A Psychological Approach to Understanding how Trust is Built and Lost in the Context of Risk

J. Richard Eiser & Mathew P. White
University of Sheffield, UK

Paper to be presented at SCARR conference on Trust, LSE, 12th December 2005

Abstract

Over 50 years of cross-disciplinary research suggests we have higher trust in those who are seen as a) knowledgeable, b) similar to ourselves and c) honest/transparent. Doctors are trusted with our medical care because they are perceived to: be competent, share our concerns, and tell us what they think. Politicians tend to be distrusted for the opposite reasons. However, we know much less about 'marginal trust', i.e. how trust is built or lost as a result of new information. Many policy makers already know that public trust in them is low. What they now want to know is how to improve this situation. Greater understanding of marginal trust is critical for such change. In its exploration of marginal trust, the current paper draws on four fundamental psychological processes and related empirical evidence. First, in line with a general *negativity bias*, people tend to trust bad news more than good. Second, in line with the desire for *cognitive consistency*, people tend to trust news that is congruent with their prior attitudes more than news that is incongruent. Third, in line with research into *information diagnosticity*, news with greater breadth (policy-related) has a bigger effect on trust than news with less breadth (event-related). Fourth, by extending a classification of performance types developed from psychophysical experiments (*Signal Detection Theory*) to issues of trust, we identify a type of error that actually increases trust. Taken together, our research provides a more optimistic view on how to improve marginal rates of trust than is commonly believed.

Key words: Trust, risk perception, negativity bias, confirmatory bias, information diagnosticity, signal detection theory.

1.0 Introduction

Imagine it's a Sunday afternoon and you feel like going for a walk. You look out of the window and see patches of blue sky amongst the odd bit of cloud and a moderate Westerly breeze. You ask yourself "should I perhaps take an umbrella?". Now, if the Sunday afternoon in question is in August, we suspect you will conclude "no I don't need one, it hardly rains in August when there's some blue sky" and you'll go for the walk without further ado. However, if the afternoon is in April, we suspect you may be a little more hesitant. You know from experience that in April the showers can come in from the West very quickly, and now you're not so sure its going to stay dry. How do you cope with

this greater level of uncertainty? Well one approach is to turn on the radio and listen to the weather forecast. Fortunately for you the forecaster says there's only a 5% chance of rain today so you set off without the umbrella and end up enjoying a lovely stroll through the countryside. Although mundane, this little scenario encapsulates many of the same processes that are involved in more serious examples of risk perception and trust in decision makers.

- First, and most importantly, we are dealing with a decision made under conditions of *uncertainty.* Although risk is commonly defined as the *probability* of an unwanted event, thinking in terms of uncertainty is more instructive. All probabilities between the extremes where $p = 0$ or 1 reflect some degree of uncertainty. Numerical expressions of probability are attempts to quantify such uncertainty. Indeed, the level of uncertainty can reflect external regularities, such as seasonal variations in rainfall. Nonetheless, the apparent precision of such numerical expressions can disguise the fact that (in all natural or open systems) they are *estimates*, the reliability of which depends on a number of factors. These include the extent of randomness in the actual events considered, the quality of the data available about these events, and the processes and assumptions used to derive estimates from data. In other words, such estimates are not merely expressions of uncertainty, but uncertain in themselves. Moreover, when probability estimates of separate events are combined (e.g. multiplied) – as in fault tree analysis or other forms of probabilistic risk assessment applied to risks of complex events or system failures – the unreliability of such estimates can propagate and be magnified throughout a whole series of calculations. Uncertainty thus can be attached both to events and to estimates of such events. Beyond this, though, uncertainty is a state of mind. What matters for the decisions we make is not simply the numbers we are offered, but how we feel about them.
- Second, the risk arises not merely from the natural 'hazard' of rainfall, but from the interaction between that hazard and the decisions we make. None of this is to minimize the potential danger of more extreme climatic or geophysical events, such as floods, earthquakes, tsunamis and hurricanes. But what most influences the consequences of such events – what mitigates or exacerbates the worst effects – is human decision-making at all levels from the individual through the institutional to the international. To a large extent, therefore, risk perception involves an implicit or explicit evaluation of such decision-making.
- Third, the need to make any kind of decision at all only arises because our actions may lead to more or less preferred *consequences*. Our scenario rests on the implicit norms that a walk in good weather is pleasant, getting soaked in the rain is unpleasant, staying indoors is boring and carrying an umbrella is a bit of a nuisance. So the decision involves not just estimating what is likely to happen, but a trade-off of likely costs and benefits.
- Fourth, although we can make decisions by ourselves, we can also seek advice from other people. In many instances we make comparative judgements about our own capacity and that

of others to make good decisions. As uncertainty increases (i.e. a walk in April) our propensity to turn to the judgements of others increases (Siegrist & Cvetkovich, 2000).

- Fifth, in such instances it is not just anyone's opinion we seek, but that of someone who we believe a) knows more about the situation than ourselves and b) is motivated to tell us what they believe to be true (i.e. the weather forecaster). In this way, our own belief about the likelihood of rain in April (our risk perception) will depend on whether or not we feel the weather forecaster is willing and able to make an accurate assessment of the likelihood of rain. That is, risk perception becomes based on a *social judgement* about the trustworthiness of the risk communicator. If our *baseline trust* (i.e. our general level of trust) in weather forecasters is high we probably won't take the umbrella, but if our baseline level of trust is low we might take one anyway.
- Sixth, the kind of trust we are talking about here is not the kind of trust we have in friends and family (i.e. interpersonal trust, e.g. Rempel Holmes & Zanna, 1985) or in other people in general (i.e. social trust, Helliwell & Putnam, 2004), but trust in specific individuals whose role it is to assess, manage and communicate information about risk. Such trust has been referred to as "role-based trust" since "it is not the person in the role that is trusted so much as the system of expertise that produces and maintains role-appropriate behaviour of role occupants" (Kramer, 1999, p.578).
- Seventh, although turning to more knowledgeable others regarding the risk of rain is an example of a natural hazard, we believe the same processes apply to human made hazards (e.g. nuclear power, vaccinations etc.) and even to other humans as potential hazards (e.g. Is this leader planning military action?; is this online trader genuine or bogus?; Is this person a suicide bomber?). Although our main focus in the current paper is on trust and human-made hazards, we will return to the issue of human hazards later when discussing on-going research. Our basic model can be seen in Figure 1. The arrow from the self to the hazard reflects 'risk perception' - that is the person's judgment of the risks from the hazard. The arrow from the experts to the hazard reflects the, usually, more formal 'risk assessment' of the same hazard by people who have greater knowledge than the self (e.g. weather forecasters, seismologists, nuclear engineers, regulators, consumer advisers, security forces and so on). The arrow from the experts to the self reflects 'risk communication', that is the processes whereby the experts report (more or less transparently) their risk assessment. The last arrow, from the self to the experts, reflects the kind of 'role-based trust' we are primarily interested in. Clearly, the model is closely associated with Heider's (1958) Balance Theory and Hardin's (2001) perspective on trust as a 'three-part relation' (see Section 3.2).

Insert Figure 1 about here

- Finally, returning to our example of a Sunday stroll, we also suspect that how you respond to the forecast will in part depend upon whether or not the last forecast you heard was correct or

not. For example, if the last forecast correctly predicted rain even though it looked unlikely, then we believe you are more likely to trust the forecaster now than if the last forecast said it would be clear but it did in fact rain. That is, our baseline level of trust can be influenced by the accuracy of specific previous judgements. Such changes reflect the *marginal rate of trust* as a function of specific judgements. One of our central arguments is that the marginal rate of trust will depend not only on the accuracy of previous decisions but also on the type of correct or incorrect decision that was made.

To date, previous research into role-based trust, has focused predominantly on *baseline trust*. Specifically, most researchers have tried to identify the correlates of such trust and this research has provided us with a relatively good picture of the kind of people who are generally (dis)trusted and why this is the case. We briefly review some of this literature in Section 2 below. Far less attention has been paid to marginal rates of trust. Specifically, there is relatively little research addressing questions such as: "how do single instances of 'good' or 'bad' performance affect trust?"; "How is it that some mistakes undermine our confidence in doctors more than others?"; "What kinds of performance should policy makers endeavour to perform if they want the public to trust them more?". In section 3 we discuss four psychological processes that we believe will help to shed some light on these and related questions.

