A draft of a chapter for

David N. Rapp & Jason L. G. Braasch (Eds.) "Processing Inaccurate Information: Theoretical and Applied Perspectives from Cognitive Science and the Educational Sciences" The MIT press, 2014

Discounting information: When false information is preserved and when it is not

Yaacov Schul and Ruth Mayo

The Hebrew University

June 2013

Preparation of this manuscript was funded by ISF Grants 124/08 (YS) and 594/12 (RM).

Correspondence concerning this article should be addressed to the authors at the Department of Psychology, The Hebrew University of Jerusalem, Israel, Email: ruti.mayo@huji.ac.il, yschul@huji.ac.il

Discounting information: When false information is preserved and when it is not

Although people often assume that communicators are cooperative (Grice, 1975), they are also well prepared for deception. Evolutionary theory assumes that deception is inherent to living in groups, and there are empirical demonstrations indicating that lying is common in everyday interactions (DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996; DePaulo & Kashy, 1998; Feldman, Forrest, & Happ, 2002). It is therefore not surprising that consumers distrust product information provided by sellers (e.g., Dyer & Kuehl, 1978; Prendergust, Liu, & Poon, 2009) or that voters are suspicious of messages coming from political candidates (Schyns & Koop, 2010). Hence, it seems reasonable to conclude that in many if not most of their dealings with others, people are aware of the possibility of being misled (Schul & Burnstein, 1998; Schul, Mayo, & Burnstein 2004).

After so many generations of coping with the need to identify deception, one might think that human beings would have evolved into highly accurate social perceivers. Yet, as dozens of studies suggest, the accuracy of interpersonal perception is modest at best. In particular, as a recent review (Hartwig & Bond, 2011) suggests, liars seem to win the Darwinian 'arms race' between senders (who attempt to deceive) and receivers (who strive to detect deception).

Because using false information from others might be very costly, and at the same time, the detection of falsehoods communicated by others proves very difficult, one expects that receivers would have developed skills that allow them to discount false information once such information is identified. However, as past research suggests, the

success of discounting such information is limited. This chapter discusses the obstacles that prevent people from attaining the discounting challenge, and the conditions that promote successful discounting.

Overview

We start by providing a very brief review of past research on discounting that emphasizes the importance of the nature of encoding and the strength of the request to discount for successful discounting. We note that receivers might be motivated to discount a particular testimony based on what they know about the source's motivations or abilities. That is, people attempt to discount when they discover that the source of a testimony attempts to deceive them, or when the source appears to be incompetent (e.g., Eagly, Wood, & Chaiken, 1978). Notwithstanding, we remark that discounting success might differ in the two cases because the reason for discounting matters. Then, we view the discounting challenge from three perspectives. First, we consider the literature on negations that offers some insight into the cognitive routes that might be taken when people attempt to discount. We continue by describing research on implicit truth, which suggests that failure to discount might be more likely when the to-be-discounted information feels true, even when the receiver knows it to be false. We end by considering the mind-set of receivers. Specifically, we compare the states of trust and distrust and argue that when one distrusts, discounting might be more successful than when one trusts. The trust/distrust comparison provides some understanding of the obstacles to discounting.

Early research on the success of discounting

Early explorations into the phenomenon of information discounting run along two main lines: Research about the success of discounting of invalid testimonies in court settings (e.g., Elliott, Farrington, & Manheimer, 1988; Hatvany & Strack, 1980; Kassin & Wrightsman, 1980, 1981; Thompson, Fong, & Rosenhan, 1981; Wolf & Montgomery, 1977), and the research about belief perseverance (see reviews in Anderson, New, & Speer, 1985; Schul & Burnstein, 1985). Interestingly, whereas the bulk of the early research on discounting in court settings suggests that individuals can successfully ignore an invalid testimony, studies done within the belief perseverance paradigm suggest the opposite. This inconsistency points to several key differences between the two paradigms which are important in understanding the process of information discounting. The two paradigms differ in the nature of the encoding of the to-be-used (TBU) and the to-be-discounted (TBD) information, the strength of the request to discount, and the motivations of the decision makers to succeed in appropriate discounting. These key factors will be discussed below. A more comprehensive analysis of the success of discounting in court settings published recently by Stelbay, Hosch, Culhane, and McWethy (2006) shows that in contrast to the conclusion gleaned from the early studies, inadmissible evidence has a reliable impact on verdicts or judgments of guilt, that is, that discounting fails even in courtroom simulations. Still, their meta-analysis reveals that a strong admonition by a judge to disregard the inadmissible evidence can nullify this effect so that discounting of invalid testimonies succeeds.

What is so unique in the discounting phenomenon relative to other cases in which people have to avoid biases? In the typical discounting situation, the receiver does not

know at the time the information is encoded whether it would have to be used (TBU) or discounted (TBD) at the time the judgment will be made[1]. In contrast, in the cases of having to avoid known biases such as those associated with stereotypes, one can identify the bias-related information and its implications very early in the process, even before that information has been fully processed. Therefore, receivers are better able to prepare themselves, if they are motivated to do so, to discount the biased information. For example, receivers might be aware of their tendency to treat others differently due to the way they speak (e.g., foreigners' poor command of the language). If they desire so, they can overcome the bias by considering the essential aspects of the communication (e.g., its content) while effortfully ignoring the incidental aspects (e.g., the style) at the time of encoding. Although by itself this challenge might be very hard to attain, the challenge of information discounting is even harder. This is because the biases in the typical discounting situation are discovered only after encoding, because during encoding one does not know which piece of information would have to be discounted (see Schul, Burnstein, & Bardi, 1996).

