Examining Trust, Forgiveness and Regret as Computational Concepts
Marsh, Stephen; Briggs, P.

*Examining Trust, Forgiveness and Regret as Computational Concepts **

Marsh, S., Briggs, P.
2008

* published in Computing with Social Trust (Book, Springer, ed. J. Golbeck). 2008. NRC 49906.

Examining Trust, Forgiveness and Regret as Computational Concepts

Stephen Marsh[1] and Pamela Briggs[2]

Abstract The study of trust has advanced tremendously in recent years, to the extent that the goal of a more unified formalisation of the concept is becoming feasible. To that end, we have begun to examine the closely related concepts of regret and forgiveness and their relationship to trust and its siblings. The resultant formalisation allows computational tractability in, for instance artificial agents. Moreover, regret and forgiveness, when allied to trust, are very powerful tools in the Ambient Intelligence (AmI) security area, especially where Human Computer Interaction and concrete human understanding are key. This paper introduces the concepts of regret and forgiveness, exploring them from social psychological as well as a computational viewpoint, and presents an extension to Marsh's original trust formalisation that takes them into account. It discusses and explores work in the AmI environment, and further potential applications.

Keywords: Computational Trust, Forgiveness, Regret, Ambient Intelligence, Security.

1 Introduction

Agents, whether human or artificial, have to make decisions about a myriad of different things in often difficult circumstances: whether or not to accept help, or from whom; whether or not to give a loan to a friend; which contractor to choose to install a new kitchen; whether to buy from this online vendor, and how much money to risk in doing so, and so on. As has been pointed out [15], invariably trust is a component of these decisions. Trust is a central starting point for decisions based on risk [61], and that's pretty much all decisions involving putting ourselves or our resources in the hands (or whatever) of someone, or something, else.

National Research Council Canada, Institute for Information Technology. e-mail: steve.marsh@nrc-cnrc.gc.ca · Northumbria University, School of Psychology and Sport Sciences. e-mail: p.briggs@unn.ac.uk

We can see trust as exhibiting something of a duality – in many decisions one trusts or does not, whereas in others one can trust *this much*, and no more, and decide where to go from there. To make the study of trust somewhat more interesting, as well as this duality, trust exhibits a behaviour that is singular in that it is seen only in its consequences – that is, I may say ‘I trust you’ but until I do something that *shows* that I trust you, the words mean nothing. Like light, trust is seen in its effect on something, and inbetween truster and trustee there is simply nothing to see.

Trust provides a particularly useful tool for the decision maker in the ‘shadow of doubt’ [51], and indeed ‘distrust‘ provides useful analysis and decision making tools in its own right (cf [71, 30, 72, 68]. That said, the dynamics of trust seem far from simple, in that trust is itself influenced by, and influences, many other social phenomena, some of which are the subjects of study in their own right, including, for instance, morality, apologies, and ethics [18, 92, 21, 44, 77, 8, 5]. The recent interest in trust (brought on mainly from the eCommerce field, and the rapid convergence of technologies and ‘social’ connections of people across distance) has resulted in a much greater understanding of trust as a single object of study. We believe that the time is right to expand that study into the counterparts of trust.

We are working towards an over-arching theory and model of artificial trust-based behaviour for a specific culture[1] in a human-populated social setting. In such a setting, it is important to be able to reason with and about the social norms in operation at that time. As such, while there exist many promising angles of study, for us the logical next step is to attempt to incorporate into our understanding of trust a deeper understanding of what it can lead to, and what happens next.

What trusting (or failing to trust) can lead to, amongst other things, is *regret*: regret over what was, or might have been, what was or was not done, and so on. What comes next may well be *forgiveness*: for wrongs done to us, or others, by ourselves or others. We conjecture that a deeper understanding of, and at least pseudo-formal model of, trust, regret and forgiveness and how they may be linked is a necessary step toward our goal. The reasons why regret and forgiveness are worth studying will be discussed further below, but that they have both been the objects of study of philosophers and psychologists for many years (or centuries) [2, 16, 7, 57, 26, 50, 27, 89, 37], and moving into the present day they provide ever more compelling topics for discussion and examination, for instance in areas where humans work and play virtually (cf [88]).

This chapter shows the enhancement of a model of trust that was first introduced in 1992 [63, 66, 64]. At that time, the model incorporated a great deal of input from the social sciences and philosophies, and although it was aimed at being a computationally tractable model, and indeed was implemented, it had its difficulties, as many observers have since pointed out. This work marks the first major enhancement of the model, and incorporates previously published work on the ‘darker’ side of trust [68] as well as recent forays into regret management and punishment [26]. Accordingly, the first part of the chapter delves into the whys and wherefores of trust itself,

[1] When one comes from a largely western-oriented judaeo-christian culture, one tends to look in that direction. However, trust and its application across cultures is not an ignored topic at all [46, 56, 75, 14]

and why it's worth studying, before setting the stage in section 3 with a parable for the modern age that brings regret and forgiveness into the frame of Ambient Intelligence and Information Sharing. Following diversionary comments related to human factors, which explain to an extent why we followed the less formal path in the first place, the original model is briefly presented in section 4.

Section 6 presents a discussion of the dark side of trust: distrust and its siblings, untrust and mistrust. They must be incorporated into the model before we can proceed further because ultimately regret and forgiveness and betrayal of trust will lead us into areas where trust is more murky. In section 7, we present a thorough examination of regret, its roots and uses, and incorporate it into the model in section 8, while sections 9 and 10 accomplish much the same for forgiveness. A worked example of the phenomena in practice is given in section 11, along with a brief discussion of some of the current work in which we are applying the concepts. Conceding that this is not the final word on trust, we discuss related work and future work in trust as is may be in sections 12 and 13 before concluding.

Caveats

This chapter is not going to state how trust works. What is, to some extent, being looked at in this work is not actually *trust* at all, as we see it in everyday life, but a derivative of it. In much the same way that Artificial Intelligence is not *real* intelligence, the computational concept of trust isn't *really* trust at all, and if the reader wants other views of real trust, there are many excellent tomes out there (for instance, [74, 73, 87, 24, 15, 33], to name a few). For some time, we felt that the artificial concept needed a name that was different enough from the original to remove the need for preconceptions, and introduce the opportunity for ingenuity (and let's be fair, some cost-cutting in definitions), and a foray into other names resulted in the *Boing* concept [67] as an attempt to do just that. Still, there is something to be gained from using a term close to the original from a human factors point of view, as is discussed later in this chapter. That given, at this time, we remain with the term 'trust'.

Some of this chapter may seem overly philosophical, or even quasi-religious, for a book on Computing with Trust. We make no apology for this — the foundations of trust go a long way back into many spheres, and when one considers the concept of forgiveness in particular, there is much to learn from a great many of the world's religions. If our ultimate aim is the attainment of socially viable 'automaton' (call it an agent) of sorts, extant in the worlds, both physical and artificial. where humans also exist, a solid interpretation of human social norms is necessary. Trust, Forgiveness and Regret are merely steps along this way, but a strong model incorporating all three is necessary, and this is what we are attempting here. This is, of course, not to say that Trust, Regret, and Forgiveness are the most *important* or even *timely* objects of study. In a world where opportunities for 'betrayal' of trust keep multiplying and the precious shadow of the future keeps shrinking [15, page 3], thinking about regret and forgiveness is at least moving in a right direction.

2 Why is Trust Important? Why a Formalization?

> … trust is a social good to be protected just as much as the air we breathe or the water we drink. When it is damaged, the community as a whole suffers; and when it is destroyed, societies falter and collapse.
>
> Bok, 1978, pp 26 and 27.

Trust is so all pervasive in all of our lives, online or not, that it sometimes seems strange to either have to ask or answer the question of why trust is important. Despite the need for more control, especially with regard to technology [15] trust remains a paramount part of our daily lives. This is especially true when other people are involved, thus, when making decisions about using a babysitter to buying a house, trust, and confidence, are elements in the decision, even if sometimes managed more by regulation. An absence of trust in a society seems to result in the death of that society, however small (cf [9, 53].)

For an artificial entity extant in a social world where humans are present, this is a fact: humans must be considered in terms of trust, and while the considerations are many, they include:

- How much they might trust an entity or what the entity gives them (for instance, information);
- How much they might trust each other in a social network;
- How much they can be trusted in a given situation, either by the entity or by someone else working with them;
- What can be done to augment the trust in order to achieve a higher level of confidence;

The key thing to note here is that allowing technology, the artificial entity, to consider trust, amongst the other factors at its disposal in decision making, can be nothing other than positive. As Gambetta states, ‘if behaviour spreads through learning and imitation, then sustained distrust can only lead to further distrust. Trust, even if always misplaced, can never do worse than that, and the expectation that it might do at least marginally better is therefore plausible.’ [34]. If this is true for us as humans, it may as well be true for the artificial entities that we allow into our societies and that, crucially, we allow to make decisions for us.

It may follow then that introducing a way for these entities to reason with and about trust, and its allied phenomena, such as regret and forgiveness, gives them a more solid footing in the human societies into which they are introduced. Any computational formalism or formalization of trust is a step in that direction. While some are used, for instance, to judge the reliability or manage the reputation of strangers; others, such as that proposed here, are more generalized (and as a result perhaps less tractable).

It is possible to argue that of course, we cannot trust machines, merely rely on them to operate as promised, indeed that ‘people trust people, not technology’ [32], but that is of course exactly the point – in designing a scheme for computational trust we are allowing the technology to reason about trust between people. That the agents in the deliberation may be human *or* technological, as Cofta states [15], and can reason about trust within *each other* is an added bonus.

What should be clear from the preceding discussion is that this work is not in itself the development of a Trust Management system, where social agents consider each other using potentially shared, inferred, and transitive trust or reputation (for instance, [38, 48, 82]. While naturally, considering how much you might trust another necessitates such a system when the other is a stranger, it is not the focus of our work, which is the individual considerations and internal workings of trust. Without both, there is a lack of *completeness* for agents making social trust deliberations.

3 A Parable of The Modern Age

Consider Ambient Intelligence (AmI).

AmI is a big thing. Stemming from Weiser's [90] vision of ubiquitous computing, and from there through the European Union's Information Society Technologies Program Advisory Group, it has become a vision of 'intelligent and intuitive interfaces embedded in everyday objects [..] responding to the presence of individuals in an invisible way' [3]. A good vision, no doubt, but one that is nevertheless still some way away. The technical aspects of AmI are within our grasp, but what of the social? In particular, since in such a vision information will be flowing around us – invisible, and highly personal – how are we to ultimately trust what is said about us? More, how are we to ensure that the systems that 'represent' us will reflect our values when deciding whether and how to share information.

These questions are not new, but in general, the assumption is that it'll all be alright on the night and we can convince people to trust the systems because they'll be built to be trustworthy. A great many advances have taken place in the field of trust management dedicated to exactly this concept, and they will result in better and more trustable systems. However, we are forgetting that the other side of the trust coin is risk.

