Draft 19/02/2012

MODELS OF COMMUNICATION, EPISTEMIC TRUST AND EPISTEMIC VIGILANCE

Mikhail Kissine & Olivier Klein
Université Libre de Bruxelles

1. Introduction

Since most social interactions involve routine use of language, one of the questions that stand prominently on the agenda of social psychology is how people come to believe what they are told. More particularly, it is the bread and butter of persuasion research. When one inspects this voluminous literature (for a recent review, see Albarracin & Vargas, 2010), it appears that experimenters have mainly relied on situations in which people are presented with explicitly persuasive statements: typically, a text advocating a previously unpopular standpoint, measure or behaviour (such as eating giblets, raising tuition fees or shutting the air conditioner). But we also easily believe what we are told even though the speaker has not been pursuing a persuasive goal. When your neighbour evokes the persistent rain during his holidays in the Périgord, you will probably unquestioningly consider his description of the weather as accurate, and so even if you were yourself in Indonesia at the time and had no idea whatsoever about the weather in France. In such mundane examples, although the speaker is not pursuing any specific persuasive strategy, for the listener, believing the communicated information is a routine activity that constitutes the fabric of social interaction — and makes it possible. And yet, in spite of their importance to social life, such ordinary instances of belief validation have largely fallen out of the scope of social psychology.[1]

The most straightforward issue such unexceptional validation processes raises is that of the connection between grasping the content of a statement and believing it. Obviously, this distinction is not only conceptual as one can mentally represent the reference of false statements (e.g., Brussels lies under Mediterranean sun) while knowing that they are false. The question is rather how hearers switch from one to the other. Is grasping without believing always prior to believing? Or do we automatically believe whatever we understand, so that realizing that a statement is false entails 'unbelieving' it?

One might expect this issue to occupy a central place in linguistics, and more specifically

[1] An exception is the work on communicational grounding (Clark, 1996), which considers how people elaborate a 'common ground' to pursue cooperative projects. However, in this kind of research the focus is more on an interpersonal level of analysis, viz. on how people manage to incorporate new knowledge through communicational behaviour. The cognitive underpinnings of this 'incorporation', by contrast, are hardly considered.

in pragmatics, a subfield devoted to the study of language *use*. But many, if not most, linguists adhere to a dominant view, inherited from the work of Paul Grice (1989), according to which verbal understanding follows a tortuous inferential route, based on assumptions about the speaker's communicative intentions. As we will see below, such a take on utterance interpretation entails that any communicated information has to be assessed before being translated into belief. This consequence has largely gone unquestioned, except by a minority who argue that, just as you (generally) believe what your eyes see, you (generally) believe what you are told until proven otherwise. Recently, however, proponents of the classic, Gricean position have undertaken to actively endorse and defend the hypothesis that there is no communication-based beliefs without prior 'filtering' (Sperber, et al., 2010).

While it remains unsettled to a large extent, the question of the validation of communicated information is thus central both to social psychology and to pragmatics. In this paper, we will attempt to clarify the debate. The position that we will defend is essentially one of the minority view. Yet, while we like to defend the meek, we will do so in a nuanced way. The main claim of the paper will be the following. Acquiring beliefs from speech is as direct as perception; however, this process may be mediated by a series of domain-specific, and independent filters, so that in some cases information is rejected without having been previously integrated among the hearer's beliefs. In the next section, we will expose in more detail the two competing models of communication, the inferential and the direct perception one, and focus on their consequences regarding vigilance towards communicated information. The relationship between communication and epistemic vigilance can be assessed from at least two standpoints. First, there exists experimental research that impacts on this issue. Second, any claim about such a link should be evolutionarily plausible. In Section 3, we propose a critical survey of the relevant experimental evidence, and argue that existing data is incompatible with the inferential model. Next, we will argue that, from an evolutionary point of view, the most plausible model of communication is that of direct perception, gradually supplemented with various epistemic filters. Finally, we will flag two important programs of research in social psychology — the 'saying is believing effect' and Schwarz' work on the role of conversational inference in judgment — which could be profitably reinterpreted in the light of the direct perception model.

2. Two models of communication

Let us now consider the two conflicting pragmatic models that we mentioned above. The first — and by far the most widespread one — stems from Grice (1989). The key idea is

that interpretative processes can be reconstructed as an attribution of complex communicative intentions to the speaker. Although it not clear that Grice himself conceived of this inferential mechanism as a psychologically valid model, and not as a mere rational reconstruction (see Saul, 2002), it has been subsequently promoted as a cognitive claim. The most paradigmatic transposition of Grice's ideas into a cognitive model is Sperber and Wilson's (1995) Relevance theory. According to Sperber and Wilson, the hearer (H) infers the meaning communicated through a linguistic utterance by attributing to the speaker (S) the informative intention to make manifest to H a certain piece of information *p*. H infers the informative intention he attributes to S by attributing her the communicative intention to make S's informative intention manifest to H. In other terms, H assumes that the literal meaning communicated by S's utterance is *p* because H attributes to S the intention to make H believe that S has the intention that her utterance causes H to believe that *p*. In what follows, we will refer to this model of interpretation as the Inferential Model — IM, for short.