Still, in spite of the difficulty of the challenge to discount, there are conditions that facilitate peoples' attempts to remove the influence of the TBD information from their judgments. Research suggests that when people are aware of the potential influence of the TBD testimony, and especially its direction of strength, when they have cognitive resources to operate on this knowledge, and when they have the motivation to avoid the bias, they can avoid the impact of the TBD testimony (e.g., Martin, Seta, & Crelia, 1990; Schwarz, et al, 2007; Strack & Hannover, 1996; Wilson & Brekke, 1994), at least when

[1] As suggested later on, a state of distrust may lead receivers to encode information spontaneously as if it is TBD information.

they make explicit judgments. This is done by mentally "subtracting" the bias from the initial judgment response. Going back to our previous example, if one knows that he or she tends to discriminate against foreigners with poor command of English, he or she can mentally add positive valence to an initial (negatively-biased) impression of the foreigner, thereby attempting to overcome the bias. The example illustrates that discounting can be achieved by a correction at the response level.

The response-level correction is one of the major mechanisms that might be used for discounting. Research shows that under well-specified conditions people might be able to undo the bias brought about by the TBD information in making their judgments. However, it is important to note that the correction is done at the response (judgment) level rather than through reinterpretation of the information. As a consequence, what appears to be a successful discounting in judgments that are made in close temporal proximity to the discounting request, turns out to be failure to discount in judgments made when the request to discount is no longer active in memory. Later on we shall describe other mechanisms of discounting and discuss the shortcomings of correction in more detail. However, before doing so, let us describe three studies from our laboratory that demonstrate how predictions derived from considering the correction mechanism can shed light on the success of discounting.

Schul and Goren (1997) found that individuals who were asked to ignore a strong testimony adjusted their judgments more than those asked to ignore a testimony with a milder persuasive impact. They proposed that as individuals consider the implication of a testimony for the judgment that they are making, they also create a meta-cognitive assessment of the impact of that testimony. When a testimony has to be discounted, the

meta-cognitive assessment of its impact is utilized, and the judgment is adjusted (corrected) accordingly. The findings reveal that when a critical testimony was weak, participants tended to underestimate its persuasive impact. As a result, they under-corrected, and consequently fail to discount when instructed to ignore it. In contrast, when the critical testimony was strong, participants tended to slightly overestimate its persuasive impact. Consequently, discounting succeeded.

The research reported in Schul and Manzury (1990) highlights the importance of the reminders of the requirement to discount. Participants in their study were asked to discount a testimony and make three types of judgments: with respect to the guilt of the accused, his aggressiveness, and his likability. The question of interest has to do with the differences between the three judgments. Making a judgment of guilt is the essence of the decision-maker's activity in court. Therefore, the norms of judgments imposed by the court setting become maximally relevant. Judgments of likability or aggressiveness are less central to court decision making. Therefore, the pressure to conform to the standards of judgments in court becomes weaker. Indeed, the participants failed to discount the TBD testimony properly when they made judgments about the aggressiveness and likability of the accused person. However, when the same participants were making judgments of guilt, they were unaffected by the TBD testimony; that is, discounting succeeded. Schul and Manzury interpreted this finding to mean that although people know what they need to do, they act on this knowledge only when the judgment reminds them of the need to correct (see also, Schul, 1993).

The third example we describe involves the nature of the motivation of participants to discount appropriately. Saar and Schul (1996) manipulated this motivation

in two complementary ways. First, they reasoned that decision makers should be more motivated to discount when they are held publicly accountable for their answers (see Lerner & Tetlock, 1999, 2003 for a more comprehensive analysis of the conditions under which accountability is likely to reduce biases). Second, they noted that the motivation to discount properly is stronger when the reasons given for discounting are substantive rather than procedural (see Golding & Hauselt, 1994; Kassin, & Sommers, 1997; Steblay et al, 2006; see also Demaine, 2008).

In order to investigate the joint impact of these two motivational forces, respondents were randomly assigned to one of six cells. There were two conditions in which respondents were given reasons (either procedural or substantive) to ignore an argument. Respondents in a third condition (termed *no-request* control condition) received the same TBD argument, but were not asked to ignore it. These three conditions were crossed with an accountability manipulation. Half the respondents were told at the onset of the experiment that they would be asked to explain their responses to the experimenter. These respondents were also instructed to write their name on the questionnaire. The remaining respondents were not instructed to identify themselves on the questionnaire, nor did they anticipate having to explain their responses.

The stimulus material involved a fictional protocol of a meeting whereby the decision about a fee reduction at the University's recreation center was discussed. The Center's head trainer presented the TBD argument, arguing against a fee reduction. The substantive request to ignore his testimony attacked the cogency of the TBD by showing that the center's head trainer omitted important details, which would have made his argument false. The procedural request challenged the formal qualifications of the head

trainer to present financial data. In both cases the chair of the meeting instructed the meeting discussants to ignore the TBD argument. Participants were asked to assume the role of someone in the meeting and rate their agreement with fee reduction. The Table below presents the mean judgments after standardization. Higher positive numbers reflect more agreement with fee reduction.

Participants who used all the information (no-request control condition) found the fee reduction request relatively unacceptable, indicating that the critical testimony (that was discounted in the other conditions) was by itself persuasive. Participants in the two other conditions who were asked to discount the critical testimony found the proposal for fee reduction acceptable. Notwithstanding, whereas the accountability manipulation had no effect on participants who received the substantive request to discount, it affected those who received the procedural request to discount. That is, participants who received a request to discount based on procedural grounds and were not held accountable managed to discount the TBD argument and thus, they agreed with fee reduction. In contrast, when the participants who received the procedural request anticipated having to justify their judgment (the accountability condition), they were reluctant to ignore the TBD testimony, and therefore, these participants opposed the proposal of fee reduction. This pattern of findings is consistent with the suggestion that accountability sensitizes decision makers to the goal they want to satisfy (Lerner & Tetlock, 2003). The goal of making an appropriate judgment dictates that TBD information should be disregarded when there are good reasons for doing so, but should not be disregarded when the reasons are unconvincing.