The question is, what really is going on? Ultimately, the AmI goal is about sharing, and reasoning with, enough knowledge and information that sensible, socially appropriate things can happen for individuals. As individuals, then, we can help the AmI environment make the correct decisions, and potentially avoid some of the worst pitfalls. Trust, as has been pointed out elsewhere [58], is an excellent tool in this regard.

Consider then, this conceptual story about granting access to information. At most levels, Steve is happy with lots of people to see lots about him. Some information is more private, yet more is more private still. Steve can say all this to his agent in the AmI environment and let it get on with things. In a truly artificial society, the likes of which AmI aims to support, his agent will share data with, and get data from, other agents (Steve need not see this happening, and Steve need not see the data). The sharing of this data can be based on trust.

Imagine that Steve's friend Alice also has a device on which is an agent. Alice is a close friend and so their agents are also in close contact. One day Alice's agent requests some information about Steve that is private, but for a legitimate reason. Maybe it's a credit card number, maybe it's health information, or maybe it's just a password for Steve's private photo site. The point is not that the information is *necessarily* harmful, it's just private as far as Steve is concerned. Given their (and Alice and Steve's) closeness, Steve's agent reasons (we'll get to that) that it can trust Alice's agent with that information. Now, Alice's agent knows this fact.

One fine weekend, Alice, Bob and Steve are out canoeing and kayaking in Algonquin Park. Alice's agent, still humming away back in town, inadvertently lets that snippet of private information out into the wider community. There's obviously no bringing it back, the damage is done.

Steve do not know about all of this, but his agent, being a member of that community itself, does. Feeling that trust has been betrayed. It contacts Alice's agent to request an exlanation. Alice's agent expresses regret, stating that the leak was unintentional and a result of a faulty decision in response to a request made by an external agency.

Based on, and mitigated by, Alice's agent's expressions of regret, Steve's agent reduces the amount of trust it has in that agent, effectively removing it from his closest circle of friends. This trust information is propagated to other close friends' agents (say, those in a direct contact with Steve). Because this information is shared, the other agents in Steve's circle of friends respect this decision and act accordingly, ensuring that no further information about him is shared with Alice's agent, and revising their own trust levels (although it is reasonable to assume that some might not alter their levels at all whilst still respecting the information block.)

Time passes, and depending on the severity of the leak, and the regret Steve's agent feels, coupled with that of Alice's agent, a forgiveness process kicks into play. Slowly, Alice's agent is trusted with more information (but monitored more closely, because of this), and eventually, at least potentially, allowed back into the circle of close friends. All is well.

Until Steve comes home and find out what happened.

Steve may wish to censure the agent myself, or even Alice for buying such a terrible implementation of an agent (she always was a cheapskate) but he has two choices here. His agent can explain its reasoning, he can accept its judgment and carry on. Otherwise, he can instruct it to censure Alice's agent once more, and be subject to his own decision about forgiveness. Steve will have to sort it out with Alice herself, but that's what people do. Alice can express her own regret, and ultimately the relationships, both human and artificial, can be repaired.

As an aside, Steve can also censure his own agent, revisiting the amount of 'trust' he had in it and revising it as he sees fit. Perhaps next time Steve won't be daft enough to let a connected machine hold sensitive personal information.

3.1 A Brief Sojurn to 'Human Factors': Why Not Call it **Trust** *After All*

Of course, one may argue that the agent isn't *really* trusting others, or even regretting what it did in sharing that information, but it certainly acts *as if* it is. So what's the difference? While that is largely a discussion for philosophy, there is one thing we can learn here: the agent can justify its decision in terms that Steve can understand. He knows what trust is, or at least, like most people, has an idea of how it works *for him*. So when he requests an explanation of why his agent shared this data with Alice's agent in the first place, as well as how it handled the situation, it uses words and concepts Steve readily understands – regret, forgiveness, and trust. The agent may or may not be trusting, or feeling forgiveness, that's ultimately for Steve to decide, but the explanations are understandable. There are some parallels here with expert systems and the way in which they justify their own decisions via backtracking, but ultimately the use of the loaded terms of trust and other human understandable phenomena is, we conjecture, a more comfortable 'relationship' between user and technology.

There is much at stake here. As noted above, the acceptance of using trust as a means of helping make decisions is that sometimes trust gets misplaced. Mistakes are made. Risk is inherent in the consideration. As well, trust is vague: it's not the same thing to all people, and even if it was, my high trust may be equivalent to your low trust, depending on our personalities, because trust is seen in action, not thought. It may be possible to set up a 'soft secure' system using nothing more than trust values, but the risk is that they may be misinterpreted, or interpreted differently than the way Steve would have liked. This will happen. Trust and its siblings are not a panacea for technological ills.

4 Trust as Was

Trust has been extensively studied as a computational phenomenon in the past decade or so, and various models exist (e.g., [1, 76, 84, 39, 52]. Marsh's model appeared first in 1992 [63] and in revised form in 1994 [64]. While it has its problems, it remains as a standalone model capable of being adapted, revised, and revisited. This chapter in fact revisits and alters the formalisation for ease of incorporation of regret and forgiveness, and in line with what we have learned in the past few years. However, it seems prudent to explore what is being revised before actually doing so. This section, then, presents the model before we move on to the matter of making it in some way different.

As we have mentioned in section 1, the key here is to note that the purpose of this formalisation is not to accurately model social trust, but rather to give a piece for discussion and better understanding of the behaviour of the phenomenon, either artificial or real.

Bearing in mind that distinct trust levels are ambiguous at best (at least in terms of semantics and subjectivity [1, p.124]), we'll use them anyway. We believe benefits far outweigh their disadvantages, and include the ability to narrow down and discuss subconcepts (as is shown below), (computational) tractability and the ability to discuss and compare to some extent, and given a limited amount of space here, we'll argue the point at length elsewhere.

From [64] we use the notation shown in table 1. For more information discussions on the use of values and their ultimate frailties, see [64, 76, 85], amongst others.

Description	Representation	Value Range
Situations	$\alpha, \beta, \ldots$	
Actors	$a, b, c, \ldots$	
Set of Actors	$\mathscr{A}$	
Societies of Actors	$\mathscr{S}_1, \mathscr{S}_2 \ldots$ $\mathscr{S}_n \in \mathscr{A}$	
Knowledge (e.g., x knows y)	$K_x(y)$	True/False
Importance (e.g., of α to x)	$I_x(\alpha)$	$[0, +1]$
Utility (e.g., of α to x)	$U_x(\alpha)$	$[-1, +1]$
Basic Trust (e.g., of x)	T_x	$[-1, +1)$
General Trust (e.g., of x in y)	$T_x(y)$	$[-1, +1)$
Situational Trust (e.g., of x in y for α)	$T_x(y, \alpha)$	$[-1, +1)$

Table 1 Summary of notation ('Actors' are truster, trustee and others).

Some explanation is in order before continuing. We see time in this system as a set of discrete states, at each of which an agent may find itself in a given *situation* – a need to carry out some task, get some information, send some, and so on. In this situation, an agent has decisions to make about who it might trust, and how much, in order to carry out its task. The passage of time, the introduction of new agents, the changing of priorities, and more, can all have an effect, and create what is ultimately a new situation for that agent.

We do not think that this is ultimately very different from 'real' life. Others may disagree.

The formalisations in [64] attempted to answer questions about trust in cooperative situations. That is, given the choice between cooperation and non-cooperation, whether to cooperate with a specific trustee or not. We make a simplifying assumption, for the purpose of the consideration, that there are two protagonists. The systems works for more, however: just figure out which you trust the most in this situation.

Two formulae are used, the first being to estimate Situational Trust, the second to estimate a Cooperation Threshold. To estimate situational trust, an entity x uses:

$$T_x(y, \alpha) \; = \; U_x(\alpha) \times I_x(\alpha) \times \widehat{T_x(y)} \tag{1}$$

The $\widehat{T_x(y)}$ here emphasises that x can use previous trust-based knowledge in y in this calculation, whether related to this situation or not [64]. Thus, *at this time, in this* *situation*, x has *this much* trust in y. It's important to make this clear because in a different situation, this may be very different – if the situation is more important to x, for instance.

This is, though, only half the story. Regardless of how much x might trust y, any given situation might put x in an interesting decisional position. The consideration is how much do I *need* to trust you to cooperate with you in this situation? The answer lies within the *Cooperation Threshold*:

$$\text{Cooperation\_Threshold}_x(\alpha) = \frac{\text{Perceived\_Risk}_x(\alpha)}{\text{Perceived\_Competence}_x(y,\alpha) + \widehat{T_x(y)}} \times I_x(\alpha) \quad (2)$$

This gives us a means of seeing what is necessary for x to accept any cooperation with (help from) y in this situation. We can state that,

$$T_x(y,\alpha) \geq \text{Cooperation\_Threshold}_x(\alpha) \Rightarrow \text{Will\_Cooperate}(x,y,\alpha)$$

It is a truism to say that, when trust is upheld, it is strengthened. When betrayed, it is weakened. Most practitioners accept this statement, with caveats here and there. In our earlier work [64], we proposed that: If:

$$\text{Helped}(x,y,\alpha)^{t-\delta} \wedge \text{Defected}(y,\beta)^{t} \quad (3)$$

Then:

$$T_x(y)^{t+1} \ll T_x(y)^{t}$$

Informally, if x helped y in the past, and y responded at this time by defecting, the trust x has in y will reduce by a large amount. The converse is if:

$$\text{Helped}(x,y,\alpha)^{t-\delta} \wedge \text{Cooperated}(y,\beta)^{t} \quad (4)$$

Then:

$$T_x(y)^{t+1} \geq T_x(y)^{t}$$

Informally, if x helped y in the past, and y reciprocated at this time with cooperation, then the amount of trust x has in y will remain the same or increase only by a small amount.

In other words, the amount of trust x has in y substantially decreases following y not reciprocating [10]. However,y's reciprocation merely confirms to x that she (x) was correct in helping y in the first place [53]. This being the case, x had every right to *expect* y to help. So, although y's reciprocation may lead x to trust her *judgement* of people more, she may revise her trust in y only slightly, if at all [53].

However, beyond these musings, little was said about how much was a lot, or a little, with respect to how to alter trust values. We revisit this below.

5 What Can't Trust Give Us?

It would be wise to consider trust as a part of a solution for any artificial (or natural) socially or culturally embedded entity. Just as humans are more than trusting or untrusting creatures, and use trust as a part of their decision making process, the same applies to artificial agents.

Most importantly, trust cannot give us *certainty* – it is a judgment based on evidence, potentially 'irrational' feelings (in humans), and is often skewed in one way or another. In fact, to trust inherently holds with it the risk of betrayal [61] – if certainty was what was sought (and achieved), trust would not be necessary.