Section 3.1 discusses what we could call the standard line on marginal trust. That is, due to an inherent negativity bias in human psychology, bad news about someone else has a stronger impact on our trust in them than good news. Since negative information is thought, from this perspective, to be stronger than positive, the view on marginal trust is rather negative, it is said to be easier to lose than gain. Section 3.2 examines the role of an alternative, and perhaps equally strong psychological process, namely the desire for cognitive consistency. From this perspective, new information that is congruent with prior attitudes will tend to be accepted more than incongruent information. In the latter case, other processes tend to kick in that help to discount the new information and thus it loses some of its power in terms of marginal trust. Section 3.3 looks more closely at the specificity of the new information. In keeping with the basic principle that more information has greater diagnostic power than less, we report on new research showing that information about implemented policies (good and bad), which cover a broader range of behavioural instances, has a greater impact on trust than information about specific behavioural instances or events (good or bad). Moreover, as we shall see, this factor seems to interact with message valence.

Section 3.4 discusses a new approach to examining marginal trust by arguing that the assessments an expert can make with regard to any given hazard under conditions of uncertainty will result in the same taxonomy of outcomes made by participants in psychophysical tasks where they need to say

whether a signal is present or not. Marginal trust in experts following these well known outcomes, such as Hits, Misses and False Alarms, is then examined to see whether there are differences in the extent to which different types of good and bad news affects trust. Again, the results are somewhat intriguing and when combined with the findings presented in Sections 3.2 and 3.3 paint a picture of marginal trust that is rather more positive than that painted if we rely on the negativity bias findings presented in Section 3.1 alone. Some concluding comments are presented in Section 4.

2.0 Baseline trust

The two key questions with respect to baseline trust are "who is (dis)trusted?" and "why?". Answers to the first question have generally been gathered using public surveys and interviews which simply ask people how much they trust various actors with respect to a given hazard or hazards. We look at the basic outcome of such research in Section 2.1. To address the second question these surveys also tend to ask people what they think about the same actors on a range of other dimensions such as competence, honesty, values and so forth. When significant positive correlations are found between these dimensions and baseline trust (i.e. more of x is associated with higher trust) the relevant dimensions are interpreted as antecedents or dimension of trust. Putting aside the possibility that the causal direction may also work in the opposite direction (for more on this issue see Eiser, Frewer & Miles, 2002; Poortinga & Pidgeon, 2004), Section 2.2 reviews some of the antecedents that have been proposed and argues that there is a certain order to the apparent chaos.

2.1 Who is trusted and who is distrusted with regard to risk-related information?

In terms of baseline, or general levels, of role-based trust in various actors involved in assessing, managing or communicating risk, the picture is relatively clear and familiar. Our own data on trust in different actors is largely consistent with research carried out in other safety contexts ranging from food (Frewer, Howard, Hedderley, & Shepherd, 1996) to bathing water (Langford, Marris, & O'Riordan, 1999). Specifically, we carried out two postal surveys of members of the public enquiring about their risk perceptions associated with the potential risks from mobile phone technology (White & Eiser, in press) and the development of 'Brownfield' sites. For present purposes, the most relevant questions concerned the level of baseline trust people had in the various actors (from 0 'no trust' to 6 'complete trust'). The basic pattern can be seen in Figure 2.

****Insert Figure 2 about here****

In line with many other surveys, scientists and medics tend to record high levels of trust while business interests and government (especially national government) tend to be less trusted. Perhaps the most surprising finding - although also consistent with prior research - is the high level of trust in family and friends. This is especially interesting given that in our mobile phone survey, people reported knowing more, on average, about the risks than their peers. Why should people be prepared to listen to peers who they think know less? We suspect the answer to this apparent paradox lies in the

possibility that people are effectively answering two different questions. For all the other actors they are considering 'role-based trust' as planned since there are no interpersonal relations. However when we ask them about friends and relatives they are considering 'interpersonal trust'. Although we ask about their trust in this category in terms of 'information about a certain risk', we believe in classic heuristic fashion (Kahneman & Frederick, 2002) they actually answer a slightly different question which is something like "do you trust your friends in general?". If nothing else this finding serves to remind us that the questions we ask aren't always interpreted in the expected way.

2.2 Correlates of baseline trust

The literature on the correlates of role-based trust is considerable and growing. These correlates include: care, competence, concern, consensual (or shared) values, consistency, expertise, fairness, faith, honesty, knowledge, objectivity, openness, past performance, predictability, reliability and sympathy (e.g. Kasperson, Golding and Tuler, 1992; Maeda and Miyahara, 2003; Renn and Levine, 1991; Siegrist, Earle and Gutscher, 2003). Although the list of possible dimensions seems daunting, a number of factor analytic studies (e.g. Jungermann, Pfister & Fischer, 1996; Frewer, et al., 1996; Mishra, 1999; Poortinga & Pidgeon, 2003) and a conceptual review (Johnson, 1999) suggest that the list can be reduced to two or three key dimensions. These three dimensions are well summed up by Peters, Covello and McCallum's (1997) distinction between a) knowledge and expertise, b) care and concern, and c) openness and honesty. A very similar tri-partite distinction composed of ability, benevolence and integrity was proposed by Mayer, Davis and Schoorman (1995) in relation to trust more generally.

In short, we tend to trust actors, at least in the context of risk, who we believe know what they are talking about, care about public safety, and are open and transparent about their operations. While actors such as scientists and doctors tend to score quite highly on these dimensions, politicians and industry tend to score much lower, especially on the care and concern and honesty dimensions. However, as we argued at the very beginning of this paper, we suspect that most policy makers already have a pretty good idea of these things. This is not surprising since the basic ideas were already identified by Yale psychologists studying persuasion processes in the 1950s (e.g. Hovland, Janis & Kelly, 1953). What is far less obvious, and has been far less researched, is the impact of new pieces of information. Since it is generally only through new information that trust levels will change what is needed is greater understanding of the processes underlying such shifts in trust. We refer to changes in trust as a function of new information as changes in 'marginal trust' for short. The remainder of the paper is concerned with this issue of marginal trust and explores four psychological processes that might shed some light on how trust is gained, maintained and lost.

3.0 Marginal trust: Specific events

In this section we explore the role of four fundamental psychological processes and the potential role they might play in marginal trust changes. As noted above, these relate to a) a general *negativity bias*, b) the desire for *cognitive consistency*, c) breadth of information and *diagnosticity*, and d) outcome types from *judgments under uncertainty*.

3.1 Negativity bias "Bad is stronger than good"

There is widespread reference in the trust and risk perception literatures to the notion that while trust is generally hard to establish, it is relatively easy to lose and that once lost it will take a long time (if ever) to become re-established (Barber, 1983; Burt & Knez, 1996; Dasgupta, 1988/2000; Levi, 1998; Rempel, Holmes & Zanna, 1985; Rothbart & Park, 1986). Slovic (1993) coined the term 'Trust asymmetry' and described it thus: "trust is fragile. It is typically created rather slowly, but it can be destroyed in an instant by a single mishap or mistake" (p.677). Moreover, the observation that favourable traits (e.g. honest) require more behavioural instances for confirmation than unfavourable traits (e.g. dishonest) and *vice versa* for disconfirmation suggests that "favorable traits are hard to acquire but easy to lose, while unfavourable traits are easy to acquire and hard to lose" (p.137; Rothbart & Park, 1986). Such a perspective on trust is also consistent with lay perspectives as exemplified by the saying 'Trust comes on foot but leaves on horseback' (Calman, 2002) and by organizations' concern about losing their 'reputation' (e.g. Meyerson, Weick & Kramer, 1996). Slovic and colleagues (Slovic, Flynn, Johnson & Mertz, 1993, cited Slovic, 1993) were amongst the first empirically to test the trust asymmetry hypothesis and develop a theoretical account of it. They presented participants with 45 positive or negative statements concerning the management of a nuclear power plant. Positive statements suggested that the plant was well run (e.g. 'Good records are kept of plant operations, fuel shipments etc.'); negative statements that the plant was not well run (e.g. 'Record keeping in the plant regarding plant operations, fuel shipments etc. is found to be poor'). Using trust in the management of the plant as the dependent variable they found that negative statements decreased trust more than positive statements increased it. Although Slovic (1993) only reported the proportion of respondents using the most extreme category responses in his original paper, our re-analysis (White & Eiser, 2005, Study 1) suggests that this asymmetry still exists when all responses are taken into account and overall means are compared (negative items M = -4.73; positive items M = 3.07; $F(1, 102) = 82.64$, $p < 0.001$, ${}_p\eta^2 = .45$). This asymmetry can be seen in Figure 3 where the trust decreasing items tend to decrease trust more than the trust increasing items increase it.