Table 1: Participants agreement with fee reduction

	No accountability	Accountability
Use all information (including TBD argument)	-.69	-.32
Substantive request to discount	.43	.59
Procedural request to discount	.23	-.31

Note: More positive numbers indicate more agreement

Summary

Taken together, the three studies described above are consistent with the suggestion that the task of discounting could be done through a deliberate process of correction whereby decision makers assess their potential biases and their incentive to avoid them. Discounting succeeds when (1) people have a good estimate of the size and direction of the bias that the TBD information induces, (2) they are motivated to remove the bias, and (3) they have cognitive resources to do so.

In the reminder of this chapter we offer three extensions to the conceptualization of discounting as deliberate correction. First, we apply findings from research on the processing of negations in an attempt to shed light on the cognitive activity enacted when people discover that a particular message should be discounted. In particular, we examine the role of an alternative schema for the success of discounting. Second, we examine the properties of the outcome measures (e.g., judgment of guilt). Specifically we emphasize differences between corrections done superficially, at a response level, and reinterpretation done at the level of representation. Finally, we discuss the readiness of individuals to cope with the challenge of invalid messages. We argue that some mental states, and particularly a state of distrust, facilitate successful handling of false

information. Facilitation is manifested not only by discounting success, but also by the spontaneity of the discounting process.

Negation and discounting

The correction perspective discussed above suggests that under well specified conditions people might be able to undo the bias brought about by the TBD information. However, as we noted earlier, the correction is superficial – it is done at the response (judgment) level rather than through reinterpretation of the information. The research on negation is informative about the potential for reinterpretation of the information.

Instructions to discount tell a person that a particular claim, X, is suspect (or even false), and thus, it should be ignored. To illustrate, consider a description of a political candidate. After learning about her positions on 10 issues you are informed that the information about the third issue comes from an unreliable source and therefore you should disregard it when evaluating her. What do respondents do when they are asked to ignore the third issue? Ideally, respondents should place themselves in an alternative world in which they reprocess the information without receiving the TBD issue. Indeed, a control condition which is used as a benchmark for discounting success had just this format. Unfortunately, however, the alternative-world scenario is only possible in a between-respondents design. For better or worse, our mental system is affected by past exposure, so that the bias that the TBD information causes to the interpretation of the TBU information during the original encoding cannot be undone by merely asking people to reprocess the information as if the TBD has never been shown (Ecker, Lewandowsky,

Swire & Chang, 2011; Schul & Burnstein, 1985; Schul & Mayo, 1999; Schul & Mazursky, 1990).

As an illustration for the nature of this bias in interpretation, imagine that Jim is applying for the position of copywriter at an advertising agency. The members of the selection committee are considering two recommendations about Jim. One indicates that Jim is hardworking, while the other states that Jim is uncooperative. Normally, when two testimonies about Jim are processed, they are interpreted jointly so that their meanings become interdependent (Schul, Burnstein, & Martinez, 1983; Schul & Mayo, 1999). To illustrate, the positive characteristic "hardworking" might be fitted with the testimony about Jim's uncooperativeness by interpreting hardworking as a somewhat negative characteristic, projecting, for example, an image of a person who is not willing to learn from others and as a result often has to rediscover the wheel. Accordingly, if later the respondents consider "hardworking" by itself (e.g., because the testimony about "uncooperativeness" has been declared inadmissible) their judgments become overly unfavorable compared to respondents that received only the "hardworking" recommendation. Parenthetically, simply trying to reprocess the TBU information without doing anything about the TBD information is like trying to avoid thinking about a white bear (Wenzlaff & Wegner, 2000) – it is not likely to work.

This bias in encoding reflects the selection of meaning of the TBU information, which occurs during the early encoding of the information. In this sense it resembles a primacy effect. Can such a bias be nullified when people are told to re-encode the information as if the TBD has never been presented? The answer is "yes." It can be achieved through correction processes, as we have suggested earlier. But our focus in this

section is on effects that have to do with the interpretation of the TBU information. Specifically, we ask about reinterpretation, namely, whether the original interpretation afforded to the TBU information by the TBD information can be changed upon learning that the TBD testimony is false. Our lesson from the research on negation is that the answer to this question is "it depends." Successful reinterpretation depends on the nature of the TBD testimony and the way it is encoded.

Prior to discovering that the TBD testimony is false, one thinks about the TBD testimony as an affirmation - a testimony that is phrased in a positive way. An affirmation (abstractly, "A is X") tends to activate the core of the message (X) with its associations. For example, the assertion "John is intelligent" activates associations of intelligence and the assertion "Jim is hardworking" triggers associations of industriousness. But what happens when the request to discount is introduced. For example, one is told that the testimony about John's intelligence is based on a completely invalid test. Some people may infer that the opposite of the assertion is true; that is, that John is stupid. Others may entertain both possibilities; that is, the possibility that the opposite of the assertion is true, and the possibility that the original assertion is true. Still, there are cases (see below) in which receivers may maintain the original set of associations. The research on negations offers predictions about the prevalence of these alternatives and their consequences for the challenge of discounting.

Mayo, Schul, and Burnstein (2004) explored the kind of associations that come to mind when one process negations. To illustrate, imagine being told that "John is simply not a romantic person." Do you think about associations that are congruent with the intended meaning of the negation (e.g., unromantic gestures that John makes), or

associations congruent with what is being negated (e.g., romantic gestures that John doesn't make)? Mayo et al (2004) distinguished between two types of negation processes: fusion and schema-plus-tag.