Trust cannot give us *control*. Control is the antithesis of a trusting relationship because it implies that one is not putting oneself into another's hands (which is what trust is), but that one has the right and the power to enforce behaviour in others. That is not to say that trusting others does not bring some form of control, at least in a moral sense (as in fact does forgiveness, if taken to extremes). Thus, if I say 'I trust you' and you are trustworthy and a moral person, you will feel obligated to work in my best interests. Needless to say this control is flimsy and easily ignored if you're not a morally righteous person!

Trust can't give us *confidence*. It can give us a sense of risk-laden comfort about the path we have chosen, but it isn't the same as knowing (being confident) that someone you are buying online from will deliver the goods and not overcharge (cf [15]). Confidence is often achieved through rules and regulations that are backed up by a trustworthy legal or social system (the irony in that sentence is not lost to us).

In short, trust gives us little more than a soft relationship with another entity. If that entity values the relationship, understands the meaning and culture of trust *and is trustworthy*, we're likely okay to trust. If any of these pre-requisites fails, we might well be in trouble. Ultimately the same applies to other soft notions as regret, forgiveness and morality. That doesn't make them useless – they have a power that is not physical or, usually, legally binding, and convey upon the trustee weighty responsibilities which, all things being equal, are not easily ignored. Certainly, we feel that they can be the *basis* for social behaviour and decision making in a moral social world, and potentially have strength even in a darker society, as long as there are some trustworthy agents out there. Various experiments with trust uphold this view well enough (see [91, 4, 80], amongst others).

6 Trust As Is, Part Zero: The Dark Side

In [68] we presented to the trusting agent the concepts of distrust, untrust and mistrust. Distrust has in fact become a much more popular object of study, although given the number of definitions of trust [15, 70, 71], distrust is at least as difficult to pin down. Distrust is often considered as the "negative mirror-image of trust" [86, page 26], a "confident negative expectation regarding anothers conduct" [55, page 439] in a situation entailing risk to the trusting party. In situations where trust ‘betrayals’ happen, trust can easily move towards distrust or untrust. Thus, since they bear relevance to our discussions of forgiveness and regret, the concepts are summarised here.

Trust is a continuously assessed, continuously variable measure, or relationship, between actors. It has positive and negative aspects, and indeed positive and negative values, at least in our model. Indeed, we can see trust as a continuum (see also Cofta’s trust cube [15, page 109] for a more dimensional model). Figure 1 illustrates the continuum, with negative trust values being seen as ‘distrust’ while positive trust values are seen as ‘trust’. But there are gaps, in both the figure and our understanding, that are addressed in this work.

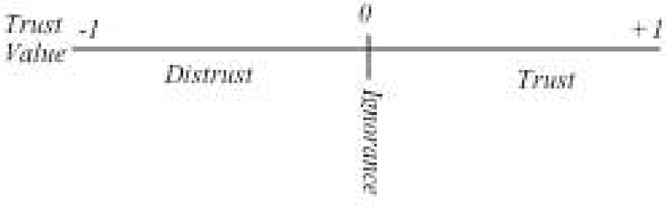

Fig. 1 Trust Continuum: From Distrust to Trust

In [64] we stated that distrust was negative of trust. Here, we’re evolving that definition because of the work that has been done in the area, and a greater understanding of the concept because of this work. That given, it’s still surprisingly difficult to find definitions of distrust that don’t use mistrust as synonymous. In fact, we believe this is a mistake because it removes a tool for trust researchers to be able to focus on what they are researching. For clarity, in [68] we used a comparison with the concepts of *misinformation* and *disinformation*. From the Oxford English Dictionary, we find that the term ‘misinformation’ can be taken to mean information that is incorrect. This can be a mistake on the part of the informer, and generally speaking, it can be spotted after the fact. The term ’disinformation’ removes all doubt – it iss information that is deliberately false and *intended* to deceive. That is, disinformation is misinformation that is deliberately and knowingly planted. From this, we moved to a better understanding of distrust and mistrust, and what untrust is.

A simple comparison between the concepts is probably necessary. For the sake of argument, following [10, 59, 79, 22, 61, 64], let’s say that trust, in general, is taken as the belief (or a measure of it) that a the trustee will act in the best interests of the truster in a given situation, even when controls are unavailable and it may not be in the trustee’s best interests to do so. Given this, we can now present untrust, distrust and mistrust.

6.1 Distrust

If we are to take disinformation as deliberately planted, that is, intentional and active misinformation, our extension with distrust is that it is also an active phenomenon, that is to say, when x distrusts y, it is because x has considered the situation, and actively *believes* that y has negative intentions towards her. We can put this semi-formally as:

$$T_x(y,\alpha) \ < 0 \Rightarrow \text{Distrust}(x,y,\alpha) \tag{5}$$

So, for this situation, x believes that y does not have her best interests at heart. Not only that, but y will actively seek to work against those best interests (this is not a failure of omission, in other words). As with a measure of trust, the greater the magnitude, the more the certainty and the greater the strength of belief that y will be actively against x's best interests.

6.2 Mistrust

Accepting that Misinformation is passive in some form (that is, it may or may not be intentional, and is a judgment usually attained after the fact), we similarly conjecture that *Mistrust* is misplaced trust. That is, following a decision in which there was a positive estimation of trust, and where one is betrayed, we can say that trust has been misplaced (not always 'betrayed,' since the trustee may not have had bad intentions). Thus, the truster *mistrusted* the trustee. As we see in [1] mistrust is defined so "When a trustee betrays the trust of the truster, or, in other words, defaults on trust, we will say that a situation of mistrust has occured, or that the truster has mistrusted the trustee in that situation." (p.47).

Note that this works both ways, and one can mistrust by assuming the other is 'distrustworthy' when in fact they are 'on our side' [15, especially chapter 6], although it's harder to recover from that, or at least spot it, since in such a situation we're unlikely to give the other the chance to prove it.

This is perhaps something of a departure from traditional english usage, which tends to confuse distrust and mistrust, but we feel that for a computational model, some degree of accuracy and definition is required!

6.3 Untrust

As complicated as life is, it's unlikely that there are black and white aspects of trust without a little grey. The reader will have noticed that, given a specific situation, the cooperation threshold puts an artificial barrier somewhere along the trust continuum. It's likely that this barrier exists within the positive side of the spectrum, and so we have a situation where a trustee is viewed positively but not positively enough to cooperate with. Given Barber's [6] view of trust based on continuity, competence and motivation, evidence against any of those may well result in this situation, as we noted in earlier work [64] – I may trust my brother to drive me to the airport, but flying the plane is a different matter. This isn't because I distrust him, it's because I know he can't fly planes. In previous work [68] we stated that 'if we say a trustee is untrusted, then the truster has little confidence (belief, faith) in the trustee acting in their best interests in that particular situation', but that's not strictly true, as my brother's example shows (I presume he has my best interests at heart!). Of course, if he really did have my best interests at heart and knew he couldn't fly a plane, he wouldn't ordinarily offer...

This grey area is what Cofta calls *Mix-Trust* [15], and what we have chosen to call *untrust*.

We can present untrust formally as:

$$T_x(y,\alpha) > 0 \; \& \; T_x(y,\alpha) < \text{Cooperation\_Threshold}_x(\alpha) \Rightarrow \text{Untrust}(x,y,\alpha) \qquad (6)$$

That is, if $T_x(y,\alpha)$ is less than the Cooperation Threshold but larger than 0, x is in a state of *untrust* in y. That is, x 'doesn't trust' y, but bear in mind that in fact the amount of trust is positive, which perhaps gives x some incentive to try to find a way to cooperate with y. Section 6.5 revisits the trust continuum to put untrust in its proper place on the map. The story isn't over for untrust yet either, as we'll see.

6.4 Ignorance is...

Ignorance, the state where x knows nothing of y at all, or the situation she finds herself in, is classed as a zero state, thus $T_x(y,\alpha) = 0$. This is both very unusual and difficult to handle, but nevertheless needs to be acknowledged. It's unusual because, in general, we can conjecture from previous experience either about potential trust in others – so called *Basic Trust* (cf [64]) – and in situations (although this may be more difficult for an artificial entity.)

6.5 The Continuum, Revisitied

We are now beginning to see how distrust and it's siblings in fact are present on the continuum of trust. As we see in figure 2, we still have negative trust being distrust, and now we can have a new section on the continuum, where untrust lives – below the level, for any given situation, of the cooperation threshold, yet still in the realms of positive trust.

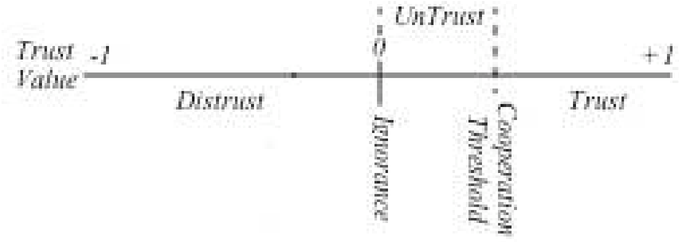

Fig. 2 Trust Continuum: Untrust

The figure does not include *mistrust*, for the simple reason that mistrust is everywhere – in other words, it's possible to make errors in trusting (or distrusting) estimations throughout the continuum. Thus, one can think of mistrust as an overarching *possibility* across the continuum of trust.

We will revisit this continuum later in the chapter.

6.6 Continuing a Difficult Relationship

As we have noted previously, distrust and untrust are important not because they may stop some relationships, or cooperation, but because they may in fact allow something to continue [68, page 21]. Consider a situation where *x* has little choice but to cooperate in some way with *y*, even while she distrusts *y* – the measure of distrust allows *x* to make hedges in order to achieve a greater comfort (and control [15]) over the errant *y*. More importantly, if *x untrusts y* there is evidence somewhere (for instance, using Barber's assessment classes [6]) that there is a positive relationship to work on in order to achieve greater comfort or control for *x*.

Trust and its siblings are not, then, the only decision, or control tool available to an agent, they are just some of many. In the final analysis, they may indeed be little more than pointers to the need for remedial work on a relationship, or a legal contract, or letters of reference, and so on. If this is so, their value is no less than if they were the ultimate arbiters of relationships.

That said, we do believe that trust is simply a part of the decision making puzzle, and that other psycho-social phenomena and/or emotions play a large part in the decisions people make. We have chosen to focus in this work on two of these, regret and forgiveness, and how they interact with trust and each other.

7 Regret

'Trust is only required if a bad outcome would make you regret your decision.'

Luhmann, 1978, page 98.

Regret has been studied in psychology and economics for some time. In 1982, both Bell [7] and Loomes and Sugden [57] independently introduced the concept of regret theory in economics, itself based on the social psychological theory of counterfactual thinking [83]. In game theory, the Savage/Regret Minimax theory has existed for some time [60], itself based again on interpretation of psychological theories. Further, regret continues to be an active field of study in psychology and economics, as well as philosophy and health care [2, 16, 7, 57, 26].