****Insert Figure 3 about here****

At the heart of Slovic's account of these findings is the notion that people pay more attention to and are more influenced by negative than positive information. This proposition, i.e. the primacy of negative information or a 'negativity bias', has received substantial support in the broader psychological literature (for reviews see, Baumeister, et al., 2001; Rozin & Royzman, 2001). Evidence

comes from fields as diverse as attention (Pratto & John, 1991), learning (Seligman, 1970), social judgements (Peeters & Capinski, 1990; Taylor, 1991) attributions (Kanouse & Hanson, 1972) and decision-making under conditions of uncertainty (Kahneman, & Tversky, 1979). This account of marginal trust also received further support in the domain of technological risk in a study Siegrist and Cvetkovich (2001, Study 1). These authors presented people with hypothetical information suggesting that a food colourant was either safe or harmful for health. They found that on average people were more likely to trust the harmful than the safe messages. Moreover, as regards trust in public institutions, it may be that the media focus undue attention on negative events and stories such that the public is exposed to more potentially trust decreasing information than trust increasing information (e.g. Koren & Klein, 1991; although see Freudenburg et al., 1996).

In short, according to the negativity bias account, trust is easier to lose than gain because negative information is more attention grabbing, more powerful and often more readily available than positive information. Moreover, there is also evidence that people who are low in dispositional trust tend to avoid others and thus limit the number of opportunities they have for acquiring new positive information about others (Yamagashi, 2001). In other words, in line with Slovic's rather pessimistic trust asymmetry hypothesis, once trust is lost, it might be very hard to re-establish. This picture certainly seems resonant with those who express concern that social capital and well-being are being undermined by a reduction of trust between people in modern societies (e.g. Helliwell & Putnam, 2004; Layard, 2005; Uslaner, 2003).

While we do not dispute the idea of negativity dominance in general, we suspect that it is *not the only psychological process at work*. If this were the case, and people really had a "one strike and you're out" heuristic surely trust would become extinct very quickly. Yet there is also evidence that after a relatively steep downturn in trust in public institutions from the 1950s to the 1970s, things have leveled out somewhat in recent years and even increased in some institutions over time (Kasperson, Kasperson & Golding, 1999). Moreover, there is also research suggesting that trust can be built rather quickly, for example in highly motivated, taskoriented groups (Meyerson, Weick & Kramer, 1996) and that rather than turning to distrust, trust may instead turn to scepticism, a more adaptive stance in complex societies where cooperation is imperative but a lack of vigilance might result in exploitation (O'Neill, 2002). Finally, looking back at Slovic's own evidence for a negativity bias in Figure 3, it is also clear that while the average impact for negative information is stronger than that for positive information, there is considerable variance as a function of valence. That is, there are some positive events which seem to increase trust considerably and even some negative ones that have relatively little impact. In other words, the negativity bias explanation is only one part of a more complex story. In the next three sections we investigate other psychological processes that seem to be involved and as we shall see they all offer a more optimistic perspective than a pure negativity bias account.

3.2 Cognitive consistency

Like negativity dominance, cognitive consistency also appears to be a fundamental psychological process demonstrated in psychological research for over 50 years (e.g. Festinger, 1957; Heider, 1946; 1958). Put simply, the theory of cognitive consistency argues that people are motivated to maintain consistency between their attitudes, beliefs, values and so on. New information that is highly similar to current beliefs will be readily assimilated and accepted. Information that is deeply inconsistent with prior beliefs will tend to be rejected and in some cases explanations will be sought to 'explain away' the new information. To accept such information at face value would result in the need to re-organise one's views of the world and such accommodation takes considerable mental effort, something which the cognitively bounded thinker is generally reluctant to do (Simon, 1957). These processes effectively lead to a *confirmatory bias*. That is, we tend to believe new information that appears to confirm our prior beliefs more readily than information that appears to disconfirm them. For example, if a highly trusted source is quoted as saying something that conflicts with our prior attitudes about them, people may simply, “deny that the source actually was responsible for the communication or may re-interpret the “real” meaning they believe the message to have” (Hovland, Janis & Kelley, 1953, p.43). In other words, negative information about a trusted source need not necessarily result in the kinds of "catastrophic" drops in trust that have been claimed (e.g. Burt & Kenz, 1986). Instead, the negative information may be dismissed as unreliable, a misinterpretation or an exception to the rule.

Returning to the issue of role-based trust and technological risk, there is now considerable evidence that such cognitive consistency processes are at work. For example, research carried out by one of us in the 1980's demonstrates these processes for nuclear power. As chance would have it, we sent out a survey about attitudes towards nuclear power shortly before the Chernobyl accident (Eiser, Spears & Webley, 1989). Seeing the opportunity to examine how attitudes might have changed following the incident we contacted our initial respondents a second time. As might be expected, we witnessed an overall shift against nuclear power following the accident. This was reflected in more opposition to an existing nuclear plant (approx. 16 km away) and opposition to any new nuclear plant, either locally or nationally. However, the size of the shift, though reliable, was modest, and individuals’ attitudes after the accident remained highly predictable from their attitudes before. On the same topic, Cvetkovich, Siegrist, Murray & Traegasser (2004) found that the power of Slovic's (1993) negative and positive items (see 3.1 above) was moderated by prior levels of trust in risk managers in the nuclear industry. Specifically, positive information about nuclear power plant managers led to greater trust among those with high levels of prior trust while bad news was judged as more informative by low prior trustors. Indeed, this tendency for initial trust or distrust to “color our interpretation of events, thus reinforcing our prior beliefs” was already recognized by Slovic (1993).

Another aspect of this issue that we have recently investigated was whether the negativity bias witnessed in Slovic's research is to some extent due to strongly negative attitudes towards nuclear power. That is, perhaps the greater effect of negative information in that study was due to the fact that as well as being negative it was also congruent with prior attitudes - i.e. that nuclear power is bad and dangerous. Building on an unpublished aspect of the Cvetkovich et al. study (2004), we constructed 12 events (6 positive and 6 negative) and suggested to participants that these events occurred *either* in the nuclear power industry (as in the original study) *or* in the pharmaceutical industry. In line with predictions this latter industry was viewed less negatively than nuclear power and the results on trust in the two industries of the negative and positive pieces of new information can be seen in Figure 4. (N.B. for ease of comparison, the absolute effects on trust are shown, negative events naturally decreased trust while positive ones increased it). While the negativity bias was re-confirmed for positive vs negative information in the nuclear industry (left side of Figure 4), the reverse was found for the pharmaceutical industry (right side). In this instance, positive information tended to increase trust more than negative information decreased it.

****Insert Figure 4 about here****

A similar pattern of results was found when we unpacked Siegrist and Cvetkovich's (2001, see 3.1 above) demonstration of negativity bias for trust in messages about food additives (White, Pahl, Buehner & Haye, 2003). For example, we found that the effect of negative information about food colourants on trust was greater than positive information only when participants had negative prior attitudes (Study 1). Moreover, when presented with positive and negative information about a more positively viewed food additive (i.e. vitamins, Study 2), positive information tended to have a large effect than negative. In other words, the negativity bias found in earlier work appears to have been, at least in part, due to the selection of hazards about which the majority of participants had prior negative attitudes. Indeed, this tendency for our initial attitudes to “color our interpretation of events, thus reinforcing our prior beliefs” was already recognized by Slovic (1993) and clearly extends to issues of marginal trust. In short, in many instances it seems that rates of marginal trust may actually be more sensitive to good rather than bad news and thus trust may not be as easily lost as a purely negativity bias account would suggest. Why? Because people are motivated to maintain their prior attitude structures and as such they are relatively reluctant to accept information, even trust-related information, that would upset the balance of these structures.

3.3 Information diagnosticity

A third psychological mechanism we believe is important for understanding marginal rates of trust is the breadth of the new information being received. For instance is the new information related to a single behavioural instance (e.g. last week the police accidentally shot an innocent suspect) or is it indicative of performance over repeated instances (e.g. last week the police adopted a more lenient policy towards shooting suspected criminals)? According to many researchers, impression formation

(including trustworthiness judgements) is essentially a categorisation process and people use new information to help them *diagnose* which category a person falls into. So, for example, “someone who once stole money probably belongs in the dishonest category, but there is a real, though lesser probability, that they instead belong in the honest category” (Skowronski & Carlston, 1987, p.689). Importantly, and not surprisingly, more information is better than less for these diagnostic purposes (Rothbart & Park, 1986).