Negation through a process of fusion is performed by activating an affirmative schema that entails the meaning of the negated message. For example, upon being told that "John is not smart," receivers may activate spontaneously inferences and implications that are congruent with being stupid. Thus, when negating according to the fusion model, receivers are able to accommodate the intended meaning of the negation as a whole. Note, however, that a necessary condition for the utilization of the fusion model is having in mind an affirmative alternative schema that entails the meaning of the negation.

Negation through the schema-plus-tag model is different. The receiver does not access an opposite schema, but rather, represents the negation as "A is Not(X)." For example, John is Not(romantic). Here, one thinks of romantic associations and negating each of them. Consequently, a boomerang effect might occur (Mayo et al, 2004). Because receivers activate during comprehension associations that are opposite to the intended meaning of the negation (e.g. romantic), in the long run receivers of the negated description might remember the description as if it had not been negated (e.g., "John is NOT(romantic)" is remembered as "John is romantic"). In short, whereas the fusion model of negation leads to reinterpretation (e.g., instead of thinking about not-intelligent, one thinks about stupid), the schema-plus-tag model, in contrast, resembles correction. One thinks about being romantic, and adds a negation marker – a mental instruction to modify the judgment.

The schema-plus-tag encoding is particularly likely when one negates a unipolar description, namely, in the case of negation of messages that have no clear alternative schema. To illustrate, consider the message "John harassed the secretary." What is the alternative schema? Not harassing can take many forms and no form is particularly dominant. Therefore, upon being asked to negate the message receivers are likely to activate various associations of harassment and negate each of them. Consequently, they will actually have multiple associations and inferences related to harassment in mind. Negation markers often become dissociated from the core attribute. When this happens, John would be incorrectly remembered as someone who *did* harass the secretary. It is important to note that most negative behaviors are unipolar, as there is no clear-cut alterative schema to represent their negation.

Returning to the issue of discounting success, we can rephrase our question: Can receivers reinterpret the TBU information when they process the request to discount the TBD information? We have highlighted the challenge of reinterpretation which stems from bias in the interpretation of the TBU information that the TBD information induces. Our theoretical analysis suggests that reinterpretation can occur when an alternative interpretation of the TBD testimony is readily available. The research on negation suggests that an alternative schema might be activated when the receiver thinks about the discounting request according to a fusion model, that is, with a schema that can accommodate the alternative of the TBD information. Accordingly, we propose that discounting of messages that have clear opposites (bipolar messages) is more likely to be successful than discounting of messages that do not (unipolar messages). This is because thinking about the negation of the unipolar TBD message brings to mind inferences that

are congruent with that message and incongruent with its negation. Therefore, in such cases the likelihood of reinterpretation of the TBU information during discounting becomes small.

The analysis of negation might have important implications for the well-known "sleeper effect" (Cook, & Flay, 1979; Mazursky & Schul, 1988; Pratkanis, Greenwald, Leippe & Baumgardner, 1988). Research shows that a persuasive message attributed to an untrustworthy source is completely discounted in the immediate judgment condition. However, when the impact of such a message is not measured immediately, the message is dissociated from the source and discounting fails. Our analysis suggests that when the message which comes from an untrustworthy source can be interpreted within a well-defined schema with an alternative meaning (as in the case of bi-polar negations), a sleeper effect would be less likely to occur than when the untrustworthy source provides a uni-polar message. In that case, the immediate negation of the message may still leave behind inferences associated with the message rather than its negation. The later dissociation of the message and the source, therefore, is likely to bring about a strong sleeper effect.

Finally, let us explicitly caution the reader that having the alternative schema is not a sufficient condition for successful discounting. Rather, we consider discounting as a struggle between competing schemata so that the schema that has an accessibility advantage at the time of judgment wins. To illustrate this struggle, we showed participants in a recent experiment two versions of the same face that differed in one feature: One version had narrow eyes, while the other version had round eyes. Past research suggests that narrow (vs. round) eyes tend to be associated with

untrustworthiness (Schul et al, 2004; Zabrowitz et al 1997). After being shown one version of the face, participants were shown the other version, and told how the version of the face they were seeing differed from the original version, and how it might affect their impressions from the faces. After a short filler task, all participants were shown one of the two versions of each face, that is, either the original or the modified. They were instructed to move forward if the original version was trustworthy and to move backwards if the original face was untrustworthy. The findings revealed that participants' movements were dominated by the face they were seeing at the time of movement rather than by the original version of the face. This failure, however, was not a failure of memory. Participants were highly accurate when they were asked to identify which of the two faces was the original face. We interpret these findings to mean that when people are making judgments they tend to rely on the most highly accessible schema information (in the above research, the face they see). Accordingly, in the battle between what they were seeing and what they knew, the former won. The lesson to those who wish to trigger an appropriate discounting is clear. It is not sufficient to highlight the nature of the bias that should be undone in order to undo the impact of a TBD testimony. Rather, the alternative has to be fully available and made at least as accessible as the TBD version.

Implicit and explicit senses of truth

The outcome of the battle between the two interpretations of the TBU information, one that contains the meanings implied by the TBD and the other without it, might be influenced by feelings of truth with respect to the TBD version. We have already noted this when we discussed the importance of the reason given for discounting.

That is, discounting is less successful when it is motivated by procedural concerns than when the reason is substantive (e.g., Kassin & Sommers, 1997; Saar & Schul,1996; Sommers & Kassin, 2001). The procedural/substantive comparison highlights a reasoning process: Because a request on procedural grounds implies that the TBD testimony might be actually true, whereas a request on substantive grounds implies that it is false, substantive requests lead to more successful discounting. However, as the research described below suggests, the feeling of truth might impact the ease of discounting through non-reasoning processes as well.