In one form, regret is a form of cognitive dissonance that an actor feels when what is done is in dissonance with what the actor feels should have been done [31, 43]. After a decision is made, if it is not supported by what we think is 'right,' we will feel 'bad' about it. This is of course not the only aspect of regret that is of importance here, although it does give additional decision making aspects for trusting agents. Another aspect of regret is experienced following a trusting decision that is betrayed by the trustee. We will attempt to formalise both aspects of regret here.

When a decision is made to trust, effectively in error (a mistrusting decision, here), what happens next is a betrayal of that trust. It is important to consider two types of betrayal – first, where the trustee *knew* they were trusted, and second, where they *did not know*. In the first instance, we propose that the regret felt by the truster is greater than in the second, because the truster has reason to believe the trustee made a conscious decision to betray the trust. Of course, there are both mitigators and exacerbators for both kinds of betrayal and these include:

- The regret felt by the trustee post decision;
- (Not the same), acknowledgment of the betrayal by the trustee;
- The magnitude of the 'betrayal' – whether the trustee knows of it or not;
- Reparations;

7.1 What Regret Is

Regret allows an action or happening to be looked upon as negative, and further allows the actors, or observers, to reinforce behaviours or associated feelings or emotions (such as trust) to ensure that the likelihood of such a thing happening again is reduced. It is, therefore, a powerful motivational force in interactions with others. Further, because it can have an effect on trust, it is necessary to study, formalise, and concretise regret to the extent that it becomes a computational tool similar to the current status of trust.

However, regret, while a tool for hindsight, is also a predictive tool. For instance, it is possible to say I am going to regret this (and then do it anyway!). In this way, regret, as with trust, allows a consideration of possible alternatives in a situation in order to choose the best, or most likely not to cause regret, for instance. Thus regret, like trust, is a powerful tool in the consideration of actions and alternatives. When allied to trust, it becomes much more powerful and predictive.

7.2 The Many Faces of Regret

Like most terms that encompass human feelings (including trust), regret is somewhat overloaded. It is possible to regret something that one is personally involved with (or did), and it is possible to regret something that was done, or happened. There are, then, several valid uses of the term. Additionally, it may be possible to feel that something should not have been done without necessarily regretting that it was. Moreover, it is possible to regret something but have seen no choice. For instance, "I wish I hadnt done that" is not the same as "I feel bad for having done that," which might be the same as "I regret having done that." But regret can encompass other things too: "It is with regret that I must announce the death of" is for example not an admission of fault, but is an admission that what has happened is regretted, that the feelings involved are negative. Similarly for statements such as "That was a regrettable incident."

7.3 Modeling Regret

Largely, in this work, we are concerned with answering the questions:

- What was lost (κ)
- What it meant (λ)
- How it feels (μ)

While the latter question is harder to estimate for an artificial entity, the first two are relatively straightforward if we can also consider what has come before, thus if there is a potential measure of utility, we can use this in determining what was lost, and if there is a measure of importance (and perhaps trust) we can use this in determining what it meant to the agent concerned. The rest, we may have to leave to the owner of an artificial entity, rather than the entity itself.

I Regret That You Did That

A truster, when betrayed, can feel regret that they were betrayed. In simple utilitarian terms we can say that the regret felt is based on opportunity cost, or the amount of utility lost from the betrayal as compared to what could have been gained (this is in fact similar to the Savage/regret Minimax criterion [60]). We suggest that there's something more to it than that, simply because there was in fact a trusting decision made. Bear in mind that in much extant work this decision would imply that in fact there is much more to lose than would have been gained in the decision to trust [61, 23] (but this view is somewhat mitigated in [40, 64]). In any case, the decision to trust has put the truster in the trustee's hands at least to some extent [10] and thus the betrayal (whether the trustee knows of it or not) is felt more personally (as a caveat, though, consider that 'trust can only concern that which one person can rightfully demand of another' [45, page 319] when thinking about regret).

Thus, we add considerations not only of utility, but also of the trust that was originally placed in the situation to our regret function:

$$\text{Regret}_x(\alpha) = (U_x(\alpha) - U_x(\alpha^-)) \bullet f(\kappa, \lambda, \mu) \tag{7}$$

Where:

- The $\bullet$ denotes some operation (presently, we use multiplication);
- $U_x(\alpha^-)$ is the utility gained from what happened (the 'betrayal' situation) as opposed to what was originally estimated *could have been* gained $(U_x(\alpha))$;

Note that we see regret as a primarily situational phenomenon. It not only simplifies the agent's considerations of what is regretted, but allows a significant amount of control over what is assessed in the regret function.

The function addressing our primary questions (what was lost, what it meant and how it feels) is addressed partly here. We are working continuously on refinements.

There are considerations. Firstly, that the amount of trust that existed, as well as the Cooperation Threshold, are important aspects in the regret measurement, and secondly, that the relationship itself is of potential importance in the final analysis. This is consistent with [54]'s analysis of Calculus-Based, Knowledge-Based, and Identification-Based Trust, and goes a little way towards not only answering *what it meant*, but also *how it feels*, for our agent.

Thus we propose:

$$f(\kappa, \lambda, \mu) = \text{C\_T}_x(y, \alpha)^t + \text{I}_x(xy) \tag{8}$$

- $\text{I}_x(xy)$ is the importance, to x, of the relationship xy – see below for more discussion of this;
- $\text{C\_T}_x(y, \alpha)^t$ is the Cooperation Threshold for x at that situation.

Here, *what it meant* (Cooperation Threshold, which took into account trust in the first place) and *how it feels* ($\text{I}_x(xy)$) are addressed, with what was lost being taken into account via the incorporation of utility.

The importance of the relationship features prominently here

Clearly there are times when nothing is known of the other, and so we use, very simply:

$$f(\kappa, \lambda, \mu) = \text{Importance}_x(\alpha) \tag{9}$$

Hence, the more important the situation was, the more regret is felt that a betrayal occurred. Don't forget utility is also taken into account. Again, we strive to answer *what was lost* (Utility) and *what it meant* (in this case, Importance).

There is much work to be done here, and we are addressing it. For instance, even when nothing is known of the other, sometimes the relationship is still important (for instance, when dealing with authority).

You Regret That You Did That

As noted above, the other side of regret is where the transgressor (the trustee) expresses (or feels) a sense of regret for what they have done (this is related to the idea of post-decisional dissonance [81]). We have suggested above that this feeling of regret need not in fact be accompanied by some form of acknowledgment of wrong done. This is more applicable when the regret expressed is over something that was outside the control of the transgressor, for example. More plausible, perhaps, is acknowledgment without regret, which we do not cover here. For the sake of simplicity, we will not be considering these different angles here, focusing only on the expression of, and feeling of, regret, and how it potentially effects trust.

The formula is similar to equation 7:

$$\text{Regret}_y(\alpha) = (U_y(\alpha) - U_y(\alpha^-)) \bullet \text{I}_y(yx) \tag{10}$$

Note that firstly, the consideration of regret here must be taken from the point of view of the transgressor, but in some what calculated (or transmitted to) the truster (who was betrayed). This is something of a problem area (y could lie, to try preserve the relationship and benefit, perhaps, from more transgressions) and needs to be further addressed.

Here, y may regret having done something, but again expressing this in a purely economic sense is not acknowledging the role of the relationship (and trust in some way). The inclusion of the importance of the relationship to y mitigates any benefit y may have gained from y's betrayal.

Given these measures of regret, it is up to x to decide how to use them in mitigating the initial, and continuing, effect on trust of the transgression. In this work, the regret calculations are simply a part of how forgiveness, the repairing of the trust relationship, works in a computational sense.

I Regret That I *Didn't* Do That, and Derivatives

Consistent with findings from counterfactual thinking and regret [36], there is evidence to suggest that we often regret that we *didn't* take more risks, do more things, at least in specific, and so on, as we grow older – a lost opportunity is something none of us appreciate. While this seems somewhat odd to think about in terms of AmI and trust, in fact, the *Anticipated (Anticipatory) Regret* (AR) of not doing (or of refraining from doing) something is potentially a powerful motivational force in actually getting us to take risks and trust more. That being the case, we can incorporate AR into the trust considerations of an agent. Indeed, we feel that AR has a role in determining the Cooperation Threshold for an agent in a given situation.

A development from [64] taking this into account gives a simple proposal for a derivative of AR:

$$\text{Cooperation\_Threshold}_x(\alpha) = \frac{\text{Perceived\_Risk}_x(\alpha)}{\text{Perceived\_Competence}_x(y,\alpha)+\widehat{T_x(y)}} \times (I_x(\alpha) - \text{AR}(\alpha^-)) \tag{11}$$

Thus, here, the utility of α^- can be taken as a positive motivational force, because α^- may be regretted if *not* done. Note that, the determination of AR in this circumstance is not necessarily different from in, for example, equation 7, but a negative regret from that equation would be a positive AR in equation 11. Equation 11 is in fact more properly *I Will Regret it if I* Don't *Do That*, a much more useful tool for the computational trusting agent.

There is also much work on using this as a tool for trying to avoid doing something that we will regret later (see for example [7, 57, 2]). I may decide against smoking another cigarette, or going out drinking the night before an exam because I know that I'll regret it later (bad health, bad grades, and so on). Once again this can be a powerful tool in decision making, although it's more properly characterized as *I Will Regret it if I* Do *Do That*. The calculation is similar to that in equation 11.

8 Trust as Is, Part One: Building Regret into Trust

Now, it is possible to think about how regret can be used to both mitigate the behaviour of others and to respond to it. We have in the past [64, 65] considered the adjustment of trust values following transgressions or cooperation, particularly as regards optimism and pessimism. It is the adjustment of trust, in fact, that both forgiveness and regret will have an impact on.

As a start, consider the following:

$$T_x(y)^{t+n} = T_x(y)^t \pm f(\text{Cooperation\_Threshold}_x(\alpha)^t, T_x(y,\alpha)^t) \tag{12}$$

Thus, the amount of trust x will have in y at a subsequent timestep ($n > 0$) will be dependent on the situation x was in, via some analysis of the relationship between the cooperation threshold and situational trust – intuitively, and for the sake of argument, the greater the difference in one direction or another between these thresholds, the more the effect on the adjustment. In fact, for an upwards movement, this may be a reliable method, but there is general agreement, at least for what [54] call *Calculus Based Trust* and *Knowledge Based Trust*, that trust is in fact relatively fragile – that is, hard to build up, and easy to lose. A sensible function in equation 12 will naturally have to take this into account. In the past we have also used a simple percentage calculation, thus the more y was trusted, the more the movement in trust (downward, at least).