The IM puts communication apart from other information channels, such as perception. Visual perception is direct, and devoid of any epistemic gap. When you see that there is a chair in front of you (a) you do not come to the conclusion that there is a chair in front of you by inference from the processes — such as the reflection of the light on your retina, your spatial position, etc. — underlying your perceiving the chair, (b) your believing that there is a chair in front of you is not preceded by an internal deliberation about your acceptance or not of this information (although you can subsequently reassess this belief in the light of other information, and, perhaps, eliminate it from your 'belief box'). In other words, nothing comes between the visual experience of the chair and the belief that there is a chair over there; visual perception is direct.[2] By contrast, according to the IM, extraction of meaning from an utterance is inherently indirect. From this theoretical standpoint, understanding that the content of S's utterance is *p* amounts to grasping S's informative and communicative intentions. An extra step is needed to arrive at the belief that *p*; understanding that a speaker wants us to believe that *p* does not automatically causes us to believe that *p*. The output of the interpretation mechanism is limited to the content of S's communicative intentions, and besides this, it does not provide any information about the world. Integrating

[2] To be sure, with some optical illusions — those that you recognise as such — you do not believe what you see. But note that conscious effort is needed: you eliminate the belief you acquired. Moreover, understanding that what you see is an illusion requires explicit training, or at least, a time-consuming detailed examination of the stimulus from different points of view and/or through different perceptual modalities.

communicated content within one's 'belief box' would thus never be automatic. As emphasized by Sperber et al. (2010), this means that 'epistemic vigilance' is part and parcel of the IM. Believing or not the content that has been communicated depends on a filtering mechanism of some sort, which checks received information for accuracy and consistency with other beliefs. In absence of epistemic assessment, grasping the communicated content falls short from leading to the belief that it is true.

The second view of information transfer through language we will examine can be called the 'Direct Perception Model' (DPM, henceforth). According to the DPM, the mechanisms that allow hearers to derive the literal meaning of an utterance are sub-conscious and as direct as those underlying visual perception — that is, they are not adequately modelled as inferences to what the utterance means. When you see a car in front of you, you do not perceive the proximal stimuli, such as the light stimulation of the retina, which constitute your visual experience; similarly, according to DPM when told that *p*, you directly form the belief that *p* without any Gricean reasoning about the speaker's intentions. A central prediction of the DPM is therefore that once the contents of communicative stimuli are grasped, they do not need to go through another assessment mechanism to get into the interpreters' 'belief boxes'. As far as the cognitive mechanisms underlying the retrieval of literal meaning go, you believe everything that you are told.

It is worth emphasizing from the outset that the DPM does not imply that no epistemic barrier filters the communicated contents that can get into the 'belief box'. The crucial difference between the DPM and the IM is that for the IM the gap between interpretation and belief is presupposed by the cognitive mechanisms assumed to underpin belief acquisition from linguistic stimuli, whereas for the DPM any epistemic filtering of hearsay information is independent from the interpretation process.

3. Experimental evidence

The DPM has a respectable, if somehow marginal, tradition in philosophy (see, for instance Burge, 1993; Millikan, 2004, p. chapter 9; 2005, pp. 207-219; Recanati, 2002), but there also exists an experimental side to the debate. An important set of empirical evidence that may be invoked in favour of the DPM comes from experiments by Gilbert and colleagues (Gilbert, Krull, & Malone, 1990; Gilbert, Tafarodi, & Malone, 1993). The aim of these experiments was to evaluate two competing models of belief acquisition. According to the first, dubbed 'Cartesian' by Gilbert, validation of a statement is never concomitant with its comprehension: a 'filter' — some kind of internal deliberation — is necessary before

endorsing communicated content as a valid description of the world. It is clear that the IM is very similar, at least in spirit, to this Cartesian model. The contrasting, 'Spinozean', model predicts that any belief is acquired automatically; if there is assessment, it takes place after information has been integrated within the 'belief box'. The DPM is fully compatible with the Spinozean view.

A central prediction of the Spinozean model is that belief rejection thus operates post hoc and is an effortful process. By contrast, the Cartesian model predicts that subjects can prevent a proposition from getting within the belief box, and that, if no deliberation takes place, being told that *p* will never result in the belief that *p*. Since a central prediction of the DPM is that acquiring information through the communicative channel is not intrinsically mediated by epistemic filtering, results favouring the Cartesian model would disprove the DPM.

In a paradigmatic study confronting these two models (Gilbert, et al., 1990: study 1), participants read statements about the meaning of word in an unknown language (Hopi, such as, e.g., 'monishna is a star'). After an 8-second presentation of each statement, the words 'true' or 'false' appeared on the screen. For some of the statements, participants had to perform a secondary task (i.e., responding to a tone), which mobilized additional cognitive resources. In a subsequent, testing phase, participants were presented a list of statements (most of which were semantically equivalent to those presented in the learning phase) and asked to identify their truth-value. The variable of primary interest is the rate of correct recognition as a function of the truth of the statements and the presence of an interfering task in the learning phase. When identifying statements that had not been interrupted during the learning phase, people did not make more errors on false than on true statements. This secondary task did not influence performance on true statements either but, crucially, it led to more errors on false statements, which were more often identified as true than in the absence of such interruption. According to Gilbert et al.'s reasoning, the latter result suggests that cognitive resources are necessary to correct the default endorsement of the sentence's content as true. By taxing these resources, the secondary task prevents such correction from taking place. Such a pattern of findings cannot be properly explained if a 'Cartesian' filter operates between the encoding of the statement and a (hypothetical) subsequent judgment of truthfulness.