In terms of marginal trust, we suspected that people would therefore be prepared to make more extreme trust-related categorization judgements (i.e. this actor is highly (un)trustworthy) when the information is more rather than less diagnostic in terms of breadth. We tested this idea by once again returning to Slovic's (1993) original data. Using a within-subject design, Slovic and colleagues presented participants with 45 hypothetical events that could have taken place in a nuclear power plant and were asked whether or not their trust in the management of these plants would increase or decrease and by how much. Recall, that in general the trust decreasing items led to greater falls in trust than the trust increasing events led to increases in trust. However, on closer scrutiny it was apparent that while some of the 45 items were related to single behavioural instances e.g. " an accident occurs at a plant in another state", many others were related to more general policies spanning many specific instances, e.g. “There is careful selection and training of employees at the plant". We thus re-coded each piece of information as reflecting either a general policy or a specific event and predicted that marginal trust would be influenced more by policies than events related because they have greater breadth and are more diagnostic (White & Eiser, 2005). The results, collapsing across all examples of policies and events, can be seen on the left side of Figure 5 (as with Figure 4, absolute impacts on trust, rather than their direction, are presented).

****Insert Figure 5 about here****

Evidently, the results were not exactly as predicted. True, in overall terms, the impact of policy related information (M = 3.99) was marginally more powerful than event related information (M = 3.80; $F(1, 102) = 3.89$, $p = 0.051$, $_p\eta^2 = .04$). However this overall difference was not large and was clearly moderated by valence $F(1, 102) = 118.17$, $p < 0.001$, $_p\eta^2 = .54$. While, as predicted, positive policies (M = 3.64) had greater impact on trust than positive events (M = 2.50), the reverse was the case for negative pieces of information (events M = 5.10; policies M = 4.35). Since this may have been in part due to the fact that the original study was not conducted to investigate policy vs. event related information, we carried out a second study into the same topic which deliberately manipulated the features of new pieces of information in these terms.

The results can be seen on the right side of Figure 5. This time, the combined effect of policy related information (M = 1.68) on marginal trust in nuclear power plant managers was clearly stronger in overall terms than the event related information (M = 1.19, $F(1,35) = 12.19$, $p = 0.001$, $_p\eta^2 = .26$).

However, once again, this main effect was moderated by valence such that while positive policies (1.69) had a stronger impact than positive events (.83), the difference was much smaller, though not actually reversed, for negative information (events M = 1.55; policies M = 1.68; $F(1,35) = 13.26$, $p < 0.001$, $_p\eta^2 = .28$). The moral of the story seems to be that while more information is generally better than less, specific negative events do seem to have a stronger impact on trust than specific positive events. Moreover, this finding also extended to an industry that was viewed more positively (i.e. the pharmaceuticals industry, see White & Eiser, 2005 for more details) and thus doesn't seem to be explicable simply in terms of a confirmatory bias.

Why might this be the case? Slovic (1993) suggests that it is because “negative *events* often take the form of specific, well-defined incidents such as accidents, lies, discoveries of errors or other mismanagement. Positive *events*, while sometimes visible, more often are fuzzy or indistinct. For example, how many positive *events* are represented by the safe operation of a nuclear power plant for one day?” (italics added). Although, he does not talk about policies, we could argue that both positive and negative policies are relatively "fuzzy" and thus the lack of difference between the impact of positive vs negative policies might be due to their similarity in this respect. The fact that overall they have a greater impact reflects our suggestion that ultimately they are more diagnostic. The take home message for policy makers? If you want to build trust, it might be better to outline the implementation of positive policies that effectively constrain behaviour over a series of events rather than trying to provide people with information about particularly positive instances of performance. Again, this is a more positive conclusion than that offered by a pure negativity bias account of marginal trust, because it suggests that some types of positive information can have a similar or even stronger effect of trust than some types of negative information.

3.4 Assessment of performance under uncertainty: People as Intuitive Signal Detection Theorists

The fourth fundamental psychological process that we consider to be important for marginal trust concerns the results of decisions made under uncertainty. As early as the mid 19th century, pioneering experimental psychophysicists such as Fechner and Weber interested in how changes in objective states of the physical world were experienced psychologically. So, for example, these 'psychophysicists' were interested in how systematic changes in the intensity of light, pressure, noise etc. were subjectively experienced and interpreted. Of particular relevance here was their interest in a phenomenon referred to as the 'just-noticeable-difference'. This concept refers to the amount of change needed in an objective stimulus before it is subjectively detected by a person. So, in a typical experiment a participant might be asked to sit in a pitch black room and say whether or not they think a tiny light had been switched on. Crucial to the situation is uncertainty. At these very low levels of intensity people are often unsure whether the small blinking they see is actually a light in the external world or the kind of light experienced due to the random firing of nerve cells in the brain (i.e. neural

'noise'). When asked if they believe a light had been presented within a certain epoch of time, they can answer either "yes" or "no". Moreover, the experimenter will know whether or not their subjective assessment was correct or incorrect. Thus in a simple study of this type there are actually four different outcomes (see Table 1).

****Insert Table 1 about here****

First, a participant can be correct in two different ways. If there is a light and they correctly identify its presence it is referred to as a True Positive or Hit. If they correctly deduce that no light had been shown it is referred to as a True Negative or All Clear. Secondly, they can be wrong in two different ways. If a light was actually present but they failed to detect it this is referred to as a False Negative or Miss. Finally, if a light was not present but they claim it was this is called a False Positive or False Alarm. Subsequent research across a broad range of phenomenon showed that people vary systematically in both their ability to correctly discriminate signals from noise (i.e. discrimination ability) and in their bias towards responding by saying that a signal was either present or not (i.e. response bias). These response patterns have been incorporated into a theory known as Signal Detection Theory (SDT, Green & Swets, 1974/1988) and have been examined a wide range of applied settings such as radiology, air traffic control, eye witness testimony, clinical assessments and industrial safety (for reviews see Swets, 2000; Swets, Dawes, & Monahan, 2000).

So what has all this got to do with marginal trust? Well, we believe that for many risk-related contexts members of the public look at decision-makers and assess their performance in terms of Signal Detection Theory. That is they are able to distinguish between the four different decision outcomes (though they might not use the explicit labels) and use this information to update their levels of trust in the decision-maker (DM) accordingly. A DM who makes a correct judgement will be rewarded with increasing trust. A DM who makes an error will be penalised with falling trust. For instance, a police officer who shoots a suicide bomber before he can detonate the bomb will be a trusted hero. The officer who shoots an innocent suspect will be seen as a trigger-happy menace to society. In other words, DMs who show the ability to correctly discriminate safe from dangerous situations are likely to be rewarded with higher trust.

However, in addition to being sensitive to a DM's discrimination ability, we believe lay observers are also sensitive to whether or not they betray a certain type of response bias. We suspect, for example, that a doctor who fails to detect the presence of cancer in their patient (a Miss) will be trusted less next time round than one who mistakenly confused a benign growth for a cancerous one (a False Alarm). While a higher tolerance of False Alarms over Misses will be present for many risk related contexts, we suspect that it will not exist for all. Legal systems that adopt an 'innocent until proven guilty' stance are effectively saying that they are prepared to tolerate Misses (i.e. the release of criminals) to avoid False Alarms (i.e. the sentencing of innocents). It is an empirical question, however, whether or not

the public would trust a magistrate more following evidence of a Miss or a False Alarm. It probably depends on the seriousness of the crime. Our main point, however, is that rates of marginal trust are likely to be affected by the outcome of previous decisions in systematic ways and exploration of these patterns will help decision-makers understand how best to build and avoid losing trust.

As originally formulated, SDT assumed that discrimination ability was a function of skill or training, whereas response bias reflected considerations of costs and benefits. In a medical context, both over-treatment (False Alarm) and under-treatment (Miss) can have potentially serious consequences, but obviously this will vary as a function of many factors, such as the urgency of the patient's condition and the aggressiveness of the treatment. Health economics also attempts to calculate the net benefit (in units known as 'Quality Adjusted Life Years') of a successful treatment (Hit), to set against the monetary cost of providing such treatment. All these factors can enter into a doctor's treatment recommendation, over and above the actual diagnosis. Likewise, in our more trivial example of a walk in the rain, the decision to go for a walk will depend not just on the forecast, but on how bothered one would be about getting wet. Since such costs and benefits provide the motivation for a riskier or more cautious response bias on the part of a DM, they also provide a basis for *attributions* of bias by observers of that DM's choices. A company that stands to make a profit out of some industrial activity or product is likely to be seen as biased towards supporting it, even if there are associated risks or costs to other people. And of course, if we are prescribed medication by a doctor, we feel entitled to expect that this prescription will be determined by our clinical need, rather than any incentive given to our doctor by a pharmaceutical company.