Taking into account a transgression, it's now possible to enhance this equation to take into account regret:

$$\begin{aligned} T_x(y)^{t+n} = T_x(y)^t - f(&\text{Cooperation\_Threshold}_x(\alpha)^t, \\ &T_x(y,\alpha)^t, \text{Regret}_x(\alpha), \text{Regret}_y(\alpha)) \end{aligned} \tag{13}$$

In our current work, for this important function, we use:

$$f = \frac{\text{C\_T})_x(\alpha) + T_x(y,\alpha)^t)}{\Xi_x} \times (\text{Regret}_x(\alpha) - \text{Regret}_y(\alpha)) \tag{14}$$

The value of Ξ_x is anything x chooses. The higher it is, the more 'volatile' the agent is, and the less 'understanding.' For this reason, we call Ξ the *understanding constant* for an agent. The lower the understanding constant, the more understanding the agent. In our work we use a value between 1 and 10, but really, most values go, as long as the result isn't too (rationally) challenging.

The outcome of such a calculation is that the agent may pass from 'trust' through untrust and on to distrust. The magnitude of the change is dependent on the magnitude of (or importance of) the situation, betrayal, regret, and so forth. Here. the more y is trusted, the greater the loss of trust. However, it's not a complete loss as postulated by many. That could easily be handled by, for example, stating that if $T_x(y,\alpha)^t$ was above a certain threshold, dependent on the agent, then $T_x(y)^{t+1}$ could simply be reduced by, for instance the value of $\text{Regret}_x(\alpha)$. There is no easy answer here, as in [54] we find that Identification-Based Trust is potentially strong enough to absorb transgressions without necessarily major alterations. It all depends, in other words, on how you want your agent to behave (and some variability makes for a much more interesting world).

For honest trustees who do not transgress, we continue to use a percentage increase, with the percentage value itself decreasing as we approach a trust limit of 0.99[2].

9 Forgiveness and The Blind and Toothless

> If we practice and eye for an eye and a tooth for a tooth, soon the whole world will be blind and toothless
>
> Gandhi.

If one is to assume that regret can be expressed, shown, or made to be felt, it would appear that we have arrived at a situation where there is a great deal of the stuff, but very few things to do with it — we can make decisions, review them, and even come to a different understanding of trust. But it's not enough if a 'next step' is not considered. In our work, we see this next step as that of forgiveness.

[2] We have discussed elsewhere [64] why trust values of 1, indicating blind trust, are not trust at all (since, being blind, they do not by definition take any consideration by the agent about the situation or others in it into account).

9.1 What Forgiveness Is

> To err is human; to forgive, divine.
>
> Alexander Pope

Forgiveness is something of an enigma. While social psychologists appear more comfortable defining what it is *not* [27] (it isn't forgetting, for example, and it doesn't imply reconciliation, but it is a conscious decision), there is some evidence that they may be out of step with what people actually think it *is* [47]. A good start is given by, Vasalou and Pitt see forgiveness as a 'prosocial decision to adapt a positive attitude towards another' [88, page 146], which neatly removes the need to say what it actually results in.

It is through forgiveness that trust can be restored in relationships, and that, consequently, things that were impossible before can become possible. The act, and expression, of regretting, which explicitly acknowledges that some bad thing has happened (but not necessarily culpability), is a major step on the road to forgiveness, and thus to restored trust. Forgiveness is not always required or justified where regret is voiced.

In Vasalou and Pitt's recent work, [88], the concept of forgiveness has been examined in the context of a reputation system. In their DigitalBlush system, expressions of shame, embarrassment, and so on are used to elicit potential forgiveness by others in the society. While acknowledging the fact that, applied too swiftly, or incorrectly, it may in fact be more problematic than if it were not applied (especially online), the system reinforces the idea that regret (however expressed), is a precursor to a potential forgiving act. In fact, there is a lively debate on the ethics of forgiveness in psychology [27, 78], but evidence to suggest that forgiveness is good for the forgiver and the forgivee [12, 89, 11].

There is little doubt that forgiveness is a particularly important area where trust is concerned. It is through forgiveness that trust can be repaired, and it is through forgiveness that cooperation can as a result be re-opened. We acknowledge, along with [88], the potential problems forgiveness may create, but we feel that it is too important, and too beneficial, to ignore as a computational concept. In our own work, we are concerned less with helping people forgive that allowing artificial or analytical systems to consider forgiveness as a tool, for example when making trusting decisions.

9.2 A Model of Forgiveness

> The weak can never forgive. Forgiveness is the attribute of the strong.
>
> Gandhi.

For our own model, we are less concerned with the precursors to forgiveness than the mechanisms of the act of forgiving in and of itself. Naturally, we assume the precursors must exist (regret, as we have already discussed, is one of them, and used heavily here), but we see forgiveness as a step along the road to the re-establishment of trust. As a step, it can be seen in its own light. This view of forgiveness may be something of a departure from some views of the topic (it's not always seen as a restorative process, for instance), but it serves well here.

While making no judgment on whether or not forgiveness can happen in any given circumstance, we see two major aspects of forgiveness in an autonomous agent:

- The Forgiveness Trait
- The Forgiveness Function

First, note that these are individual to each agent, and therefore can differ radically between agents.

Consider the Forgiveness Trait for an agent. Put in its simplest form, this trait is an expression of the length of time after a transgression that must pass before the agent will even begin to consider forgiving. When this length of time has passed, the Forgiveness Function can come into play. This is in fact quite a simple parameter to set up in an artificial system, but also slightly proscribed. In fact, it makes much more sense to relate this length of time to the severity of the transgression, coupled with the Forgiveness Trait as an expression of the 'strictness' of the agent's 'moral code', represented as a percentage (this is important in the following equations), and once more this is simple enough to accomplish. Thus the length of time before forgiveness is:

$$t_{\mathrm{Ft}_x} = \mathrm{Ft}_x \times \mathrm{Regret}_x(\alpha) \tag{15}$$

With Ft_x expressed as a number between 1 and 100 - more forgiving agents have lower Ft values.

From this, then, we can calculate a number of timesteps between transgression and forgiveness that is related to the Forgiveness Trait of the agent, coupled with how much the agent regrets what happened (the more regret, the longer it takes to think about forgiving). As will be discussed further below, agent time and human time are subjectively very different things. You own mileage may vary.

The Forgiveness Function is likewise straightforward. Ultimately it is an expression of the transgression's severity, regret felt and expressed (the concept of shame and embarrassment is similar in [88], and the relationship that the agents have had before the transgression occurred. Formally (and normalised in some sensible way, which we discuss further below), the Forgiveness Function for a (very) simple agent is:

$$\mathrm{Fk}_x = \frac{(\mathrm{Regret}_y(\alpha) - \mathrm{Regret}_x(\alpha) + \mathrm{I}_x(xy))}{\mathrm{Ft}_x} \times T_x(y) \tag{16}$$

Thus, the more regret x has, and the less (or even negative) regret y has, the less forgiveness is forthcoming. Note also that the forgiveness is mitigated by the amount of trust that exists in the relationship (the higher it is, the more forgiveness). This trust could in fact be that which now exists as a result of the transgression, or what existed before - different results will be obtained from each. Consider for example a high trust relationship which after a transgression becomes a very low trust relationship - using the original trust value may be more forgiving, and a reflection of the value of trust, than using the post transgression value. These are considerations, however, for individuals (both agent and human).

10 Trust As Is, Part Two: The Incorporation of Forgiveness

Now that the Forgiveness Trait and Function are clarified, we can look at how forgiveness, when adapted, can enter the alteration of trust following a transgression.

Formally, over time, an agent who transgressed *may be* forgiven:

$$T_x(y)^{t+t_{\mathrm{Ft}_x}} = T_x(y)^{t+t_{\mathrm{Ft}_x}-1} + \mathrm{Fk}_x(\mathrm{Regret}_x(\alpha), \mathrm{Regret}_y(\alpha), \mathrm{I}_x(xy), \alpha^-) \qquad (17)$$

Where:

- Ft_x is x's *Forgiveness Trait*;
- Fk_x is x's *Forgiveness Function*;
- t is some time step in the future from the trangression;
- α^- represents the situation in which the trangression took place, and can be used to calculate other aspects, such as thresholds, etc.

We have introduced the Forgiveness Trait and Function above. Note here that any forgiveness consideration must take into account the situation in which the transgression took place, as well as the players (there may be more than two) in the situation.

10.1 The Trust Contunuum, Revised: The Limits of Forgiveness

Our considerations of forgiveness allow us to revisit what we may have known about the trust continuum – previously we considered untrust to be a positive trust, yet not enough trust for cooperation (see figure 2). However, it is possible to imagine a situation where a negative trust is not in fact distrust, but the result of a transgression that propels a specific agent's trust values into negativity. It's possible the transgression was minor, or even an honest mistake (consider Alice's agent sharing Steve's information), and it's possible to consider the agent standing in potential of forgiveness and remediation, regardless of the fact that the current trust we have in them is negative.

There are, however, limits. Some things cannot be accpted, and some agents are malicious or non-redeemable. To take this into account we introduce a new concept, the *Limit of Forgivability*, beyond which we might say the agent is truly distrusted, and cannot be considered as other than acting against our best interests. In considerations of forgiveness, this Limit will be used by agents to determine the worth of entering into redemption strategies with others. Note that a single transgression may well put another beyond this pale. Note that the Limit of Forgiveness is individual and personal (private). Keeping the limit private is in fact important in any situation where knowledge of how trust and forgiveness work can be used against an agent.

The Limit of Forgivability introduces the concept of untrust as a potentially negative phenomenon. This is shown in figure 3.

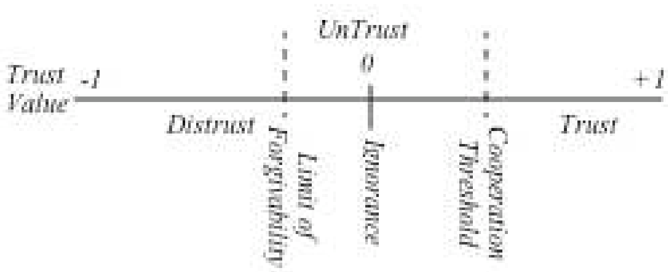

Fig. 3 Trust Continuum: The Limit of Forgivability

11 Applications: Revisiting the Parable and Imagining the Future

While interesting in its own right, often the theory and associated descriptions are note enough to shed light on what is imagined. Our current work is concentrated on the use of trust, regret and forgiveness in information sharing architectures, both as expressed in the parable in section 3 but also in more complex environments where trust amalgamation is key. Much of this work is theoretical, but it is possible to show worked examples, and this section is devoted to such an endeavour. In the first place, we revisit our parable, and work through the example showing how the individual agents work. Following this, we briefly present the wider information sharing concept, and how trust, regret and forgiveness, amongst other social norms, can work to enhance human experiences in this domain. Finally, we present our ideas for Soft Security and Regret Management.

In the course of this work, much of what is conjectured works well enough at first blush, but on application, even in such a limited sense as a worked example, errors and omissions are found. It's worth mentioning here that this is indeed the case in this work, and that the formalisations above have benefitted from such worked examples. In addition, the examples below illustrate some of the 'behind the scenes' considerations that agents and others must make in their deliberations, some of which are trust related, others of which are not.