Let us start with a simple example. Assume that you try to ascertain if Bob told you that you are self-centered. Did he actually do it? Is it your imagination? Research conducted within the Source-Monitoring framework (see review in Johnson, 2006; see also Johnson & Raye, 1981) suggests two routes through which questions of this sort might be answered. Truth might be uncovered by reasoning. You might consider the situation involving Bob's statement, the details of interaction, your reaction to Bob, and the provocation that triggered his unkind assertion. You may also try to compare this statement to Bob's past behaviors toward you and others, and/or to the ways people other than Bob evaluate you. In a sense, you are trying to determine whether the assertion is reasonable within whatever else you know about the interaction. Research on reasoning (Evans 2008; Johnson Laird 2006) and causal attribution (see Jaspers et al, 1983) provide important insights into the systematic reasoning processes that allow people to evaluate the truth value.

Still, in trying to determine whether it is true that Bob made the assertion, you might also be influenced by the properties of the active representation, based on a toolbox

of heuristic rules (Gigerenzer, 2001). These properties do not involve the content of the information. Instead, they are based on features of the representation that the mental system learns implicitly while processing externally-generated facts and internally-generated fiction. The Source-Monitoring model suggests that people capitalize on these properties in separating truth (memories reflecting facts) from falsehoods (internally-generated memories). Thus, for example, rich memory representations, or an absence of cues about the operations that gave rise to the representation, might cause perceivers to err and evaluate an internally-generated image as an actual or "real" experience (e.g., Johnson, et al, 1993; Fiedler et al, 1996).

One of the strongest heuristic cues for inducing a bias of truth is the fluency of processing (e.g., Begg et al 1992; Hansen, et al, 2008; Henkel & Mattson, 2011; Winkielman, et al, 2003). It has been repeatedly shown that other things being equal, statements which are easy to process (e.g., due to perceptual or conceptual facilitation) are rated as more true than less fluent statements. This effect ties well with the findings of Mayo and Schul discussed above. It is much easier to process the face you see than to reconstruct a face from memory. Such differences may make the seen face "more real" allowing respondents to react to it as if it were true.

The research briefly reviewed above indicates conditions and processes responsible for the failure of people to separate truth from fiction. Shidlovski, Schul and Mayo, (under review) have recently begun investigating a complementary question. Assume we focus only on events that have been recognized explicitly as false; do these events vary in their propensity to feel as true? Stated differently, can an explicitly false event feel like a true event?

The short answer is "yes." Shidlovski et al. investigated imagined events and assessed their truth value both explicitly and implicitly. Explicit truth was assessed by judgments on a true/false continuum, as is typically done in research on veracity. Implicit truth was assessed in a variant of the autobiographical Implicit Association Test (aIAT; see Sartori, et al., 2008) that has been developed recently as a lie-detection tool. This procedure enabled us to test the extent to which it is easier to associate sentences about an imagined event with true sentences than with false ones and, therefore, the extent to which it is easier to associate the sentences about the imagined event with truth rather than with falsehood. Note that the implicit truth value (ITV) of events indicates the perceiver's tendency to categorize events that are implicitly true together with other events solely on the basis of their truth value. Without going into details, it was found that the imagined event (as well as an event that was actually experienced) gave rise to a higher ITV compared to a similar event that was not experienced or imagined. Significantly, this effect occurred even when participants acknowledged explicitly that the imagined event was false. Finally, it was found that the influence of the imagination manipulation on the ITV is mediated by the vividness of the representation of that event.

At the most general level, the dissociation between implicit and explicit senses of truth implies that perceivers who cognize (explicitly) that a specific act was false might still be influenced by it as if it was true; and conversely, people who acknowledge something as true might be unable to accept it as such and react to it as if it was false. Accordingly, the distinction between the explicit and the implicit senses of truth may help us understand the huge array of phenomena in which people behave as if they are inconsistent or irrational.

In particular, discounting can fail because people may feel that the narrative which contains the TBD is implicitly truer, even though they acknowledge explicitly that the TBD is false. To cope with this, one can either make the alternative narrative – the one without the TBD – feel more implicitly true, or one can weaken the tendency of decision makers to rely on their feelings of implicit truth (e.g., Pham, Lee, & Stephan, 2012).

Trusting versus distrusting states of mind

Our introduction refers to the duality of motivations that participants in social interactions have: They need to cooperate with each other, and at the same time they need to protect themselves from being exploited by the other. The former need is associated with trust, the latter with distrust. In this section we propose that the mental states of trust and distrust trigger cognitive processes that have opposite implications for the success of discounting. Specifically, trust impairs successful discounting and distrust facilitates it.

Trust connotes safety and transparency; individuals believe there is nothing to be feared in transactions between them and others. Distrust, in contrast, is associated with the perception that the other person intends to mislead the perceiver (Schul, Mayo, Burnstein, & Yahalom, 2007). Therefore, unlike situations that trigger trust, when people distrust they attempt to search for signs that the other's behavior is departing from what is normal in the situation and prepare to act upon finding out that deception had occurred.

What are the implications of this to the thought processes triggered under trust and distrust? Other things being equal, when a state of trust is active, one tends to believe, to follow the immediate implications of the given information. As a result,

information is encoded in integrative fashion, whereby early information influences the narrative within which later information is being processed. Moreover, when they trust, perceivers do not question their gut reactions. They trust not only others, but also their internal cues (Schul, Mayo, & Burnstein, 2008). Accordingly, trust might impair successful discounting for two main reasons. First, it enhances integrative encoding, which makes it difficult to separate the TBD from the TBU information. Second, it leads one to trust gut feelings and in particular the implicit sense of truth of the narrative that contains the TBD testimony.