In this initial set of studies, it is unclear whether participants' judgments reflected their memory of the information presented in the learning phase or actual belief in such information. In their second set of studies, Gilbert et al. measured more reliable indicators of

actual beliefs. For example, in one experiment, they presented information about a defendant in the context of a criminal case, such that some bits of this material were explicitly tagged as false. Furthermore, the information thus communicated was either exonerating or aggravating for the defendant. In addition, half of participants had to simultaneously perform a secondary task, whereas the other half were not interrupted. In the interrupted condition, when aggravating information was false, participants judged the defendant more harshly (i.e., by proposing a more severe penalty) than when exonerating information was false. This was not the case when they were not interrupted. Thus, when interrupted, people assimilated the false information and failed to correct it. Taken together, these findings suggest that people not only encode false information as true when their cognitive resources are taxed but that they act on it.

As is suggested by the title of the 1993 paper ('You can't not believe everything you are told'), Gilbert et al. seem to consider that people are incapable of suspending belief. As hearers, we would be capable to consider statements as false only after having previously endorsed them. This — extreme — position has been challenged by Hasson et al. (2005). A straightforward concern about Gilbert's experiments is that, from the participant's point of view, false statements used in the experiment (such as the ones about meaning of Hopi words) are uninformative: accordingly, knowing that they are false (e.g. that it is not true that 'moshina' means 'star' in Hopi) has no informative value for the participant. But of course being told that a statement is false may prove informative per se. For instance, when a statement like 'Tom is generous' is tagged as false, the participant may directly encode the information 'Tom is greedy'. That is, it is not the statement itself that would be encoded — as a strict application of the Spinozean model would suggest — but rather an inference drawn from its being tagged as false. If this happens, the sentence 'Tom is generous' should be readily identified as false in the second phase of the experiment even when cognitive resources are depleted. In order to test this hypothesis, Hasson et al. replicated Gilbert et al.'s first experiment but manipulated the informativeness of false statements. They reproduced the same results when the statements where uninformative when false (viz. the distracting task during exposure phase led to consider false statements as true) but, critically, when the false statements were informative, recognition performance was not altered by the secondary task — recalling the truth-value of informative statements is not affected by parallel cognitive overload.

In a second study, Hasson et al. primed participants with sentences describing a person (e.g., 'John is generous') whose photo was presented. Either these stimuli were tagged as 'true'

or as 'false', or no truth-value was provided. Next, participants had to perform a lexical decision task on a target word that was either related to the true version of the sentence (e.g., 'warm') or to the false version (e.g., 'rapacious'). If people directly encoded all sentences as true — as the Spinozean model predicts — the prime sentence should facilitate the recognition of true-related words regardless of this sentence truth-value. However, it turned out that people were more likely to correctly identify truth-related targets when the prime sentence was true than when it was false or when its veracity unknown. This is consistent with the assumption that people did keep track of false sentences as such (possibly in the form of a meaningful inference) during the priming phase. For false-related targets, the prime sentence veracity exerted no effect, which suggests that false sentences did not activate semantic content consistent with their implications[3].

Another challenge to Gilbert's claims has been posed by Richter et al. (2009). Drawing on research in psycholinguistics, these authors claim that people routinely (i.e., in the absence of an explicit goal) and effortlessly rely on validation processes when comprehending sentences. Contrary to both the Spinozean and the Cartesian view, it would be impossible to divorce validation from comprehension. Validation, argue Richter and colleagues, is grounded in background assumptions (stored either in working or in long-term memory) related to the topic at hand. Since the material used in Gilbert's experiments consisted in statements about unknown topics, lack of any relevant background may have therefore prevented the participants from performing routine validation. Richter et al reproduced Gilbert et al.'s 'Hopi' experiment, but with statements half of which were perceived to be true/false with a high certainty in a pilot study, while the truth-value of the other half of stimuli was seen as uncertain, in the same pilot study. Richter et al. replicated Gilbert's pattern for the latter group of stimuli (interruption during the learning phase yielded weaker recognition performance for false, but not true, statements). However, they also found that for statements with strong background beliefs (be they true or false), interruption did not affect performance.