11.1 The Parable at Work

Let us revisit the agents of Alice and Steve. Ordinarily they consider us to be friends and trusted. Let's say then that the amount of general trust Alice has in Steve (strictly speaking, Alice's agent has in Steve's own) is 0.85 – quite high[3] in other words, and Steve's in Alice is 0.80. We can express this as:

$$T_{Steve}(Alice) = 0.80$$
$$T_{Alice}(Steve) = 0.85$$

When Alice's agent requests that piece of information, Steve's agent has certain things to consider – the importance of the information to Steve, the utility of revealing it to Alice's agent (which could be based on furthering trusting relations, getting something back later, or just being nice because Steve likes Alice), Steve's trust in Alice (as seen by his agent), Alice's (agent's) 'information handling' competence (as seen from experience, and if not, then estimated by other means), and the risks associated with distribution. In this instance, there's little to be concerned about as far as Alice is concerned (little do we know. . .) Thus:

$$T_{Steve}(Alice, \text{info\_share}) \ = \ U_{Steve}(\text{info\_share}) \times I_{Steve}(\text{info\_share}) \times T_{Steve}\widehat{(Alice)}$$

Putting in some sensible numbers (this is relatively important, and Steve stands to gain becausehe know Alice needs the info for a book she's working on). We already know how much Steve trusts Alice. . .

$$T_{Steve}(Alice, \text{info\_share}) \ = \ 0.80 \times 0.85 \times 0.80 \ = \ 0.544$$

This may not seem like much, but bear in mind there's another set of considerations:

$$\text{Cooperation\_Threshold}_{Steve}(\text{info\_share}) = \frac{\text{Perceived\_Risk}_{Steve}(\text{info\_share})}{\text{Perceived\_Competence}_{Steve}(Alice, \text{info\_share}) + T_{Steve}\widehat{(Alice)}} \times I_{Steve}(\text{info\_share})$$

Again, with sensible numbers (it's risky, because the information is personal, but we see, so far, Alice's agent as competent in all dealings thus far – little do we know

[3] Regardless of how well we may be able to justify a value system, or the choice of a particular value, others will rightly state that 0.85 is way too high for a friend, while still others might say it's not high enough, again rightly. And here we arrive once more at the problem with trust – sharing values is just not going to work. Thus, keep your own, measure it your way, and measure by action, not statement, of values or otherwise.

that this is because Alice has (human)-fielded all her agent's dealing with others):

$$\text{Cooperation\_Threshold}_{Steve}(\text{info\_share}) = \frac{0.75}{0.7+0.8} \times 0.85 \; = \; 0.397$$

And so, since clearly Alice (her agent, at least) is trusted enough in this situation, Steve's agent shares this information.

Time passes and it becomes clear that Alice's agent has transgressed on the understanding, and the information is out there. Trust must be re-evaluated, based on what has happened. Recall from section 4 that we have in the past considered a very simple means of reducing trust following a transgression [64], in most cases this being a percentage of the original trust, with a greater or lesser percentage dependent on the 'personality' of the agent concerned. With the tool of regret, we have an additional means of re-evaluating trust. In this case, we might say that I (Steve's agent) regret that you (Alice's agent) did that, and so we can take a look at equation 7 and 8:

$$\text{Regret}_{Steve}(\text{info\_share}) = (U_x(\text{info\_share}) - U_x(\text{info\_share}^-)) \times f(\kappa,\lambda,\mu)$$

Where:

$$f(\kappa,\lambda,\mu) = \text{C\_T}_{Steve}(Alice,\text{info\_share})^t + \text{I}_{Steve}(Steve,Alice)$$

We already know some of the values here, and can fill in the others now. Certainly, Steve stands to lose now that the information is out. In fact, it can potentially cost him, so there is a negative utility to the current situation (say, −0.1, because it's not a huge cost, but will take time and effort to fix). Steve and Alice are good friends, and he values the relationship. Thus:

$$\text{Regret}_{Steve}(\text{info\_share}) = (0.8-(-0.1)) \times (0.397+0.75) \; = \; 1.03$$

This is greater than 1, and that's fine, but we could normalize it if needed. In this case, we can use the regret to calculate how much trust is lost. From equation 13 and 14, we have:

$$\begin{aligned} T_{Steve}(Alice)^{t+n} = \\ & T_{Steve}(Alice)^t \; - \\ & (\tfrac{\text{C\_T})_{Steve}(\text{info\_share})+T_{Steve}(Alice,\text{info\_share})^t)}{\Xi_{Steve}} \times \\ & (\text{Regret}_{Steve}(\text{info\_share}) - \text{Regret}_{Alice}(\text{info\_share}))) \end{aligned}$$

Now, Steve is a nice guy and understands that mistakes happen, and he doesn't like to punish people unnecessarily for that, so the *understanding constant* for his agent is set to 5. As noted above, it's an arbitrary choice, and to be determined by the agent's owner (we use anything between 1 and 10). So:

$$T_{Steve}(Alice)^{t+n} = 0.8 - (\frac{(0.397+0.544)}{5} \times (1.03 - \text{Regret}_{Alice}(\text{info\_share}))$$

We're almost there. We just need to figure out how much Alice's agent regrets what happened. Of course, we needn't, and can set this to 0, giving an adjustment of 0.19. That's okay, but we give the agent a chance to express regret and see.

As discussed above in section 7.3, it's not so easy to figure out if what we're *being told* is actually what *is*. We do have a potential formula from equation 10:

$$\text{Regret}_{Alice}(\text{info\_share}) = (U_{Alice}(\text{info\_share}) - U_{Alice}(\text{info\_share}^{-})) \bullet \text{I}_{Alice}(Alice, Steve)$$

It's possible for Steve's agent to estimate much of these values in a pinch, and these may tally with Alice's agent's estimates or they may not. This is in fact not as important as it might sound, since the trust we are re-evaluating is Steve's agent's in Alice's agent, and this is inherently personal. If Steve or his agent was to get feedback from Alice's agent *that he or his agent considered valid* this *may* make a difference, but he could just as easily choose to discard it. In this instance, because his agent knows the relationship is important, and the previous trust was high (for Steve), it makes the decision to believe what it is given. In addition, for consistency, it calculates its own $\text{Regret}_{Alice}(\text{info\_share})$ value, which we call $Steve(\text{Regret}_{Alice}(\text{info\_share}))$:

$$Steve(\text{Regret}_{Alice}(\text{info\_share})) = (0.5 - 0)times0.8 \ = \ 0.4$$

Alice's own calculations are similar, and show a regret of 0.45. Being a nice agent, Steve's agent takes this to be true. Then we can finally work out how to adjust the trust value:

$$T_{Steve}(Alice)^{t+n} = 0.8 - (\frac{(0.397+0.544)}{5} \times (1.03 - 0.45) \ = \ 0.8 - 0.109 \ = \ 0.691$$

This final value has an effect on how Steve's agent sees Alice's agent. For example, next time she asks to get some information, the original trust and cooperation threshold calculations above result in 0.47 for trust and 0.535 for the cooperation threshold (all other things, except for Alice's agent's competence being revised drastically down to 0.5 here). Clearly, it's not going to happen again. Distributing this data amongst Steve's agent's circle of friends is not complicated either.

So what happens next?

Alice's agent regrets what happened, that much is clear. In the final analysis this regret may be higher than the formulae here predict (there is a loss over and above that situation). Eventually, it's time, then, for forgiveness to be considered.

A brief aside is necessary here. At what time does one consider forgiveness? In equation 17, recall, it's our *Forgiveness Trait*. For some, it's a more or less instantaneous action[4], for others less so. Some never consider it. For our part, with no moral judgment on the matter, we have given our agents the *Forgiveness Trait* in order to allow agent owners, if this system is incorporated within them, to decide for themselves. However, there is a difference in subjective times for agents and humans – agents in this example if not others work on 'internet time,' and things happen fast there. 'Human time,' however perceived, is always slower than this. Thus if we visit the parable once more, we see that forgiveness is entered into before Steve emerges from the bush, kayak in hand (so to speak). This could be a week later or a month later. Perhaps even less time has passed. For Steve's agent, it seems a lot longer…

Steve's agent has a *Forgiveness Trait* of 75. Recall from equation 15 that regret mitigates the length of time to wait – the more regret, the longer the time. Here the agent's regret is 1.03, giving a timescale of 77.25 for the agent. That's 77.25 timesteps. For the sake of nothing other than arbitrariness, let's say that each step on that scale equates to one hour[5] of real time. Thus, in 77 hours and 15 minutes, Steve's agent is *ready to consider* forgiving Alice's agent. When that time arrives, his agent considers the *Forgiveness Function*, from equation 16:

$$\text{Fk}_{Steve} = \frac{(\text{Regret}_{Alice}(\text{info\_share}) - \text{Regret}_{Steve}(\text{info\_share}) + \text{I}_{Steve}(Steve,Alice))}{\text{Ft}_{Steve}} \times T_{Steve}(Alice)$$

Then, with our new level of trust, but still considering the relationship important:

$$\text{Fk}_{Steve} = \frac{(0.4 - 1.03) + 0.75}{75} \times 0.691 \; = \; 0.0011$$

So, from equation 17 Steve's agent can now trust Alice to a value of 0.6921. This can take some time.... Of course, we can now look at this in a couple of different ways - every timestep, since this is a positive value, we can reconsider forgiveness and increase trust, and should circumstances, such as cooperation in other endeavours, permit, increase the $T_{Steve}(Alice)$ value in that way also. Or, as with the grey areas of untrust, state that since the forgiveness function came up with a positive figure, forgiveness is assured and wipe the slate clean (returning the trust value to its previous figure of 0.80). We prefer the more gradual approach and use it in our work. Consider that if we re-evaluate every hour (time step), then after 53 hours, all other things being equal, cooperation can resume on the information sharing situation. Even with a more punishing revising of the competence of Alice's agent to 0.3, cooperation can resume within 217 hours (time steps). 5 to 12 days in real time.

Forgiveness then, in this circumstance, appears to do what it's supposed to – give time for reflection and the rebuilding of trust.

[4] Gordon Wilson, in life, and Michael Berg respectively were and are prominent among them.

[5] Yes, it could be a minute, or a second. This is one parameter an owner of such an agent should consider, amongst many others.

11.2 Regret Management

In [26] we introduced the concept of *Regret Management*, and here will take it a little further with the discussion of a simple Regret Management system.

The purpose of a Regret Management system is to ensure that transgressors in a previously trusted situation are held to account for their transgressions. There are many examples of where and how this could be achieved, from a posteriori access control [13] to blacklisting in online auction sites, or from our own example of the parable above taken through Alice's agent's eyes, through advanced Trust Management system techniques that ensure accountability.