In contrast, when a state of distrust is active, one tends to search for alternative interpretations of the given information. This spontaneous activity is a generalization of receivers' habitual responses to the situation of distrust which is associated with concealment of truth (cf., Fein, 1996; Schul, Burnstein & Bardi, 1996; Schul et. al., 2008). Thus, in distrust, the mental system becomes more open to the possibility that the ordinary schema typically used to interpret the ongoing situation may need to be adjusted.

We (Schul, Mayo, & Burnstein, 2004) investigated these conjectures by comparing contexts of trust versus distrust with respect to the associative links they activated in processing messages. It was predicted that when receivers trust they bring to mind thoughts that are congruent with the message. In contrast, when receivers distrusted they tend to look for hidden or non-routine associations, which are typically incongruent with the message. This prediction was tested using single words as messages and priming facilitation to indicate the associative structure activated in response to a prime word. We triggered trust or distrust by showing faces that were associated either with trust or with distrust. We found, as predicted, that the standard congruent priming effect was flipped in

a distrust context: When a prime word appeared together with a face which signaled *distrust*, it facilitated associations that were *incongruent* with it. Thus, incongruent target words were facilitated more than congruent target words (e.g., "light" activated "night" more than "dark" activated "night"). The opposite pattern was found in the context of trust: Now the prime activated associations that were *congruent* with it (e.g., "dark" activated "night") more than associations that were incongruent with it (e.g., "light" activated "night"). This has been extended by Mayer and Mussweiler (2011) and generalized to non-verbal stimuli by Schul et al, (2008).

It should be noted that the states of trust and distrust differ not only in terms of their impact on encoding processes, but also in the motivational forces that they trigger. Suspicion and distrust may raise the need to discern and identify falsehoods from truth, or, put differently, the concern for information accuracy. Accordingly, individuals concerned with information accuracy (i.e., under distrust) may seek to find out whether a witness has the ability and the incentive to report accurately, or whether the testimony fits with other reports. Moreover, trust and distrust may also differ with respect to concerns for outcome accuracy. In particular, compared with conditions of trust, conditions of distrust might trigger a greater concern with judgment accuracy. Concern with judgment accuracy can lead to increased information search, to a stronger tendency to analyze the information systematically, and to a higher likelihood of being influenced by the diagnosticity of the information (Chaiken et al, 1989; Kruglanski, 1989; Thompson et al., 1994).

The differences in encoding and in motivation suggest that a state of distrust (vs. trust) allows receivers to discount information more appropriately for several related

reasons. First, during encoding, those who distrust may encode messages with incongruent as well as congruent associations. By creating narratives that contain multiple interpretations of the message, which are either consistent with the given information or are inconsistent with it, receivers can prepare for discounting. In this sense, a state of distrust might function as a trigger for spontaneous negation, whereby the given messages are encoded together with alternative schemas that entail their negations. Such encoding prevents the tight associative structure created by integrative encoding and therefore allows more successful discounting. Second, those who distrust might have higher motivation for veridical processing of messages and for arriving at unbiased judgments. Indeed, Schul, Burnstein and Bardi, (1996) showed that people who were warned about the possibility of being misled discounted information more successfully than those who were not warned (see also, Ecker, Lewandowsky, & Tang, 2010).

Final notes

More often than not, discounting fails. The mental system seems to prefer construction to reconstruction. Accordingly, as the literature on belief perseverance shows, interpretations tend to stick, even when the evidential basis of them is undermined (see Guenther & Alicke, 2008 for a fuller discussion). The research we described above suggests some reasons for this phenomenon. We tend to create mental structures that are well integrated, and in doing so we try to account for everything that we know. We are not very good with introspecting and assessing the impact of individual messages or cues on our judgments, and we often do not care that much about being accurate. When we do

correct, however, we act superficially; namely, we use various shortcuts or heuristic rules in trying to remove a bias at a response level. Although such correction might provide decision makers with a sense of competence in being able to control biases, the biases may surface if measured by alternative means.

One may consider replacing the attempts at correction by attempts at reinterpretation. The research described above offers several ways in which reinterpretation might be achieved. However, it should be noted that attempts at reinterpretation may also induce a bias. The admonition to ignore the testimony about Bill's stupidity should not be taken as a license to assume Bill's smartness. Assuming the opposite might do as much injustice as assuming the original interpretation.

The challenge of proper discounting, therefore, involves finding a way to lead decision makers to think in a more complex way, to entertain both possibilities: Bill might be smart or stupid, one does not know. Such complex thinking requires one to delay arriving at closure, to be tolerant of ambiguities, and to resist resolving inconsistencies (Kruglanski et al, 2006). In a world of information overload and high time pressure, habitual decision-making strategies tend to work in the opposite way, to lead to immediate resolution of inconsistencies and to early freezing of conclusions. No wonder, therefore, that discounting often fails.

References

Anderson, C.A., New, L.B., & Speer, J.R. (1985) Argument availability as a mediator of social theory perseverance. *Social Cognition, 3*, 235-249.

Begg, I.M., Anas, A., & Farinacci, S. (1992) Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General, 121*, 446-458.

Chaiken, S., Liberman, A., & Eagly, A.H. (1989). Heuristic and systematic processing within and beyond the persuasion context. In J.S. Uleman & J.A. Bargh (Eds.) *Unintended thought.* (pp. 212-252). New York: Guilford Press.

Cook, T.D., & Flay, B.R. (1978) The persistence of experimentally induced attitude change. In L. Berkowitz (Ed*.), Advances in experimental social psychology* (Vol. 11). New York: Academic Pres. (pp 1-57)

Demaine, L.J. (2008) In search of an anti-elephant: Confronting the human inability to forget inadmissible evidence. *George Mason Law Review, 16*, 99-140.