In a more direct test of the presence of an implicit validation, Richter et al. (exp. 3 & 4) have relied on an 'epistemic Stroop' paradigm, in which subjects had to evaluate (by rapidly clicking on one of two buttons) the spelling of words belonging to sentences that were either consistent or inconsistent with strong background beliefs. Richter et al. assume that, if routinely triggered, belief validation should interfere with orthographical judgments. Thus,

[3] It is worth noting that when participants were presented with statements whose truth-value was undetermined, they took longer to respond to false-related targets than to true-related targets. This result is actually in line with a Spinozean theory.

people should experience difficulties both to approve the spelling of words within statements that contradict strong background beliefs and to disapprove the spelling of words within statements that conform to such beliefs. These predictions were corroborated on measures of error rates and reaction times. Participants made fewer errors and (in experiment 4) took less time to respond when words within true sentences were correctly spelled and when words within false sentences were incorrectly spelled than in the two incongruent conditions (viz. correct spelling with false statements and misspelling within true statements).

Hasson et al.'s and Richter et al.'s findings suggest that the radical version of the Spinozean view is hardly tenable: in some cases false information is not rejected a posteriori, but filtered straight at the entrance of the 'belief box'. However, that a filter is present in some context does not imply that it is a necessary condition for acquiring hearsay beliefs. If anything, the studies by Hasson et al. and Richter et al. suggest boundary conditions for the operation of a filter. But this falls short from invalidating the DPM. What remains uncontroversial about Gilbert's results is that epistemic filtering is not inherent in communication: in certain circumstances, you believe directly what you hear (or read). That this happens when the communicated message is irrelevant or does not clash with background beliefs does not change anything to the fact that epistemic vigilance is optional in acquiring hearsay beliefs. The indubitably important finding that epistemic filtering is sub-conscious, routine, and automatically triggered in certain conditions is of no use for the IM. What the IM predicts is that there is no believing of communicated meaning without epistemic check — and this is not the case.

To insist, that we are endowed with largely automatic and efficient filtering mechanisms is beyond doubt. We have just seen that epistemic filtering is triggered whenever one's salient beliefs are contradicted or when information has a high degree of relevance. In the same vein, one person's facial characteristics contribute to the assessment of trustworthiness after an exposure as short as 100 milliseconds (Todorov & Willis, 2006) which may influence the degree at which the validity of the communicated information will be checked.

This view of epistemic filtering as optional is coherent with a classic dual pathway model of belief validation (cf. Evans, 2008): the Spinozean, automatic, route to belief validation is followed unless specific conditions are present, in which case a more controlled process of epistemic vigilance comes into play. These conditions may pertain to the content of the utterance or to the cognitive and/or motivational state of the judge (e.g., through cognitive load).

This line of thought receives support from data on the ontogenesis of epistemic vigilance. The capacity to assess the reliability of a communicator is quite a precocious one. From the age of four, children are capable to discriminate between a reliable and unreliable puppet; when facing a choice between these two sources, they tend to trust the reliable one (Clément et al. 2004). Likewise, four-year-olds tend to distrust a puppet characterised as a liar (Mascaro and Sperber 2009). Yet, this capacity is by no means part and parcel of the processing of communicative behaviour (as the IM would have it). To begin with, studies by Clément et al., (2004) and by Mascaro and Sperber (2009) also revealed that at the age of three children fail to adopt such selective trust. Furthermore, Vanderbilt et al. (in press) show that explicitly identifying an adult as an unreliable deceiver in three consecutive communicative exchanges does not prevent four-year-olds from trusting the information communicated by this same person just afterwards.

The point is not that young children are blindly gullible: that they are not is revealed, for instance, by the fact that children below five years tend to privilege first-hand, perceptual information over verbal claims made by an adult (Robinson, Mitchell, & Nye, 1995).[4] However, while such findings unveil precocious mastery of effective heuristics for managing conflicting information, there is no evidence that acquisition of various 'sceptic' strategies is inherent to the development of the capacity to interpret communicative stimuli.

3. The evolution of epistemic vigilance: some speculations

A major finding of the neo-Darwinian paradigm has been that cooperative behaviour — and, even, to a certain extent, altruism — proves evolutionarily rewarding. It is also widely accepted that such strategies must encompass a mechanism aimed at the exclusion of non-cooperative 'cheaters' from interaction (e.g. R. Axelrod & Hamilton, 1981; R. M. Axelrod, 1984; Cosmides & Tooby, 1992; Dawkins, 1989; Dennett, 1995; Kitcher, 1993). It is also widely accepted that cooperative behaviour can be found, in some form or another, all over the animal kingdom. Particularly striking are the findings that to a great extent social relations among big apes are ruled by expectations of cooperation and ostracism of cheaters (e.g. De

[4] Children below four are selective about informational medium (visual or tactical perception, or hearsay), and in cases where two sources provide contradicting information, they favour the most reliable one; however, they have difficulties in reporting correctly the source of their beliefs, which suggests that once a belief is acquired no trace of its provenance subsists (Gopnik & Graf, 1988; for a related discussion, see Millikan, 2005, pp. 209-210; Mitchell, Robinson, Nye, & Isaacs, 1997; Whitcombe & Robinson, 2000).

Waal, 2006). In this light, it is not too risky a conjecture that human communication emerged among groups whose members could already reasonably expect each other to be helpful, and from where cheaters were ostracized (with all the disastrous consequence this entailed).

Let us indulge now in some 'just-so-story' kind of speculation. Imagine two different groups: the direct perceivers and the inferentialists. Members of both groups can communicate to share information, but while the former acquire information from speech following the DPM, the latter have to go through the inferential strategy, posited by the IM. Assume, furthermore, that members of each group pass their interpretative strategy to their offspring.