In [26], we postulated that a Regret Management system would have the following properties (now numbered, in no particular order):

1. It is capable of assessing to some value the amount of regret a truster has after a trustee transgresses.
2. It is capable of ensuring that the transgressor is 'assigned' that regret – that is, punished in some material (meaningful to the transgressor) way in a form proportional to the truster's regret.
3. It is open and clear enough, and 'trusted' by both parties to be able to make these things happen. An 'untrusted' regret management system is as good as no system at all.

For item 1, the formalisations for regret above correctly fulfill this role – there is an assessment of regret. Item 2 depends more on the system, but consider that given the assessment, any system where communication of some form is possible would be able to assign that regret. Note for instance that in the example the regret is assigned by my agent, and further by my agent's broadcasting this to it's own close acquaintances. Given the gamut of possibilities for systems, item 3 is more difficult, but what is necessary is for the transgressor to know that the truster is not above making their own 'failings' in trusting 'badly' known to the world in some way. *All* trust-based systems, human or otherwise, will fail in the face of reluctance in this instance – embarrassment is the con-man's greatest weapon.

If we can revisit the worked example above for a moment, we note that Alice's agent is capable of expressing regret, and that my own agent is capable of making it regret its actions (by shutting it out of my community for a time, and by broadcasting that to the community). A reasonable start for a Regret Management system. However, a more formal approach is required in order to achieve our second and third requirements.

We propose that the system should:

- Calculate regret for each agent in the relationship, independently of the agents' calculations;
- (If possible) ascertain the regret calculations from each agent for corroboration – this in fact can be used also to determine the truthfulness of each agent's deliberations;
- Ensure all agents are aware of the calculations and results;

- Apportion reparations based on the results of the calculations
- Enforce reparations;

We are currently implementing such a system based on the ACORN architecture [69] with trust additions.

12 Related Work

There is a great deal of work, this volume amongst it, that deals with the phenomenon of trust in a computational setting. Approaches range from Trust Management [39, 49, 25, 42], where there is a need to determine trust or its propagation via trusted others, to trust models [64, 1, 84], which enable individual agents to model other with trust as at least a component, if not the major one, of the model. Additionally, regret is, as has been noted above, not a new phenomenon of study, particularly in the economic sciences [7, 57, 16, 35, 93]. It is nothing new to suggest that an agent can and does use regret, anticipatory or otherwise, in decision making. Forgiveness, while extensively studied in religious and philosophical fields [27, 47, 78, 12, 89, 11, 77, 17], is much less popular a field of study in computational settings. Indeed, [88] is premiere in this field, discussing the application of forgiveness in a computer-mediated communication setting.

While to our knowledge there is no work combining the three phenomena to further the development of truly social agents, the effect obtained, at least when we consider autonomous agents working in specific environments for humans, is similar to the concept of *Adjustable Autonomy*, in which humans retain a meta-level control over their more-or-less autonomous agents [28, 41, 29]. In this work, however, the control is given back to the agents in order to allow them to adjust the amount of leeway (or autonomy) other agents have with their resources, as well as the human. Moreover, there is a built in, via forgiveness, mechanism for re-attaining autonomy when adequate guarantees are given or behaviour observed.

The quest for more socially oriented agents, and the phenomena studied here are related to Danielson's concept of Artificial Morality [19], where a game theoretical approach is used to balance rationality and morality in a social setting. As well, Dautenhahn's Social Intelligence concept [20] maintains a viewpoint similar to our final goal.

13 Trust as Will Be: Future Work and Conclusions

The model of trust presented in this chapter is not in itself new, but the way in which it interacts with regret and forgiveness is. In this work, Trust, Regret and Forgiveness form an *internal triangle*. Trust allows agents tools in making decisions, regret allows them yet more, but also gives them a means to, and a measure for, adapting to current circumstances. Forgiveness allows the society (of agents, as well as humans) to continue operating in a reasonable, cooperative fashion.

While the study of the phenomena in their own right is interesting, they point the way to powerful tools in a practical sense. To that end, we have presented the design of the trust, forgiveness, regret triangle in the context of Ambient Intelligence. Our current work is focused on further refining the models, including the incorporation of a consideration of Decision Justification (cf. [16]), implementing the triangle in an information sharing architecture (ACORN [69]), an AmI interface, and Regret Management systems.

We have stated that trust is not the panacea for all technological ills. It is, however, necessary for the development of a *complete* social agent. The same goes for regret and forgiveness as computational concepts.

Regret and forgiveness are not the only phenomena with which a social agent can be equipped – other rational decision aids exist. For our part, without presuming to theorise about what *correct*, or *moral* behaviour is, we are beginning to examine the concept of *Integrity* – doing the 'right' thing for the 'right' reason. This is something of a departure from a trusting, or even a moral, agent, but integrity is in itself a decision-making strategy. Moreover, we conjecture that trusting an integrity-based agent would be rather simpler than one who is not, since we would expect the integrity-based agent to behave in accordance with its principles (which could be public) at all times.

Current work involving trust is both promising and worthwhile, but represents only a small portion of the picture that needs to be uncovered. If the systems we build are to be capable of sustained interactions in the (our) social world, there is a need for them to be able to understand, or at least represent, much more than just trust. This work is a step along that road.

References

1. Alfarez Abdul-Rahman. *A Framework for Decentralised Trust Reasoning*. PhD thesis, Department of Computer Science, University College London, 2004 (Submitted).
2. Charles Abraham and Paschal Sheeran. Deciding to exercise: The role of anticipated regret. *British Journal of Health Psychology*, 9:269–278, 2004.
3. J. Ahola. Ambient intelligence. *ERCIM News*, 47, October 2001.
4. Robert Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.
5. Annette Baier. Trust and antitrust. *Ethics*, 96(2):231–260, January 1986.
6. Bernard Barber. *Logic and Limits of Trust*. Rutgers University Press, New Jersey, 1983.
7. D. E. Bell. Regret in decision making under uncertainty. *Operations Research*, 30:961–981, 1982.
8. George David Birkhoff. A mathematical approach to ethics. In James R. Newman, editor, *The World of Mathematics, Volume 4*, pages 2198 – 2208. Simon and Schuster, New York, 1956.
9. Sissela Bok. *Lying: Moral Choice in Public and Private Life*. Pantheon Books, New York, 1978.
10. Susan D. Boon and John G. Holmes. The dynamics of interpersonal trust: resolving uncertainty in the face of risk. In Robert A. Hinde and Jo Groebel, editors, *Cooperation and Prosocial Behaviour*, pages 190–211. Cambridge University Press, 1991.

11. Lesley A. Brose, Mark S. Rye, Catherine Lutz-Zois, and Scott R. Ross. Forgiveness and personality traits. *Personality and Individual Differences*, 39:35–46, 2005.
12. Ryan P. Brown and April Phillips. Letting bygones by bygones: further evidence for the validity of the tendency to forgive scale. *Personality and Individual Differences*, 38:627–638, 2005.
13. J. G. Cederquist, R. J. Corin, M. A. C. Dekker, S. Etalle, J. I. den Hartog, and G. Lenzini. Audit-based compliance control. *International Journal of Information Security*, 6(2–3):133–151, 2007.
14. John Child and Guido Möllering. Contextual confidence and active trust development in the chinese business environment. *Organization Science*, 14(1):69–80, January–February 2003.
15. Piotr Cofta. *Trust, Complexity and Control: Confidence in a Convergent World*. Wiley, 2007.
16. Terry Connolly and CMarcel Zeelenberg. Regret in decision making. *Current Directions in Psychological Science*, 11(6):212–216, December 2002.
17. Jim Consedine. Forgiveness as public policy. *Australian EJournal of Theology*, 9, Marsh 2007.
18. Peter Danielson. *Artificial Morality: Virtuous Robots for Virtual Worlds*. Routledge, 1992.
19. Peter A. Danielson. Is Game Theory Good for Ethics?: Artificial High Fidelity. Corrected version of a paper presented at an invited symposium on Game Theory at the APA Pacific Division meeting, San Francisco, 29 March, 1991. Author is at University of British Columbia, Vancouver, Canada., 1992.
20. Kerstin Dautenhahn, Bond Alan H, Lola Canamero, and Bruce Edmonds, editors. *Socially Intelligent Agents: Creating Relationships with Computers and Robots*. Kluwer Academic Publishers, 2002.
21. Paul Davis. On apologies. *Journal of Applied Philosophy*, 19(2):169–173, 2002.
22. Morton Deutsch. Cooperation and trust: Some theoretical notes. In M. R. Jones, editor, *Nebraska Symposium on Motivation*. Nebraska University Press, 1962.
23. Morton Deutsch. *The Resolution of Conflict*. Yale University Press, New Haven and London, 1973.
24. Mark R. Dibben. *Exploring Interpersonal Trust in the Entrepreneurial Venture*. London: MacMillan, 2000.
25. Changyu Dong, GIovanni Russello, and Narakner Dulay. Trust transfer in distributed systems. In Sandro Etalle and Stephen Marsh, editors, *Trust Management: Proceedings of IFIPTM 2007*, pages 17–30, 2007.
26. Sandro Etalle, Jerry den Hartog, and Stephen Marsh. Trust and punishment. In *International Conference on Autonomic Computing and Communication Systems (Autonomics), 28-30 October 2007*. ACM Press, October 28–30 2007.
27. Julie Juola Exline, Everett L. Worthington Jr., Peter Hill, and Michael E. McCullogh. Forgiveness and justice: A research agenda for social and personality psychology. *Personality and Social Psychology Review*, 7(4):337–348, 2003.
28. Rino Falcone and Cristiano Castelfranchi. Levels of delegation and levels of adoption as the basis for adjustable autonomy. In Evelina Lamma and Paolo Mello, editors, *AI*IA 99: Advances in Artificial Intelligence: 6th Congress of the Italian Association for Artificial Intelligence, Bologna, Italy, September 1999. Selected Papers*, volume LNAI 1792 of *LNAI 1792*, pages 273–284. Springer, 2000.
29. Rino Falcone and Cristiano Castelfranchi. The human in the loop of a delegated agent: The theory of adjustable social autonomy. *IEEE Transactions on Systems, Man and Cybernetics*, 31(5):406–418, 2001.
30. Rino Falcone and Cristiano Castelfranchi. The socio-cognitive dynamics of trust: Does trust create trust? In R. Falcone, M. Singh, and Y.-H. Tan, editors, *Trust in Cyber-Societies*, volume 2246 of *Lecture Notes in Artificial Intelligence*. Springer Verlag, 2001.
31. Leon Festinger. *A Theory of Cognitive Dissonance*. Stanford University Press, CA, 1957.
32. Batya Friedman, Peter H. Kahn, and Daniel C. Howe. Trust online. *Communications of the ACM*, 43(12):34–40, 2000.
33. Francis Fukuyama. *Trust: The Social Virtues and the Creation of Prosperity*. Simon and Schuster, 1996.