Moreover, in addition to the two dimensions of discrimination ability and response bias proposed by SDT, we believe that a third dimension of performance is also likely to be important when considering trust in risk managers. In many instances risk managers are also risk communicators. It is often not sufficient for them to simply make correct decisions or adopt an appropriate response bias, they also need to communicate the outcome of these decisions to the public. Concerns about openness and transparency in decision-making are already well-known to policy makers which is why they often call for public enquiries and the like because it is now widely known that it is not only the outcomes that matter to people but the procedures used to attain these outcomes as well (e.g. Tyler & DeGoey, 1996). So a DM who fails to be open and transparent about their decision-making processes is likely to be punished with falling trust while one that is open will be rewarded with higher relative levels of trust next time around. For example, we expect that the decision-maker who makes a Miss and tries to cover it up will be subsequently trusted much less than the one who makes a Miss but openly acknowledged their error. We refer to this third dimension of performance as *communication bias*, to reflect the fact that like response bias it represents a particular tendency towards one kind of response (to be open or not) under conditions of uncertainty.

In sum, we believe that in terms of marginal trust in decision makers, observers effectively operate as *Intuitive* Signal Detection Theorists, or Intuitive Detection Theorists (IDTs) for short. These IDTs, we argue, adjust their levels of trust on the basis of information about the levels of discrimination ability, response bias and communication bias witnessed in previous decisions. Interestingly, although derived from a theoretical approach to understanding decisions under uncertainty, it is enlightening to compare these three decisions to the correlates observed for baseline trust (see Section 2.2). Recall for example that Peters, Covello and McCallum's (1997) also made a tri-partite distinction between a) knowledge and expertise, b) care and concern, and c) openness and honesty. The first dimension appears to be very similar to discrimination ability. The second appears related to response bias and the third has obvious links to communication bias. In this way there seems to be a degree of convergent validity between our own approach to marginal rates of trust and beliefs about trust in general. What our approach offers in addition is predictions about how specific decisions that vary systematically will lead to increases and decreases in marginal rates of trust. Specifically, our model makes the following three predictions:

H1) Correct decisions (Hits & All Clears) will be associated with more positive changes in trust than incorrect decisions (False Alarms & Misses).

H2) Decisions showing cautious danger response bias (Hits and False Alarms) will be associated with more positive changes in trust than decisions showing a risky response bias (All Clears & Misses).

H3) Risk managers who are 'open' about their decisions will be associated with more positive changes in trust than risk managers who are 'closed'.

We have now tested these predictions in a number of experiments across risk contexts as diverse as nuclear power, travel vaccinations, computer viruses and the licensing of new prescription drugs and by using both within and between participant designs. Although the research program is very much in its preliminary stages, the results across these different contexts and methods are remarkably consistent (White & Eiser under review) In order to provide a flavour of the results we return again to the issue of nuclear power since in many ways it was this issue that sparked much of the discussion into marginal trust (Slovic, 1993). In this study, we again replicated Slovic's procedure of showing participants a number of positive and negative events in a nuclear power station and asked them to record how knowledge of these events would change their levels of trust in the plant's management. We were careful to make sure that all items were events - rather than policies (see section 3.3 above) to keep the level of information specificity and diagnosticity as similar as possible.

Importantly, in terms of our model, participants were asked to evaluate both open and closed examples of Hits, Misses, False Alarms and All Clears. Thus in total there were eight events to evaluate. An example of an Open False Alarm was, "Management shut down operations and warned local residents

of an incident at the plant even though it turned out to be only a faulty warning system." An example of a Closed False Alarm was "Employees were quickly evacuated and operations stopped following a suspected fire but since it was only a false alarm, managers did not inform locals". The subsequent changes in marginal trust as a result of the eight type of event can be seen in Figure 6.

****Insert Figure 6 about here****

On the left side of Figure 6 are the responses to the four types of outcome when the managers had been open and honest. On the right are the outcomes when they had been less than fully open about events at the plant. In line with Hypothesis 3, marginal trust rates were lower in the latter *(*M = -.80) than the former case (M = .22, $F(1,63) = 73.42$, $p < 0.001$, $\eta_p^2 = .54$). In other words, trust suffered more when a lack of transparency was evident. In line with Hypothesis 2, Hits and False Alarms (when combined together, M = .37) resulted in higher marginal trust than All Clears and Misses (M = -1.96, $F(1,63) = 106.31$, $p < 0.001$, $\eta_p^2 = .63$). However, there was no overall support for Hypothesis 1. Correct judgments (Hits and All Clears, M = -.19) did not result in significant improvements in marginal trust compared to incorrect ones (Misses & False Alarms, M = -.39, $F(1,63) = 2.75$, n.s., $\eta_p^2 = .04$).

The primary reason for this unexpected finding is clear. Open False Alarms actually increased marginal trust - in fact more than an Open Hit. Moreover, Closed False Alarms did not lead to a drop in trust, in contrast to all three other Closed outcomes. In other words, there was a type of error that did not decrease marginal trust. Another interesting aspect of Figure 6 is that the only instance where marginal trust really suffered heavily was for the Closed Miss. Trust fell much less in the case of an Open Miss, even though it was quite serious, i.e. "Management releases press statement saying that it was wrong to ignore signs of storage tank corrosion which led to a toxic release into the environment". In other words, while all such effects will depend to a great extent on the magnitude of costs and benefits involved, people seemed relatively tolerant of even quite serious errors as long as the decision maker was prepared to admit their mistake. It is only when such mistakes are covered up that trust really suffers in the way described by Slovic and others.

Moreover, as noted above, this basic pattern was not restricted to the nuclear power context or the particular methodology used but instead generalised across a range of situations. There was even evidence that these changes in trust were not restricted to the specific decision-makers who actually made the errors or correct decisions. Rather, marginal trust rates also generalised to the category of decision-maker. For example, in a subsequent study (White & Eiser, Study 2, under review) we assessed prior levels of baseline trust in various risk managers in general (e.g. nuclear power plant managers, doctors and computer support personnel). Then, right at the end of the study, following the presentation of event types and measurement of marginal trust changes in those who had made specific correct or incorrect judgements, we again asked for baseline trust judgements of the relevant

categories. What we noticed was that when a particular doctor, say, wrongly judged a situation to be safe, not only did trust in this doctor drop considerably, baseline trust for doctors in general fell (controlling for prior levels). Although the fall was somewhat smaller for general baseline trust, the fact that one decision-maker's error could influence trust in the general category was intriguing. Certainly, it is consistent with viewing trust in the present context as role-based rather than interpersonal (Kramer, 1999). That is, a mistake by a specific category exemplar serves to undermine trust in the system which legitimized their position and thus is more likely to generalize to others in this system.

At this stage these findings are preliminary and require replication and further investigation. A number of issues remain. First, all of our contexts focused on technological risks. Further research needs to extend the approach to other types of risk setting, especially ones where an alternative response biases might be more appropriate. For example we are currently exploring the idea that humans can be hazards too as in the case of suicide bombers. In this research we aim to examine people's trust in police officers who make a decision to shoot or not shoot a suspect. We expect that False Alarms may not result in the same kinds of positive effects on marginal trust witnessed for technological hazards. Second, we have so far only examined single instances of performance or pieces of information. What happens when people are exposed to a two, three or four pieces of information about the same actors? The story of the 'boy who cried wolf' tells us that while one False Alarm may be tolerated, several in a row will not. Further research needs to examine the dynamic aspect of decisions over time and their influence on marginal trust in order to establish a more realistic picture of what happens in applied contexts.

3.5 Summary

Section 3 has introduced four well-known psychological mechanisms that might be important for understanding marginal rates of trust. We saw that people tend to give more weight to negative than positive information and thus in some contexts this results in trust being easier to lose than gain. We all know from our own experience that when someone we know does something really daft, it might take us a long time, if ever, to trust them again. However, subsequent sections suggested that this negativity bias is not the only process at work for trust. In order to maintain cognitive consistency, people also tend to attenuate negative messages about those they already have high levels of trust in. Decision-makers are more likely to be given the benefit of the doubt if they have already built a strong foundation of trust than ones who haven't. Moreover, information varies in its specificity and hence diagnosticity. Some messages are more powerful than others for marginal trust because they relate to either single or multiple behaviours. Policy makers interested in building trust would, it appears, do well to emphasise positive policies than one off good deeds. Finally, insights from Signal Detection Theory enabled us to suggest that people are sensitive to the exact outcomes of decisions when

adjusting their levels of trust. In the context of many technological hazards, it seemed that Misses, especially ones where the actor was less than fully transparent, seemed to be exactly the kind of 'really daft' behaviours that led to catastrophic falls in trust. Single Open False Alarms, although incorrect assessments of danger, led to marginal increases in trust suggesting that not all negative information leads to a fall in trust. Moreover, these patterns appeared to extend to other related decision-makers not directly involved in the specific instances under consideration. Clearly, more work is needed to investigate how many of such errors will be tolerated and whether there are contexts where Misses will actually be preferred to False Alarms.