It seems obvious that within the kind of cooperative niche just described, direct perceivers would be clearly advantaged over inferentialists. Since any communicated content is directly added to the direct perceivers' belief boxes, they would acquire information much faster and in a more effortless way than inferentialists, who need, every time, to go through assessment and consistency checking before taking what they have been told on board. Provided that communicators are benevolent and competent, communicated information will be accurate often enough to privilege direct perceivers, because accurate information acquisition would then mobilize fewer resources than those needed by the inferentialists — resources which can thus be profitably allocated to another task. In such an environment direct perception through speech would be evolutionary stable. Sub-populations endowed with it would rapidly take over individuals that cannot communicate altogether, — and over a hypothetical inferentialist, Cartesian sub-group.

However, being a direct perceiver makes interactions with unreliable speakers very costly. The DPM predicts there is a great risk that the contents of misleading statements will automatically get into the direct perceiver's belief box, so that an exclusion or assessment process will be necessary (at risk of letting a false belief influence her decisions). But such processes take time and cognitive energy that could be better employed. Therefore, when the quantity of false utterances exceeds a certain threshold, direct perceivers would become disadvantaged with respect to inferentialists. Inferentialists, remember, never take communicated information in before checking it for accuracy and consistency. Once we form a certain belief, this belief is likely to influence other beliefs, desires and action plans; therefore, cancelling it is likely to entail a costly domino effect of revisions and checks. In other words, suspending one's judgment to accuracy check, as the inferentialists do, is more efficient than reassessing a content previously believed to be true, as the direct perceivers have to do when they realize that they have been misinformed.

Even though it is very plausible to assume that these hypothetical ancestors of ours evolved in small groups, bound by kinship and in-group cooperation links, direct perceivers were exposed to two sources of misinformation. First, encounters with deceivers, ones issued from other groups or in-group members who adopted uncooperative strategies during certain competitive circumstances, remained possible; second, even within their own, reliable group, direct perceivers must have had to count with benevolent but mistaken communicators. That is, direct perceivers were advantaged only if they had developed independent means to overcome deception and unintentional misinformation.

Regarding the first type of risk, the best evolutionary strategy whatsoever is clearly to supplement the mechanism that ensures quick and effortless integration of communicated information, with efficient and automatic epistemic filters that activate vigilance with respect to certain speakers (or in certain conditions). Such an evolutionary scenario entails that epistemic vigilance is not of one piece — it is a patchwork of adaptive strategies, shaped by heterogeneous environmental pressures. The supplementation of interpretive mechanisms with epistemic filtering is a classical case of what Krebs and Dawkins (1984) call the 'evolutionary arms race', concomitant to the development of communication systems. The aptitude to inform begets misinformation and deception, which, in turn, increases the adaptability of filtering mechanisms. Such an adaptation would be hard to explain in absence of environmental pressure to control the ingress of communicated information within the 'belief box'. It is precisely because the cognitive processes that allow us to interpret utterances as conveying informative contents do not come with an inherent epistemic safe-guard that such filtering mechanisms have been selected.

To be sure, the domain-specific epistemic filters still do not shield direct perceivers from misinformation from their benevolent fellows, viz. from being misinformed in a kind of cooperative situation where no specific vigilance should be triggered. Recall that misinformation has a higher cost for direct perceivers than for inferentialists. Therefore, the result is that, *ceteris paribus*, it is better for direct perceivers to avoid interaction with unreliable speakers altogether. Communicating false information, intentionally or not, should be seen as a non-cooperative behaviour, worth of ostracism from the group.

At this point, it may be objected that this last feature of our evolutionary scenario renders the DPM evolutionary implausible after all, for under such a view, speaking seems to be quite a risky business. If saying something false is assimilated to non-cooperative behaviour, punishable by exclusion from interaction — with all the dramatic consequences this entails — the most evolutionarily stable strategy would be to remain silent unless one is

absolutely certain about the truth of her utterance (for a related discussion, see Hurford, 2007, pp. 276-277).

However, a line of thought that gained popularity in recent years views precisely this risk as the cornerstone of the emergence of human communication. Under such view, being prone to exchanging information — with the risk of being mistaken — has a higher evolutionary value than remaining silent in most cases, and thus avoiding any risk to be treated as a cheater (for a similar point, stated in more general terms, cf. Kitcher, 1993). One such model is defended by Dessalles (1998), according to whom by providing reliable (and relevant) information speakers seek to increase their social prestige, and hence, their reproductive success. Another, compatible, position is Miller (2000, p. chapter 10), who argues that the human propensity to communication is explained in great part by sexual selection: verbal display raises the chances for mating.

