatch.org/news/report1.pdf . 2002. Consumer Web Watch.

Corritore,C.L., Kracher,B., & Wiedenbeck,S. (2001). Trust in the online environment. In M. J. Smith, G. Salvendy, D. Harris, & R. J. Koubek (Eds.), Evaluation and Interface

Design: Cognitive Engineering, Intelligent Agents and Virtual Reality. (pp. 1548-1552). Mahwah, NJ: Lawrence Erlbaum.
Corritore, C. L., Kracher, B., & Wiedenbeck, S. On-line trust: concepts, evolving themes, a model. International Journal of Human Computer Studies 58(6), 737-758. 2003.
Damasio,A.R. (1994). Descarte's Error: Emotion, Reason and the Human Brain. New York: Avon.
Dasgupta,P. (1988). Trust as a Commodity. In D. Gambetta (Ed.), Trust. Making and Breaking Cooperative Relations. (pp. 49-71). Oxford: Basil Blackwell.
Davis,J.P., Farnham,S.D., & Jensen,C. Decreasing online 'bad' behavior. In Anonymous. Extended Abstracts CHI 2002: New York: ACM Press.
Dellarocas,C.,& Resnick,P. Reputation Systems Symposium. 1st Interdisciplinary Symposium on Reputation Systems.http://www.si.umich.edu/~presnick/reputation/symposium/agenda.htm
Deutsch,M. (1958). Trust and suspicion. Journal of Conflict Resolution, 2(3), 265-279.
Economist. (2004, February 19). Branching Out. Economist.
Egger,F.N. (2001). Affective Design of E-Commerce User Interfaces: How to maximise perceived trustworthiness. In M. Helander, H. M. Khalid, & Tham (Eds.), Proceedings of CAHD: Conference on Affective Human Factors Design: Singapore.
Fehr, E. and Fischbacher, U. (2003), The Nature of Human Altruism, Nature, no. 425, pp. 785-791.

Fogg,B.J. (2003a). Persuasive Technology. using Computers to Change What We Think and Do. San Francisco: Morgan Kaufmann.
Fogg,B.J. Prominence-Interpretation Theory: Explaining How People Assess Credibility Online. CHI2003 Extended Abstracts. New York: ACM Press.
Friedman,J.W. (1977). Oligopoly and the Theory of Games. Amsterdam: North-Holland Publishers.
Fukuyama,F. Social Capital and the Civil Society. In Anonymous. 2nd Conference on Second Generation Reforms: Washington, DC: IMF.
Garau,M., Slater,M., Vinayagamoorhty,V., Brogni,A., Steed,A., & Sasse,M.A. The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual Environment. In Anonymous. Proceedings of CHI 2003: New York: ACM Press.
Goffman, E. (1959), *The Presentation of Self in Everyday Life.* Garden City: Doubleday.
Giddens,A. (1990). The consequences of modernity. Stanford: Stanford University Press.
Hails, M. Gamertag: Gecko Dance. A Newbie's Diary. All Xbox GameWatcher: Live Wire . 2003. www.allxbox.com/news/sotires/livewire122602b.asp
Horn,D.B., Olson,J.S., & Karasik,L. (2002, April 20-2002, April 25). The Effects of Spatial and Temporal Video Distortion on Lie Detection Performance.CHI 2002 Extended Abstracts. New York: ACM Press.
Horton,D., & Wohl,R.R. (1956). Mass communication and para-social interaction: Observations on intimacy at a distance. Psychiatry, 19, 215-229.
Kollock,P. (1998). Social Dilemmas: The Anatomy of Cooperation. Annual Review of Sociology, 24, 183-214.
Lahno,B. (2002a). Institutional Trust: A Less Demanding Form of Trust? Revista Latinoamericana de Estudios Avanzados (RELEA),
Lahno,B. (2002b). Vertrauen. Paderborn: Mentis.
Landauer,T.K. (1996). The Trouble with Computers: Usefulness, Usability, and Productivity. Cambridge, MA: MIT Press.

Lewicki,R.J., & Bunker,B.B. (1996). Developing and Maintaining Trust in Work Relationships. In R. M. Kramer & T. R. Tyler (Eds.), Trust in Organizations. Frontiers of Theory and Research. Thousand Oaks, CA: Sage.

Lewis,J.D., & Weigert,A. (1985). Trust as a Social Reality. Social Forces, 63, 967-985.

Luhmann, N. (1979) Trust and Power Cichester: Wiley.

Mayer,R.C., Davis,J.H., & Schoorman,F.D. (1995). An Integrative Model of Organizational Trust. Academy of Management Review, 20(3), 709-734.

McAllister,D.J. (1995). Affect- and Cognitation-based Trust as Foundations for Interpersonal Cooperation in Organizations. Academy of Management Journal, 38(1), 24-59.