DePaulo, B.M. & Kashy, D.H. (1998) Everyday lies in close and casual relationships. *Journal of Personality and Social Psychology, 74*, 63-79.

DePaulo, B.M., Kashy, D.H., Kirkendol, S.E., Wyer, M.M., & Epstein, J.A. (1996) Lying in everyday life. *Journal of Personality and Social Psychology, 70*, 979-993

Dyer, R.F., & Kuehl, P.G. (1978), A Longitudinal Study of Corrective Advertising. *Journal of Marketing Research, 15*, 39-48.

Eagly, A.H., Wood, W., & Chaiken, S. (1978). Causal inferences about communicators and their effect on opinion change. *Journal of Personality and Social Psychology, 36*, 424-435.

Ecker, U.K.H & Lewandowsky, S. Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction *Psychonomic Bulletin and Review, 18*, 570–578.

Elliott, R., Farrington, B., Manheimer, H. (1988) Eyewitnesses credible and discredible. *Journal of Applied Social Psychology, 18*, 1411-1422.

Evans, J. St. B.T (2008) Dual-Processes accounts of reasoning. *Annual Review of Psychology, 59*, 255-278

Fein, S. (1996) Effects of suspicion on attributional thinking and the correspondence bias. *Journal of Personality and Social Psychology, 70,* 1164-1184

Feldman, R.S., Forrest, J.A., & Happ, B.R. (2002) Self presentation and verbal deception: Do self-presenters lie more? *Basic and Applied Social Psychology, 24*, 163-170

Fiedler, K, Walther, E., Armbruster, T., Fay, D., & Naumann, U. (1996) Do you really know what you have seen? Intrusion errors and presuppositions effects on constructive memory. *Journal of Experimental Social Psychology, 32*, 484-511

Gigerenzer, G. (2001) The Adaptive Toolbox: Toward a Darwinian Rationality. In French, J. A., Kamil, A. C., & Leger, D. W. (Eds.). (2001). *Evolutionary psychology and motivation (Nebraska Symposium on Motivation*, Vol. 48, pp. 113–143). Lincoln: University of Nebraska Press.

Golding, J.M., & Hauselt, J. (1994) When instructions to forget become instructions to remember. *Personality and Social Psychology Bulletin, 20*, 178-183.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J. Morgan (eds.), *Syntax and Semantics.(Vol 3)* New York: Academic Press. (pp. 43-58).

Guenther, C.L., & Alicke, M.D. (2008) Self-enhancement and belief perseverance. *Journal of Experimental Social Psychology, 44,* 706-712.

Hansen, J., Dechêne, A., & Wänke, M. (2008). Discrepant fluency increases subjective truth. *Journal of Experimental Social Psychology, 44*, 687–691.

Hartwig, M., & Bond, C.F. (2011) Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin, 137*, 643-59

Hatvany, N., & Strack, F. (1980) The impact of a discredited key witness. *Journal of Applied Social Psychology, 10,* 490-509.

Henkel, L.A., & Mattson, M. E. (2011). Reading is believing: The truth effect and source credibility. *Consciousnsess and Cognition, 11*, 1705-1721

Jaspars, J., Fincham, F., & Hewstone M. (Eds.) (1983) *Attribution theory and research: Conceptual, developmental and social dimensions*. London: Academic Press.

Johnson, M.K. (2006). Memory and reality. *American Psychologist, 61*, 760-771.

Johnson, M.K., Hashtroudi, S., & Lindsay, D.S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3-28.

Johnson, M.K., & Raye, C L. (1981). Reality monitoring. *Psychological Review, 88*, 67-85.

Johnson-Laird, P.N. (2006) *How We Reason*. Oxford: Oxford University Press

Kassin, S.M., & Sommers, S.R. (1997) Inadmissible testimony, instructions to disregard, and the jury: Substantive versus procedural considerations. *Personality and Social Psychology Bulletin, 23*, 1046-1055.

Kassin, S.M., & Wrightsman, L.S. (1980) Prior confessions and mock-jury verdicts. *Journal of Applied Social Psychology, 10*, 133-146

Kassin, S.M., & Wrightsman, L.S. (1981) Coerced confessions, judicial instruction, and mock juror verdicts. *Journal of Applied Social Psychology, 11*, 489-506.

Kruglanski, A. W. (1989). The Psychology of being "right": On the problem of accuracy in social perception and cognition. *Psychological Bulletin, 106*, 395-409

Kruglanski, A.W., Pierro, A., Manetti, L. & DeGrada, E. (2006). Groups as epistemic providers: Need for closure and the unfolding of group centrism. *Psychological Review, 113*, 84-100

Lerner, J.S., & Tetlock, P.E. (1999) Accounting for the effects of accountability. *Psychological Bulletin, 125*, 255-275.

Lerner, J.S., & Tetlock, P.E. (2003). Bridging individual, interpersonal, and institutional approaches to judgment and choice: The impact of accountability on cognitive bias. In: Schneider S, Shanteau J (Eds.) *Emerging Perspectives on Judgment and Decision Research*. Cambridge: Cambridge University Press.(p. 431-457).

Martin, L.L., Seta, J.J., & Crelia, R.A. (1990). Assimilation and contrast as a function of people's willingness and ability to expand effort in forming an impression. *Journal of Personality and Social Psychology, 59*, 27-37.

Mayer, J. & Mussweiler, T. (2011) Suspicious spirits, flexible minds: when distrust enhances creativity. *Journal of Personality and Social Psychology, 101,* 1262-1277.

Mayo, R., Schul, Y., & Burnstein, E. (2004) "I am not guilty" versus "I am innocent": Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology, 40,* 433-449.