4.0 Concluding comments

Many of the risks we face in everyday life are the product of human decisions of one kind or another. Risk perception is therefore often as much a social judgment as an assessment based on various characteristics of the hazard itself. Since social judgments are fundamentally social psychological in nature we believe the discipline has much to offer further understanding of risk perception processes. Although the term *trust* is widely used when discussing these social judgments it is important to distinguish this kind of 'role-based trust' which relates to the target's abilities to fulfill designated risk management roles, from the kind of interpersonal trust we have in our friends and family. Prior research into the public's role based trust in risk managers suggests that it is influenced by perceptions of their ability to distinguish 'dangerous' from 'safe' situations, their propensity to act in the public interest when uncertain, and their openness and transparency in communications about the decisions they have made. Based on these three facets it is easy to see why general or baseline trust in doctors is usually quite high while that in industry representatives is not. Both are generally seen as knowledgeable about the risks, but only the former tend to be thought of as acting in the public interest and communicating risk information in a transparent fashion.

However, knowledge about baseline trust rankings is now widespread and, not surprisingly, those at the bottom want to know how to improve their ranking while those at the top want to know how to stay there. That is, risk managers and other key decision makers want to know how to gain trust and avoid losing it. To investigate these more dynamic issues we reviewed the potential role of four well known psychological mechanisms namely, i) a general negativity bias, ii) the desire to maintain cognitive consistency, iii) the greater diagnosticity of information with broader specificity and iv) psychophysical processes relating to the detection of signals under conditions of uncertainty. In turn each of these four processes seemed to add valuable insights into rates of marginal trust, i.e. the amount of trust that could be gained or lost as a function of a single event or outcome. Generally speaking, bad news had a larger (negative) impact on trust than good news but this was moderated by prior attitudes towards the hazard and the risk manager, by the amount of information conveyed and

the exact nature of the error or correct decision. Moreover, we would be surprised if there weren't further psychological processes that could shed additional light on these issues.

That trust isn’t as asymmetric, i.e. hard to again but easy to lose, as previously thought is good news for many but perhaps not so surprising after all. Research into the related field of the evolution of cooperation has demonstrated, for example, that ultimately a 'tit-for-tat' strategy (i.e. start off cooperatively and then mirror the behaviour of the other) generally rewards 'organisms' with higher pay offs than a vengeful strategy where a single defection leads to repeated non-cooperation (Axelrod, 1980). If we have evolved to use strategies of this sort, which seems plausible given the social circumstances of our evolutionary history, then trust can be rebuilt, maybe not as quickly as a simple tit-for-tat mechanism would imply, but certainly in the long run following repeatedly 'cooperative' behaviours. Clearly, the definition of the term cooperative depends on whose perspective we are taking but where we are concerned with public perceptions in general, cooperation often seems to be a function of performing one's allotted tasks to the best of one's ability, acting in the public interest and being open and honest. Risk managers and other key decision makers would do well to remember that cooperation and trust are, at the end of the day, two-way process however much the complexity of modern societies has separated them physically from those whose trust they desire.

References

Axelrod, R. (1984). *The Evolution of Cooperation.* New York: Basic Books.

Barber, B. (1983). *The Logic and Limits of Trust*. New Brunswick: Rutgers University Press.

Baumeister, R.F., Bratslavsky, E., Finkenauer, C., & Vohs, K.D. (2001). Bad is stronger than *Review of General Psychology, 5,* 323-370.

Burt, R. S. & Knez, M. (1996). Trust and third-party gossip. In R.M. Kramer & T.M. Tyler (Eds*.) Trust in Organizations: Frontiers of theory and research* (P.68-89). Sage: Thousand Oaks.

Calman, K.C. (2002). Communication of risk: choice, consent and trust. *Lancet, 360,* p166-68.

Cvetkovich, G., Siegrist, M., Murray, R., & Tragesser, S. (2002). New information and social trust: Asymmetry and perseverance of attributions about hazard managers., *Risk Analysis*, *22*, 359-367.

Dasgupta, P. (1988/2000) Trust as commodity. In D. Gambetta (ed), *Trust: Making and Breaking Cooperative Relations*, electronic edition, Department of Sociology, University of Oxford, Chapter 4 pp.49-72 <http:77www.sociology.ox.ac.uk7papers7dasgupta.pdf> (p.50)

Eiser, J.R. (1990). *Social Judgement*. Milton Keynes: Open University Press.

Eiser, J.R., Miles, S., & Frewer, L.J. (2002). Trust, perceived risk and attitudes toward food technologies. *Journal of Applied Social Psychology, 32*, 2423-2433.

Eiser, J.R., van der Pligt, J. & Spears, R. (1995). *Nuclear Neighbourhoods*. Exeter: Exeter University Press.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford: Stanford University Press.

Freudenburg, W.R., Coleman, C.L., Gonzales, J., & Hegeland, C. (1996). Media coverage of hazard events - analysing the assumptions. *Risk Analysis*, *16* (1) 31-42.

Frewer, L.J., Howard, C., Hedderley, D., & Shepherd, R. (1996). What determines trust in information about food related risks? *Risk Analysis*, *16*, 473-486.

Green, D. A. & Swets, J.A. (1974/1988). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Hardin, R. (2001). Conceptions and explanations of trust. In K. Cook, (Ed.) *Trust in Society*. (pp.3-40). New York: Russell Sage Foundation.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York: Wiley.

Helliwell, J.F. & Putnam, R.D. (2004). The social context of well-being. *Royal Society's Philospohical Transactions: Biological Sciences*, 359, 1435-1446.

Hovland C.I., Janis, I.L. & Kelley, H.H. (1953). *Communication & Persuasion*. New Haven, Yale Univ. Press.

Johnson, B.B. (1999). Exploring dimensionality in the origins of hazard-related trust. *Journal of Risk Research*, 2, 325-354.

Jungermann, H., Pfister, H. R., & Fischer, K. (1996). Credibility, information preferences, and information interests. *Risk Analysis*, *16* (2), p251-261.

Kahneman, D. & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin & D. Kahneman (Eds.) *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: CUP.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, *47,* 263 - 291.

Kanouse, D.E. & Hanson, L. (1972). Negativity in evaluations. In E.E. Jones, D.E. Kanouse, S.Valins, H.H. Kelley,, R.E. Nisbett, & B.Weiner (Eds.), *Attribution: Perceiving the causes of behaviour* (pp. 47-62). Morristown, NJ: General Learning Press.

Kapserson, R.E., Golding, D., & Kasperson, J.X. (1999). Risk, trust and democratic theory. In G. Cvetkovich, & R. Lofstedt, (Eds.) (1999). *Social Trust and the Management of Risk* (pp.42-52). Earthscan: London.

Kasperson, R.E., Golding, D., & Tuler, S. (1992). Social distrust as a factor in siting hazardous facilities and communicating risks. *Journal of Social Issues*, *48*, 161-187.

Koren, G. & Klein, N. (1991). Bias against negative studies in newspaper reports of medical research. *Journal of American Medical Association, 266*, 1824-1826.

Kramer, R.M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, *50*, 569-98.

Kramer, R.M. & Tyler, T.R. (1996). (Eds.) *Trust in Organizations: Frontiers of theory and research*. London: Sage.

Langford, I.H, Marris, C. & O'Riordan, T. (1999). Public reactions to risk: Social structures, images of science, and the role of trust. In P. Bennett, & K. Calman, (Eds.) *Risk Communication & Public Health* (pp.33-50). Oxford: OUP.

Layard, R. (2005). *Happiness: Lessons from a new science*. London: Penguin.

Levi, M. (1998) A state of trust. In M. Levi, & V. Braithwaite, eds, *Trust & Governance*. New York: Russell Sage Foundation.

Maeda, Y. & Miyahara, M. (2003). Determinants of trust in industry, government, and citizen's groups in Japan. *Risk Analysis*, *23*, (2), 303-310.

Mayer, R.C., Davis, J.H., & Schoorman, F.D. (1995). An integrative model of organizational trust. Academy of Management Review, 20, 709-734.

Meyerson, D., Weick, K.E. & Kramer, R.M. (1996). Swift trust and temporary groups. In R.M. Kramer & T.R. Tyler (Eds.) *Trust in Organizations: Frontiers of theory and research* (pp.166-195). London: Sage.