Dessalles's and Miller's rationales assume that communicating has a certain cost; otherwise, no prestige would be attached to such behaviour. Both authors appeal to what is known as the Handicap Principle (Zahavi & Zahavi, 1997). According to this principle, some traits that constitute a *prima facie* handicap can provide the organism that displays them with a higher chance of reproduction. Roughly, the idea is that by exhibiting a handicap the individual demonstrates ability to survive despite this handicap, which, in turn indicate a high degree of fitness. For instance, male bowerbirds build elaborate bowers of twigs whose only use is to serve as a stage for courtship displays. The more adorned and big this 'stage' is, the more is its construction energy consuming, but also the higher is the likelihood for the 'builder' to be chosen by a female to copulate. The female uses bowers as an indication of fitness in order to choose a mate. A male that can waste time and energy to build a big and elaborate bower, at the expense of looking for food, is likely to be more fitted to the environment than the one that cannot afford such a costly behaviour.[5]

The crucial component of the Handicap Principle, and a leitmotiv of Zahavi and Zahavi (Zahavi & Zahavi, 1997) book, is that in order to be a reliable indicator of fitness, the handicap must be hard to fake — otherwise any individual, not only the fittest one, could afford it. If building complex bowers were not handicapping, even an otherwise unfit male could afford it. (Accordingly, because of the lack of correlation between bowers' size and decoration and fitness, this kind of display would gradually drift out.) Therefore, if, as Dessalles (1998) and Miller (2000) claim, providing information through speech benefits the speaker by increasing

[5] Handicaps may also serve to deter predators; see Zahavi and Zahavi (1997, p. chapter 1).

social prestige and sexual attractiveness, the speaker's task should not be easy — otherwise, there would be no reason for attributing reliable speakers an increased social rank.

Now, in communities of direct perceivers one such risk is obvious: unreliable speakers run the danger of being excluded from interaction as cheaters, with all disastrous consequences linked to ostracism (cf. Williams, 2005). In other words, the DPM predicts the emergence of the policing mechanism that theories of language evolution based on the Handicap Principle need to get of the ground.[6]

The IM imposes to view the apparition of linguistic communication as a twofold and simultaneous evolutionary emergence of a new channel and a new way of information acquisition. Accordingly, it needs to posit a double — and simultaneous — environmental pressure to explain the emergence of linguistic communication: one that explains the selection of complex communicative behaviours, and another that selects for an inferential acquisition strategy, with an inherent gap between understanding and believing. The DPM, by contrast, views language as a new channel to feed information in the belief box in exactly the same way as perception. This, in itself, makes the DPM more plausible from a phylogenetic point of view. To be sure, hearsay beliefs are not as reliable as perception based ones. But, as we have argued in this Section, linguistic communication proves maximally efficient when appended with domain specific epistemic filters.

3. Impact on social psychology

Before concluding this paper, we would like to evoke some central issues in social psychology that may benefit from a critical assessment of the IM. Grice's theory of meaning is probably the most important influence from pragmatics on social psychology. It has been used to (re)interpret research on a variety of cognitive biases. One such example is base rate neglect (Kahneman & Tversky, 1973): participants receive information regarding a person displaying traits typical of a social category A (e.g., engineers) or of a stereotypically opposite category B (e.g., lawyers): for example, 'Jack loves mathematical puzzles' would be construed as more typical of engineers than lawyers. Participants are asked to estimate the likelihood that the target person belongs to category A. In addition, the target is presented as drawn from a sample containing either a majority of members of social category A (e.g., 70 %

[6] The social prestige associated with speaking cannot count in itself as the origin of a policing mechanism resulting in the exclusion of misinformants. Truthful verbal communication increases the speaker's evolutionary fitness; it follows that both no display at all and non-truthful verbal communication entail lack of increase in one's prestige. Consequently, if the gain of social prestige were the source of truth-commitment, we should expect that remaining silent deserves as much punishment as lying — which is not so (also Hurford, 2007, p. 293).

engineers vs. 30 % lawyers) or only a minority (e.g., 30 % engineers vs. 70% lawyers). Typically, people's estimations of the target's membership are little affected by this statistical information: Rather, it is the stereotypicality of the target that explains most of the variance. Tversky and Kahneman (1974) famously analyzed such biases as the effects of simple and frugal heuristics (e.g., representativeness) by opposition to more elaborate and 'rational' calculations. However, Norbert Schwarz (1994) pointed out that the conversational context in which these biases arose had been neglected. Conversational moves are governed by certain expectations — among which the ones Grice identified as 'conversational Maxims' —, and these are often implicitly violated in experimental settings. Typically, hearers expect speakers' contribution to be relevant; but in the 'base-rate paradigm', described above, a central part of the experimenter's contribution, viz. the target's stereotypical traits, should not be relevant if participants were to behave 'rationally'. This effect can be eliminated, or attenuated, when conversational expectations are neutralized. Thus Schwarz, Strack, Hilton, and Naderer (1991) showed that when information about the target was presented as selected by a computer, rather than a psychologist, participants' estimation of the target's membership is more influenced by statistical information. This is so because we usually do not expect computers to be sophisticated enough to select all, and only, relevant information about the psychological profile of a person; hence, not every bit of information is automatically taken to be relevant.[7] This kind of Gricean explanation has been conclusively applied to a great number of other experimental paradigms (for reviews, see: Holtgraves, 2010; Schwarz, 1994).