McCarthy,J.D.,& Riegelsberger,J. (in press) The Designer's Dilemma: Approaches to the Free Rider Problem in Knowledge Sharing Systems. In Anonymous. COOP 2004, in Proceedings of 6th International Conference on the Design of Cooperative Systems, May 2004, French Riviera, France:

McKnight,D.H.,& Chervany,N.L. (2000) What is Trust? A Conceptual Analysis and An Interdisiplinary Model. American Conference on Information Systems:

Meyerson,D., Weick,K.E., & Kramer,R.M. (1996). Swift Trust and Temporary Groups. In R. M. Kramer & T. M. Tyler (Eds.), Trust in Organizations. Frontiers of Theory and Research. (pp. 166-195). Thousand Oaks, CA: Sage.

Nielsen,J., Molich,R., Snyder,S., & Farrell,C. (2000) E-Commerce User Experience: Trust. Fremont, CA: Nielsen Norman Group.

O'Doherty,J., Kringelback,M.L., Rolls,E.T., Hornak,J., & Andrews,C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. Nature Neuroscience, 4(1), 95-102.

O'Neill,O. (2002). A Question of Trust. The Reith Lectures 2002. London: BBC Channel 4.

Olson,J.S., Zheng,J., Bos,N., Olson,G.M., & Veinott,E. (2002) Trust without Touch: Jumpstarting long-distance trust with initial social activities. Proceedings of CHI 2002. New York: ACM Press.

Poundstone,W. (1993). Prisoner's Dilemma. (2nd ed.). New York: Anchor Books.

Putnam,R.D. (2000). Bowling Alone: The Collapse and Revival of American Community. New York: Simon & Schuster.

Raub,W.,& Weesie,J. (2000) The Management of Durable Relations. In Anonymous. The Management of Durable Relations: Amsterdam: Thela Thesis.

Reeves,B., & Nass,C. (1996). The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Stanford: CSLI Publications.

Rempel,J.K., Holmes,J.G., & Zanna,M.P. (1985). Trust in close relationships. Journal of Personality and Social Psychology, 49(1), 95-112.

Resnick,P. (2002). Beyond Bowling Together: SocioTechnical Capital. In J. M. Carroll (Ed.), HCI in the New Millenium . (pp. 247-272). Boston, MA: Addison-Wesley.

Riegelsberger,J.,& Sasse,M.A. (2001). Trustbuilders and trustbusters: The role of trust cues in interfaces to e-commerce applications. In B. Schmid, K. Stanoevska-Slabeva, & V. Tschammer (Eds.), Towards the E-Society: E-commerce, E-Business and E-Government: Norwell: Kluwer.

Riegelsberger,J.,& Sasse,M.A. (2002). Face it: Photographs Don't Make Websites Trustworthy. In Anonymous. CHI2002: Extended Abstracts. New York: ACM Press.

Riegelsberger,J., & Sasse,M.A. (2003a). Designing E-Commerce Applications for Consumer Trust. In O. Petrovic, M. Ksela, & M. Fallenboeck (Eds.), Trust in the Network Economy. (pp. 97-110). Wien: Springer.

Riegelsberger,J., Sasse,M.A., & McCarthy,J. (2003b). The Researcher's Dilemma: Evaluating Trust in Computer-Mediated Communication. International Journal of Human Computer Studies, 58(6), 759-781.

Rocco,E., Finholt,T.A., Hofer,E.C., & Herbsleb,J.D. (2000) Designing as if trust mattered. (2000). Collaboratory for Research on Electronic Work (CREW) Technical Report

Rousseau,D.M., Sitkin,S.B., Burt,R.S., & Camerer,C. (1998). Not so different after all: A cross-discipline view of trust. Academy of Management Review, 23(3), 393-404.

Sally,D. (1995). Conversation and Cooperation in Social Dilemmas. A Meta-Analysis of Experiments from 1958 to 1992. Rationality and Society, 7(1), 58-92.

Sapient & Cheskin (1999). eCommerce Trust. http://www.cheskin.com/think/studies/ecomtrust.html

Schunter,M., Waidner,M., & Whinett,D. The SEMPER Framework for Secure Electronic Commerce. In Anonymous. 4. Internationale Tagung Wirtschaftsinformatik: Heidelberg: Physica Verlag.

Shenk,D. (1998). Data Smog: Surviving the Information Glut. San Francisco: Harper.

Shneiderman,B. (2000). Designing trust into online experiences. Communications of the ACM, 43, 57-59.

Tan,Y., & Thoen,W. (2000). Toward a Generic Model of Trust for Electronic Commerce. International Journal of Electronics Commerce, 5, 61-74.

Terry,M., Mynatt,E.D., Ryall,K., & Leigh,D. (2002) Social net: using patterns of physical proximity over time to infer shared interests. CHI 2002: Extended Abstracts, Minneapolis, MN: New York: ACM Press.

Tucker,A. A two-person dilemma. (1950). Stanford, CA: Stanford University Press.

Uslaner,E.M. (2002). The Moral Foundations of Trust. Cambridge: Cambridge University Press.

Yamagishi,T., & Yamagishi,M. (1994). Trust and commitment in the United States and Japan. Motivation and Emotion, 18, 129-166.

Zajonc,R.B. (1980). Feeling and Thinking. Preferences Need no Inferences. American Psychologist, 35, 151-175.

Zimmerman,J.,& Kurapati,K. Exposing profiles to build trust in a recommender. CHI 2002 Extended Abstracts: New York: ACM.