Mishra, A. (1996). Organizational responses to crisis: The centrality of trust. In R.M. Kramer & T.R. Tyler (Eds.) *Trust in Organizations: Frontiers of theory and research* (pp.261-287). London: Sage.

O'Neill, O. (2002). A question of trust. *Reith Lectures 2002*. http://www.bbc.co.uk/radio4/reith2002/

Peeters, G. & Capinsky, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational effects. In W. Stroebe & M. Hewstone (eds.), *European Review of Social Psychology* (Vol 1, pp.33-60). New York: Wiley.

Peters, R.G., Covello, V.T. & McCallum, D.B. (1997). The determinants of trust and credibility in environmental risk communication: An empirical study. *Risk Analysis*, *17* (1), 43-54.

Poortinga, W. & Pidgeon, N. (2003). Exploring the dimensionality of trust in risk regulation. *Risk Analysis*, *23* (5) 961-972. Poortinga, W., & Pidgeon, N., (2004). Trust, the asymmetry principle and the role of prior beliefs. *Risk Analysis*, 24, 6, 1475-1486.

Pratto, F. & John, O. (1991). Automatic vigilance: The attention-grabbing power of negative Social information. *Journal of Personality and Social Psychology*, *61*, 380-391.

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

Renn, O. & Levine, D. (1991). Credibility and trust in risk communication. In R.E. Kasperson & P.J.M. Stallen (Eds.). *Communicating risks to the public* (pp.175-218). The Hague: Kluwer.

Rothbart, M. & Park, B. (1986). On the confirmability and disconfirmability of trait concepts. *Journal of Personality and Social Psychology*, *50*, (1), pp 131-142.

Rozin, P. & Royzman, E.B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*, 296-320.

Seligman, M.E.P., (1970). On the generality of the laws of learning. *Psychological Review*, *77*, 406-418.

Siegrist, M. & Cvetkovich, G. (2000). Perception of hazards: The role of social trust and knowledge. *Risk Analysis*, *20* (5), 713-719.

Siegrist, M., & Cvetkovich, G.T. (2001). Better negative than positive? Evidence of a bias for negative information about possible health dangers. *Risk Analysis*, *21*, 199-206.

Siegrist, M., Earle, T., & Gutscher, H. (2003). Test of a trust and confidence model in the applied context of Electromagnetic Fields (EMF) risks. *Risk Analysis, 23*, 705-716.

Simon, H.A. (1957). *Models of man: Social and Rational*. New York: Wiley.

Skowronski, J.J. & Carlston, D.E (1987). Social judgement and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology*, *52*, 689-699.

Slovic, P. (1993). Perceived risk, trust and democracy. *Risk Analysis*, *13* (6), 675-82.

Slovic, P., Flynn, J., Johnson, S.M., & Mertz, C.K. (1993). *The dynamics of trust in situations of risk*. (Report No 93-2). Eugene, OR: Decision Research.

Swets, J.A. (2000). Enhancing diagnostic decisions. In Connolly, T., Arkes, H.R. & Hammond, K. (Eds.) *Judgement and Decision Making: An interdisciplinary Reader (2nd Edition,* pp.66.81). Cambridge: CUP.

Swets, J.A., Dawes, R.B., & Monahan, J. (2000) . Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1-26.

Taylor, S.E. (1991). Asymmetrical Effects of Positive and Negative events: The mobilzation-minimization hypothesis. *Psychological Bulletin*, *110*, 67-85.

Tyler, T.R. & DeGoey, P. (1996). Trust in organizational authorities. The influence of motive attributions on willingness to accept decisions. In R.M. Kramer & T.M. Tyler (Eds.) *Trust in Organizations: Frontiers of theory and research* (pp.331-356). Thousand Oaks: Sage.

Uslaner, E.M. (2003). *The Moral Foundations of Trust*. Cambridge: CUP.

White, M.P., & Eiser, J.R. (2005). Information specificity and hazard risk potential as moderators of trust asymmetry. *Risk Analysis*, 25, 5, 1187-98.

White, M.P. & Eiser, J.R. (in press). A social judgement approach to trust: People as intuitive detection theorists. In M. Siegrist & T.Earle (Eds), *Trust and Risk Management*. London: Earthscan.

White, M.P., & Eiser, J.R. (under review). Event type as moderator of trust asymmetry: Signal detection theory and social trust in decision makers.

White, M.P., Pahl, S., Buehner, M. & Haye, A. (2003). Trust in risky messages: The role of prior attitudes. *Risk Analysis*, 23, 717-26.

Yamagashi, T. (2001) Trust as a form of social intelligence. In K. Cook, (Ed.) *Trust in Society*. (pp.121-147). New York: Russell Sage Foundation.

Figure 1: Role-based trust in the processing of risk

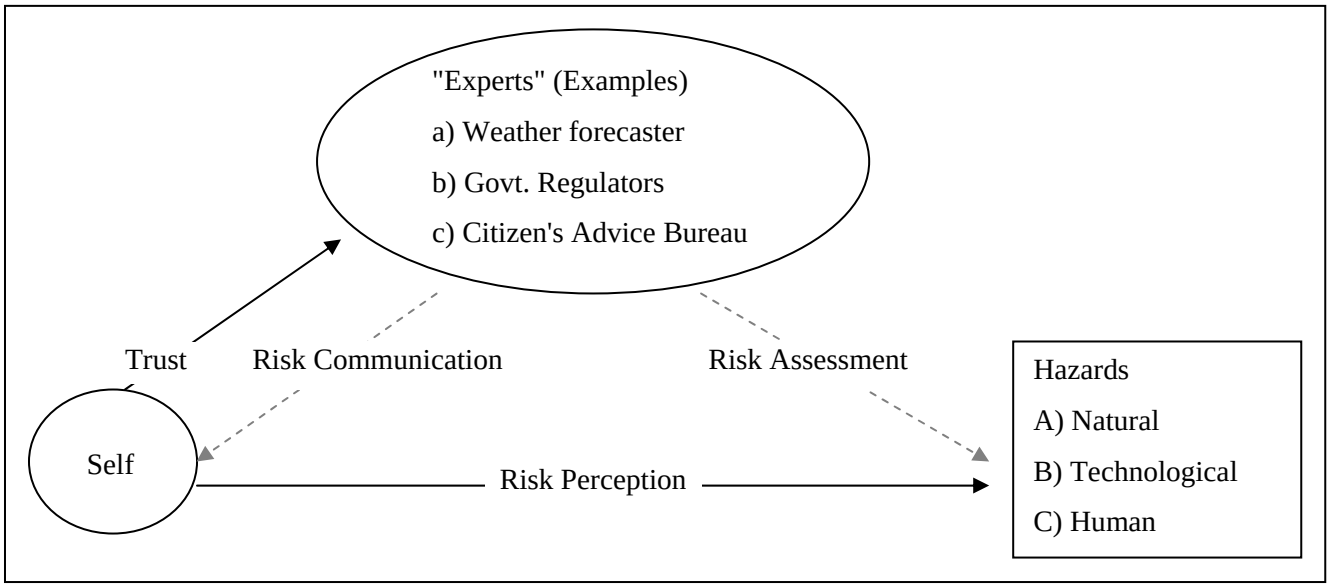


Figure 2: Baseline trust in various targets in two risk contexts (White & Eiser, in press)