While fully Gricean in spirit, this analysis of cognitive biases is fully consistent with our rejection of the IM. In the foregoing, we claimed that epistemic vigilance is not inherent in our capacity to retrieve information from speech. What Schwarz's results reveal is that how communicated information is integrated, and hence influences other beliefs and decisions, depends on the context (e.g. the nature of the problem, the identity of the speaker, etc.). That some information selected by a computer is not taken into account does not show that, when the same information is provided by a psychologist and is used within a decision process, it necessary undergoes inferential epistemic filtering.

A program of research in social psychology that coheres even more with the DPM is research on the saying-is-believing effect (Higgins & Rholes, 1978). In this paradigm, participants read a description of a person (the target). This description is crafted in such a way that the statements composing it can be interpreted as reflecting either relatively

[7] Furthermore, the opposite pattern of results obtained when the problem was presented as statistical, and not psychological (presumably, computers are better with statistical tasks than psychologists).

desirable or undesirable traits. Participants are then asked to describe the target to an audience (who is already acquainted with this person) in order to allow this audience to identify the target. Crucially, speakers are informed of the audience's attitude, which can be either favourable or unfavourable to the target. Unsurprisingly, communicators tend to describe the target in a more flattering light when the audience holds a positive than a negative attitude. More interestingly, however, is the fact that when communicators' memory for the target is probed later, it appears that their own memory is biased as well and in the same direction. This does not happen in a control condition in which speakers are exposed to the audience's attitude but do not have to communicate. Decades of research on this phenomenon have led to consider it as driven by a desire to establish a shared understanding of the target (i.e. shared reality) with the audience (Echterhoff, Higgins, & Levine, 2009). This shared reality is contingent on trust in the audience, and especially in its capacity to form an accurate opinion of the target. When this trust is present, there seems to be no barrier to simply incorporating her attitude into one's own. In other words, when trusting an audience, her view of the target seems to translate into believing that this view is correct. However, this process seems to be mediated by the active construction of an understanding of this target through communication. In other words, everything happens as if once verbalized, descriptions of the target influenced by the audience's attitude come to be perceived as true. This is a very reasonable prediction that can be made from the DPM; presumably, no filter is activated for one's own statements, which, therefore, automatically integrate the speaker's belief-box.

Another phenomenon that speaks to the DPM is belief perseverance (Anderson, Lepper, & Ross, 1980). In research on this topic, people are presented with 'facts' that are later discredited. Yet, people keep believing more in these 'facts' than control subjects who have not been exposed to them. For example, in a study by Ross, Lepper, and Hubbard (1975), participants received false feedback about their performance on a task. They were later divided in three conditions as a function of the debriefing they received. In one condition, they were informed that this feedback was random and fabricated by the experimenter. In a second, they were additionally told that the purpose of this feedback was to study the perseverance of beliefs in the face of false information. In the third (control), not outcome was provided. Participants were later asked to estimate their skill at performance the task for which they had received false feedback (i.e., identifying fake suicide notes). As expected, participants in the control condition estimated their skill as higher when having received a positive than a negative feedback. Interestingly, however, participants in the first (outcome)

condition were also influenced by the feedback, although it was irrelevant. This suggests that people accepted the experimenter's feedback at face value. Obviously, this finding is not necessarily incompatible with an inferential model since what is at stake here is not, belief validation but how one can come to 'disbelieve' information that has been initially accepted as true. However, the difficulty to undo such beliefs suggests that they are incorporated relatively automatically, without conscious control (as a Spinozean model would rather suggest).

In a similar vein Douglas and Sutton (2006) found that people exposed to information consistent with conspiracy theories regarding Lady Diana's death (e.g., 'One or more rogue "cells" in the British secret service constructed and carried out a plot to kill Diana.') later tended to overestimate the extent to which they were influenced by such theories, compared to others (i.e., this what is called the 'third person effect'). It thus seems that, in spite of their inclination to distrust information they received, participants were influenced by it, which would be perfectly in line with the DPM. In this respect this paradigm poses an even greater explanatory problem to IM that the SIB or the belief perseverance paradigm, for which trust in the audience or in the experimenter respectively seem to play a crucial role.

5. Conclusion

When the white coated experimenter in Stanley Milgram's classic experiments explained to his subjects that one them would be the 'pupil' and the other 'the teacher', he expected his subjects (1) to understand what he said but also (2) believe it as a truthful account of the situation. That the description was well understood was taken for granted (actually, if this condition were not fulfilled, the study would be considered ill-designed) and generally independent of the second part, which was of a much more interest to Milgram, as it is this very belief that made murder possible. Social psychologists tend to consider comprehension and validation as two independent processes and are actually much more interested in validation than in comprehension, which is best left to (psycho)linguists. What we hope to have shown above is that it is impossible to remain agnostic about the cognitive processes underlying utterance interpretation if one undertakes to explain how communication-based beliefs are formed. Grice's work helped researchers to realize that communication is an inter-subjective activity, whose many aspects are influenced by expectation of cooperativeness. However, it is a mistake to adopt the reconstruction of speaker's meaning in terms of intention attribution, operated by Grice, as a model of language comprehension. Such a theoretical choice forces one to posit that no belief can be drawn from linguistic stimuli

without having gone through an epistemic filter. We have argued that this consequence is hard to accommodate with available experimental data. Much more plausible, both from empirical and evolutionary point of view, is the direct perception through language model. Contrary to the extreme position the proponents of this model might have claimed, it does not compel us to assume that no information can be rejected without having being previously held as true. The cognitive equipment that allows us to acquire hearsay beliefs is supplemented with a variety of epistemic filters, which may be easily and automatically activated under certain conditions. A promising direction for future research is to get a better understanding of the typology of these filters. Meanwhile, it seems fair to conclude that epistemic vigilance is not inherent in our capacity to understand others' statements.[8]