34. Diego Gambetta. Can we trust trust? In Diego Gambetta, editor, *Trust*, chapter 13, pages 213–237. Blackwell, 1990.
35. Daniel T. Gilbert, Carey K. Morewedge, Jane L. Risen, and Timothy D. Wilson. Looking forward to looking backward: The misprediction of regret. *Psychlogical Science*, 15(5):346–350, 2004.
36. Thomas Gilovich and Victoria Husted Medvec. The temporal pattern to the experience of regret. *Journal of Personality and Social Psychology*, 6(3):357–365, 1994.
37. H. C. J. Godfray. The evolution of forgiveness. *Nature*, 355:206–207, 16th January 1992.
38. Jennifer Golbeck. *Computing and Applying Trust in Web-based Social Networks*. PhD thesis, University of Maryland, College Park, 2005.
39. Jennifer Golbeck. *Computing and Applying Trust in Web-based Social Networks*. PhD thesis, University of Maryland at College Park, 2005.
40. Robert T. Golembiewski and Mark McConkie. The centrality of interpersonal trust in group processes. In Cary L. Cooper, editor, *Theories of Group Processes*, chapter 7, pages 131–185. Wiley, 1975.
41. Michael A. Goodrich, Dan R. Olsen Jr., Jacob W. Crandall, and Thomas J. Palmer. Experiments in adjustable autonomy. In *IJCAI-01 Workshop on Autonomy, Delegation, and Control: Interaction with Autonomous Agents*, 2001.
42. Andreas Gutscher. A trust model for an open, decentralized reputation system. In Sandro Etalle and Stephen Marsh, editors, *Trust Management: Proceedings of IFIPTM 2007*, pages 285–300, 2007.
43. Diane F. Halpern. *Thought and Knowledge: An Introduction to Critical Thinking*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1984.
44. Nicolai Hartmann. *Ethics II — Moral Values*. George Allen and Unwin, London, 1932.
45. Lars Hertzberg. On the attitude of trust. *Inquiry*, 31(3):307–322, September 1988.
46. Natsuko Hikage, Yuko Murayama, and Carl Hauser. Exploratory survey on an evaluation model for a sense of security. In H. Venter, M. Eloff, L. Labuschagne, J. Eloff, and R. von Solms, editors, *IFiP Internationa! Federation for Information Processing, Volume 232, New Approaches for Security, Privacy and Trust in Complex Environments*, volume 232, pages 121–132. Springer, 2007.
47. Robert Jeffress. *When Forgiveness Doesn't Make Sense*. Waterbrook Press, 2001.
48. Audun Josang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.
49. Audun Josang, Stephen Marsh, and Simon Pope. Exploring different types of trust propagation. In Ketil Stolen, WIlliam Winsborough, Fabio Martinelli, and Fabio Massacci, editors, *Trust Management: Proceedings of the 4th International Conference on Trust Management (iTrust'06), 2006*, volume 3986 of *Springer Lecture Noted in Computer Science*, pages 197–192, 2006.
50. Daniel Krahmer and Rebecca Stone. Regret in dynamic decision problems. Technical Report 71, GESY - Governance and the Efficiency of Economic Systems Discussion Paper 71, www.gesy.uni-mannheim.de, July 2005.
51. Roderick M Kramer. Trust rules for trust dilemmas: How decision makers think and act in the shadow of doubt. In R. Falcone, M. Singh, and Y.-H. Tan, editors, *Trust in Cyber Societies*, pages 9–26. Springer Verlag, Lecture Notes in Artificial Intelligence, LNAI 2246, 2001.
52. Karl Krukow. *Towards a Theory of Trust for the Global Ubiquitous Computer*. PhD thesis, University of Aarhus, 2006.
53. Olli Lagenspetz. Legitimacy and trust. *Philosophical Investigations*, 15(1):1–21, January 1992.
54. R. J. Lewicki and B. B. Bunker. Trust in relationships: A model of trust, development and decline. In B. B. Bunker and J. Z. Rubin, editors, *Conflict, Cooperation and Justice*, pages 133–173. San Francisco: Josey Bass, 1985.
55. R. J. Lewicki, D. J. McAllister, and R. J. Bies. Trust and distrust: New relationships and realities. *The Academy of Management Review*, 23(3):438–458, 1998.
56. R. J. Lewicki, D. J. Bies McAllister, and R. J. Bies. Trust and distrust: New relationships and realities. *Academy of Management Review*, 23:438–458, 1998.

57. LG. Loomes and R. Sugden. Regret theory: An alternative theory of rational choice under uncertainty. *Economic Journal*, 92:805–824, 1982.
58. Steve Love, Pamela Briggs, Linda Little, Stephen Marsh, and Lynne Coventry. Ambient intelligence: Does public mean private? In *Proceedings of HCI 2005: The Bigger Picture. Edinburgh, 5–9 September*, 2005.
59. M. Low and V. Srivatsan. What does it mean to trust an entrepreneur? In S. Birley and I. C. MacMillan, editors, *International Entrepreneurship*, pages 59–78. Routledge, London, 1995.
60. R. Duncan Luce and Howard Raiffa. *Games and Decisions*. Dover Publications, 1957.
61. Niklas Luhmann. *Trust and Power*. Wiley, Chichester, 1979.
62. Niklas Luhmann. Familiarity, confidence, trust: Problems and alternatives. In Diego Gambetta, editor, *Trust*, chapter 6, pages 94–107. Blackwell, 1990.
63. Stephen Marsh. Trust and reliance in multi-agent systems: A preliminary report. In *MAAMAW'92, 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Rome*, 1992.
64. Stephen Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, Department of Computing Science, University of Stirling, 1994. Available online via http://www.stephenmarsh.ca/Files/pubs/Trust-thesis.pdf.
65. Stephen Marsh. Optimism and pessimism in trust. In J Ramirez, editor, *Proceedings Iberoamerican Conference on Artificial Intelligence/National Conference on Artificial Intelligence (IBERAMIA94/CNAISE94)*. McGraw-Hill, October 1994.
66. Stephen Marsh. Trust in Distributed Artificial Intelligence. In Cristiano Castelfranchi and Eric Werner, editors, *Artificial Social Systems*, pages 94–112. Springer Verlag, Lecture Notes in AI, Vol. 830, September 1994.
67. Stephen Marsh. Trust, regret, forgiveness, and boing. Seminar, University of St Andrews, Scotland, UK, October 2005.
68. Stephen Marsh and Mark R. Dibben. Trust, untrust, distrust and mistrust — an exploration of the dark(er) side. In Peter Herrmann, Valerie Issarny, and Simon Shiu, editors, *Trust Management: Proceedings of iTrust 2005*. Springer Verlag, Lecture Notes in Computer Science, LNCS 3477, 2005.
69. Stephen Marsh, Ali A. Ghorbani, and Virendra C. Bhavsar. The ACORN Multi-Agent System. *Web Intelligence and Agent Systems*, 1(1):65–86, March 2003.
70. D. Harrison McKnight and Norman L. Chervany. The meanings of trust. Working paper, MISRC, 1996.
71. D. Harrison McKnight and Norman L. Chervany. Trust and distrust definitions: One bite at a time. In R. Falcone, M. Singh, and Y.-H. Tan, editors, *Trust in Cyber-Societies*, volume 2246 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, Berlin, Heidelberg, 2001.
72. D. Harrison McKnight, Chuck Kacmar, and Vivek Choudhury. Whoops... Did I use the Wrong concept to Predict E-Commerce Trust? Modeling the Risk-Related Effects of Trust versus Distrust Concepts. In *36th Hawaii International Conference on Systems Sciences*, 2003.
73. Barbara A. Mistal. *Trust in Modern Societies*. Oxford: Blackwell, 1996.
74. Guidio Möllering. *Trust: Reason, Routing, Reflexivity*. Elsevier Science, 2006.
75. Guido Möllering. The nature of trust: From georg simmel to a theory of expectation, interpretation and suspension. *Sociology*, 35(2):403–420, 2001.
76. Lik Mui. *Computational Models of Trust and Reputation: Agents, Evolutionary Games, and Social Networks*. PhD thesis, Massechusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2002.
77. J. G. Murphy. Forgiveness, mercy and the retributive emotions. *Criminal Justice Ethics*, 7:3–15, 1988.
78. J. G. Murphy. Forgiveness in counseling: A philosophical perspective. In S. Lamb and J. G. Murphy, editors, *Before forgiving: cautionary views of forgiveness in psychotherapy*. Oxford University Press, 2002.
79. B. Noteboom, H. Berger, and N. Noordehaven. Effects of trust and governance on relational risk. *Academy of Management Journal*, 40(2):308–338, 1997.
80. Peter A. Pang. Experiments in the evolution of cooperation, masters thesis, university of stirling, department of computing science, 1990.

81. Scott Plous. *The psychology of judgment and decision making*. McGraw-Hill, New York, 1993.
82. Paul Resnick, Richard Zeckhauser, Eric Friedman, and Ko Kuwabara. Reputation systems. *Communications of the ACM*, 43(12):45–48, 2000.
83. Neal J. Roese and James M. Olson, editors. *What Might Have Been: The Social Psychology of Counterfactual Thinking*. Lawrence Erlbaum Associates, Mahwah, NJ, 1995.
84. Jean-Marc Seigneur. *Trust, Security and Privacy in Global Computing*. PhD thesis, Trinity College, Dublin, 2005.
85. Jean-Marc Seigneur and Christian Damsgaard Jensen. The role of identity in pervasive computational trust. In Philip Robinson, Harald Vogt, and Waleed Wagealla, editors, *Privacy, Security and Trust within the Context of Pervasive Computing*, volume 780 of *Kluwer International Series in Engineering and Computer Science*. Kluwer, 2005.
86. P. Sztompka. *Trust: a Sociological Theory*. Cambridge University Press, 1999.
87. Piotr Sztompka. *Trust: A Sociological Theory*. Cambridge University Press, 2000.
88. Asmina Vasalou and Jeremy Pitt. Reinventing forgiveness: A formal investigation of moral facilitation. In Peter Herrmann, Valèrie Issarny, and Simon Shiu, editors, *Trust Management: Third Internation Conference, iTrust 2005, Proceedings*, volume 3477 of *Lecture Notes in Computer Science*. Springer, Berlin Heidelberg, 2005.
89. D. F. Walker and R. L. Gorsuch. Forgiveness awithin the big five personality model. *Personality and Individual Differences*, 32:1127–1137, 2002.
90. M. Weiser. The computer for the 21st century. *Scientific American*, 265(3):66–75, September 1991.
91. D.J. Wu, Steven O. Kimbrough, and Fang Zhong. Artificial agents play the “mad mex trust game”: A computational approach. In *Proceedings of the 35th Hawaii International Conference on System Sciences*, 2002.
92. Yutaka Yamamoto. A Morality Based on Trust: Some Reflections on Japanese Morality. *Philosophy East and West*, XL(4):451–469, October 1990.
93. Marcel Zeelenberg and Rik Pieters. A theory of regret regulation 1.0. *Journal of Consumer Psychology*, 17(1):3–18, 2007.