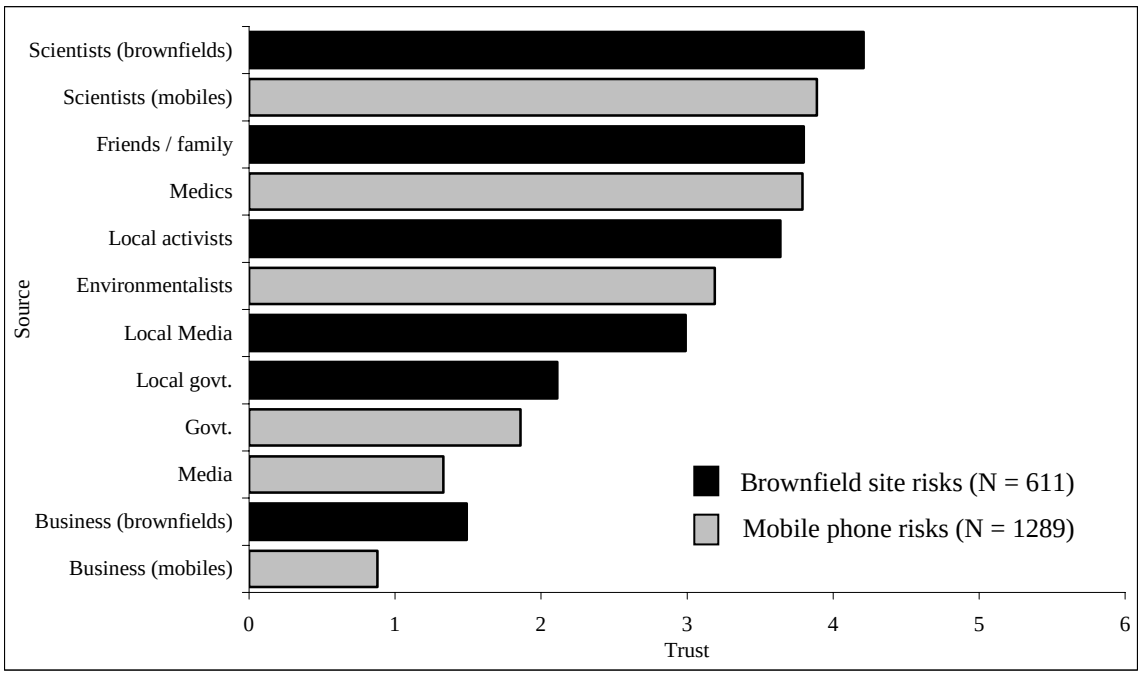
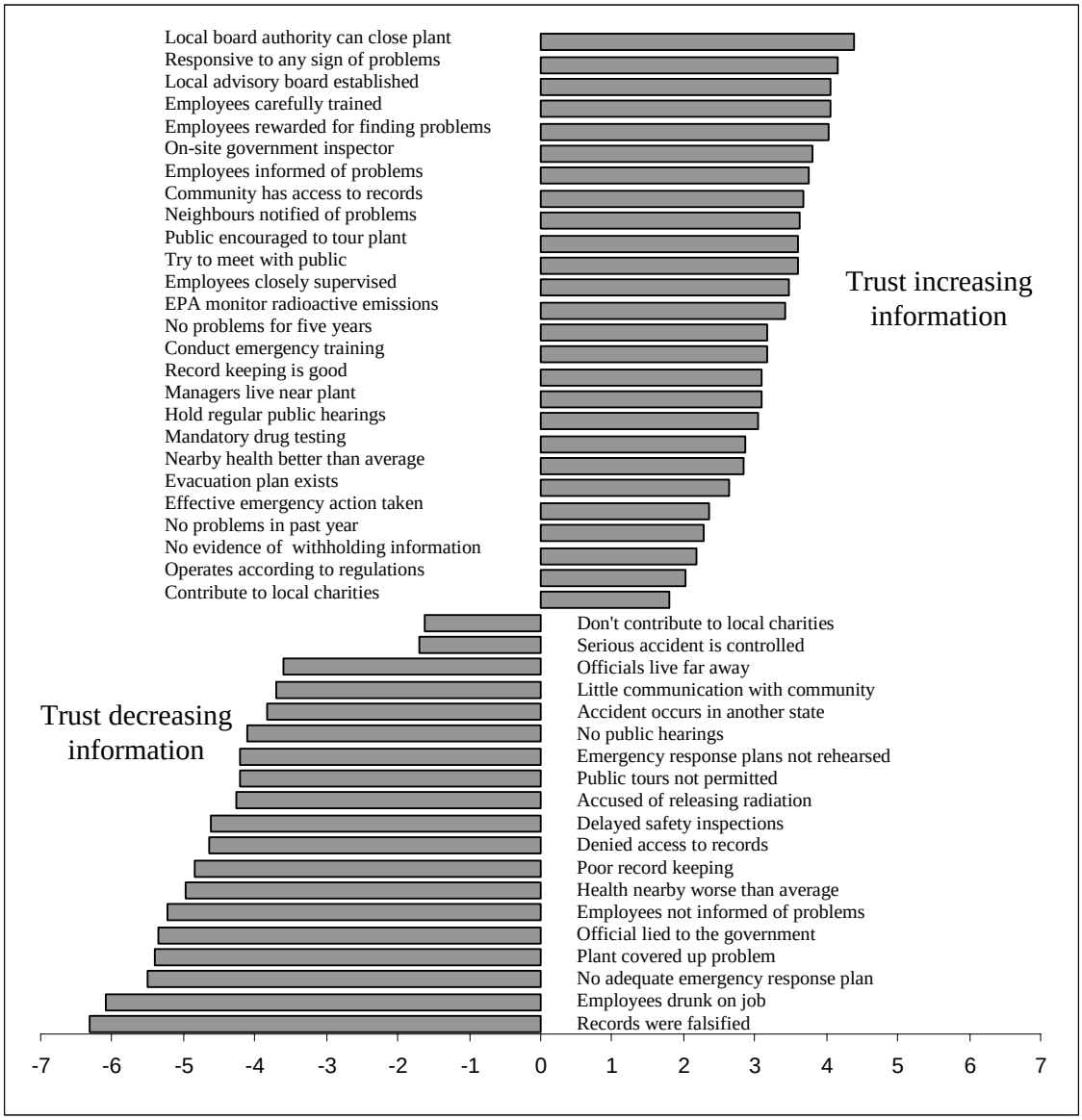


Figure 3: Slovic (1993) data re-analysed



N.B. Only annotated items are shown. For complete items please contact the authors or see Slovic (1993).

Figure 4: Confirmatory bias and heavy industry (White & Eiser, 2005, Study 2)

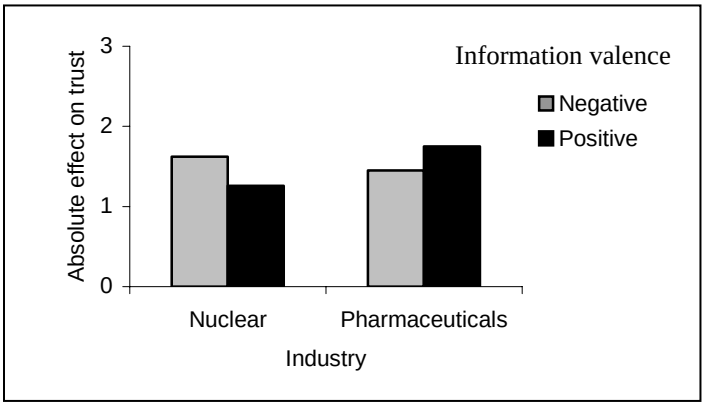
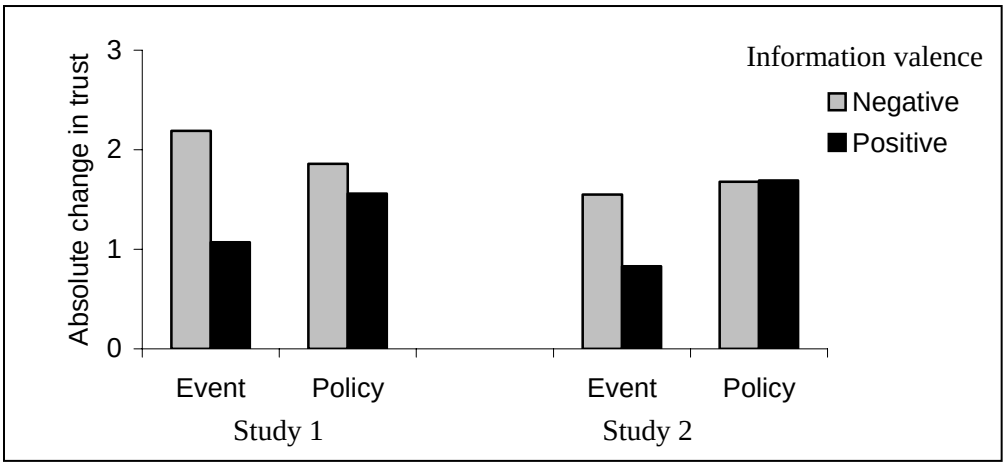


Figure 5: Information specificity: Events vs Policies in the nuclear industry (White & Eiser, 2005)



N.B. For the sake of comparisons Study 1 scores (from -7 to +7) have been rescaled to make them more comparable with Study 2 scores (-3 to +3).

Table 1: The four possible outcomes of a simple binary detection task

		Perceiver says signal is	
		"Present"	"Absent"
Signal really is	Present	TRUE POSITIVE (Hit)	FALSE NEGATIVE (Miss)
	Absent	FALSE POSITIVE (False Alarm)	TRUE NEGATIVE (All Clear)

Figure 6: IDT findings (White & Eiser, under review, Study 1)