Mazursky, D., & Schul, Y. (1988). The effects of advertisement encoding on the failure to discount information: Implications for the sleeper effect. *Journal of Consumer Research, 15*, 24-36.

Pham, M.T., Lee, L., & Stephen, A.T. (2012) Feeling the Future: The Emotional Oracle Effect *Journal of Consumer Research, 39*, 461-477

Pratkanis, A.R., Greenwald, A.G., Leippe, M.R., & Baumgardner, M.H. (1988) In search of a reliable persuasion effect: III. The sleeper effect is dead. Long live the sleeper effect. *Journal of Personality and Social Psychology, 54*, 203-218.

Prendergast, G., Liu, P.Y., Poon, D.T.Y (2009) A Hong Kong study of advertising credibility, *Journal of Consumer Marketing, 26*, 320 – 329

Saar, Y., & Schul, Y (1996). The effect of procedural versus substantive instructions to disregard information. Unpublished manuscript

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological science: a journal of the American Psychological Society*, *19*, 772-80.

Schul, Y. (1993). When warning succeeds: The effect of warning on success of ignoring invalid information. *Journal of Experimental Social Psychology, 29*, 42-62.

Schul, Y., & Burnstein, E. (1985). When discounting fails: Conditions under which individuals use discredited information in making a judgment. *Journal of Personality and Social Psychology, 49*, 894-903.

Schul, Y., & Burnstein, E. (1998). Suspicion and discounting: Ignoring invalid information in an uncertain environment. In J.M. Golding & C. MacLeod (Eds.) *Intentional forgetting: Interdisciplinary approaches*. Erlbaum. (pp. 321-348).

Schul, Y., Burnstein, E., & Bardi, A. (1996). Dealing with deceptions that are difficult to detect: Encoding and judgment as a function of preparing to receive invalid information. *Journal of Experimental Social Psychology, 32*, 228-253.

Schul, Y., Burnstein, E., & Martinez, J. (1983). The informational basis of social judgments: Under what conditions are inconsistent trait descriptions processed as easily as consistent ones? *European Journal of Social Psychology, 13*, 143-151.

Schul, Y., & Goren, H. (1997). When strong evidence has less impact than weak evidence: Bias, adjustment, and instructions to ignore. *Social Cognition, 15*, 133-155.

Schul, Y., & Mayo, R. (1999) Two sources are better than one: The effects of ignoring one message on using a different message from the same source. *Journal of Experimental Social Psychology, 35*, 327-345.

Schul, Y., Mayo, R., & Burnstein, E. (2004) Encoding under trust and distrust: The spontaneous activation of incongruent cognitions. *Journal of Personality and Social Psychology, 86*, 668–679

Schul, Y., Mayo, R., & Burnstein, E. (2008) The value of distrust. *Journal of Experimental Social Psychology, 44,* 1293-1302.

Schul, Y., Mayo, R., Burnstein, E., & Yahalom, N. (2007) How people cope with uncertainty due to chance or deception. *Journal of Experimental Social Psychology, 43*, 91-103.

Schul, Y. & Manzury F. (1990). The effect of type of encoding and strength of discounting appeal on the success of ignoring an invalid testimony. *European Journal of Social Psychology, 20*, 337-349

Schul, Y., & Mazursky, D. (1990). Conditions facilitating successful discounting in consumer decision making: Type of discounting cue, message encoding, and kind of judgment. *Journal of Consumer Research, 16*, 442-451.

Schwarz, N., Sanna, L., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in Experimental Social Psychology, 39,* 127-161

Schyns, P., & Koop, C. (2010) Political distrust and social capital in Europe and the USA. *Social Indicators Research, 96*, 145-167

Shidlovski, D., Schul, Y., & Mayo, R. (under review) If i can imagine it, then it happened: The implicit truth value of imaginary representations.

Sommers, S.R., & Kassin, S. (2001). On the many impacts of inadmissible testimony: Selective compliance, need for cognition, and the overcorrection bias. *Personality and Social Psychology Bulletin, 26*, 1368-1377.

Steblay, N., Hosch, H.M., Culhane, S.E., & McWethy, A. (2006). The impact on juror verdicts of judicial instruction to disregard inadmissible evidence: A meta-analysis. *Law and Human Behavior, 30*, 469-542.

Strack, F., & Hannover, B. (1996) Awareness of influence as a precondition for implementing correctional goals. In P.M. Gollwitzer & J.A. Bargh (Eds). *The psychology of action: Linking motivation and cognition to behavior*. New York: Guilford (pp. 579-596).

Thompson, E.P., Roman, R.J., Moskowitz, G.B., Chaiken, S., & Bargh, J.A. (1994). Accuracy motivation attenuates covert priming: The systematic reprocessing of social information. *Journal of Personality and Social Psychology, 66*, 474-489.

Thompson, W.C., Fong, G.T., & Rosenhan, D.L. (1981) Inadmissible evidence and juror verdicts. Journal of Personality and Social Psychology, 40, 453-463.

Wenzlaff, R. M., & Wegner, D. M. (2000). Thought suppression. In S. T. Fiske (Ed.), *Annual review of psychology* (Vol. 51, pp. 59-91). Palo Alto, CA: Annual Reviews.

Wilson, T.D., & Brekke, N. (1994) Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin, 116*, 117-142.

Winkielman, P., Schwarz, N., Fazendeiro, T., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 189–217). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wolf, S., & Montgomery, D.A. (1977) Effects of inadmissible evidence and level of judicial admonishment to disregard on the judgment of mock jurors. *Journal of Applied Social Psychology, 7*, 205-219.

Zebrowitz, L.A. (1997). *Reading faces: Window to the soul?* Boulder, CO: Westview Press.