References

Albarracin, D., & Vargas, P. (2010). Attitudes and Persuasion: From biology to social responses to persuasive intent. In S. T. Fiske, D. T. Gilbert & G. Lindzey (Eds.), *Handbook of Social Psychology* (Vol. I, pp. 394-427). New York: Wiley.

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology, 39*(6), 1037.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*, 1390-1396.

Axelrod, R. M. (1984). *The evolution of cooperation*. New York: Basic Books.

Breheny, R. (2006). Communication and folk psychology. *Mind and Language, 21*(1), 74-107.

Burge, T. (1993). Content preservation. *The Philosophical Review, 102*(4), 457-488.

Clark, H. H. (1996). *Using language* (Vol. 4). Cambridge, UK: Cambridge University Press Cambridge.

Clément, F., Koenig, M., & Harris, P. (2004). The ontogenesis of trust. *Mind and Language, 19*(360-379).

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides & J. Tooby (Eds.), *The Adapted Mind. Evolutionary psychology and the generation of culture* (pp. 163-228). Oxford: Oxford University Press.

Dawkins, R. (1989). *The Selfish Gene. Second edition*. Oxford: Oxford University Press.

De Waal, F. (2006). *Primates and Philosophers. How morality evolved*. Princeton: Princeton University Press.

[8] Origgi and Sperber (2000) and by Sperber et al. (2010) invoke massive contextual dependence of the literal meaning to discard the DPM, and argue that verbal understanding necessarily conforms to IM. However, there are very strong reasons, mostly from typical and atypical language and cognitive development, to dismiss the idea that all such pragmatic processes require the complex mindreading posited by IM (see, Breheny, 2006; Kissine, 2012; forthcoming, p. chapters 3 and 5; Recanati, 2002).

Dennett, D. C. (1995). *Darwin's Dangerous Idea. Evolution and the meanings of life*. London: Penguin.

Dessalles, J.-L. (1998). Altruisim, status, and the origin of relevance. In J. R. Hurford, M. Studdert-Kennedy & C. Knight (Eds.), *Approaches to the Evolution of Language: Social and cognitive bases* (pp. 130-147). Cambridge: Cambridge University Press.

Douglas, K. M., & Sutton, R. M. (2006). The hidden impact of conspiracy theories: Perceived and actual influence of theories surrounding the death of Princess Diana. *The Journal of social psychology, 148*(2), 210-222.

Echterhoff, G., Higgins, E. T., & Levine, J. M. (2009). Shared reality. *Perspectives on Psychological Science, 4*(5), 496-521.

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: some problems in the rejection of false information. *Journal of Personality and Social Psychology, 59*, 601-613.

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology, 65*(2), 221-233.

Gopnik, a., & Graf, P. (1988). Knowing How You Know - Young Childrens Ability to Identify and Remember the Sources of Their Beliefs. *Child Development, 59*(5), 1366-1371.

Grice, P. (1989). *Studies in the way of words*. Cambridge, Mass.: Harvard University Press.

Higgins, E. T., & Rholes, W. S. (1978). "Saying is believing": Effects of mes- sage modification on memory and liking of the person described. *Journal of Experimental Social Psychology, 14*, 363-378.

Holtgraves, T. (2010). Social psychology and language: Words, Utterances, and Conversations. In S. T. Fiske, D. T. Gilbert & G. Lindzey (Eds.), *Handbook of Social Psychology* (Vol. I, pp. 1386-1422). New York: Wiley.

Hurford, J. R. (2007). *The Origins of Meaning*. Oxford: Oxford University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237-251.

Kissine, M. (2012). Sentences, utterances, and speech acts. In K. Allan & K. M. Jaszczolt (Eds.), *Cambridge Handbook of Pragmatics* (pp. 169-190). Cambridge: Cambridge University Press.

Kissine, M. (forthcoming). *From Utterances to Speech Acts*. Cambridge: Cambridge University Press.

Kitcher, P. (1993). The evolution of human altruism. *Journal of Philosophy, 90*(10), 497-516.

Krebs, J. R., & Dawkins, R. (1984). Animal signals: mind-reading and manipulation. In J. R. Krebs & N. B. Davies (Eds.), *Behavioural Ecology. An evolutionary approach. Second edition* (Second edition ed., pp. 380-402). Oxford: Blackwell.

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition, 112*, 367-380.

Miller, G. (2000). *The Mating Mind: How sexual choice shaped the evolution of human nature*. London: Heinemann.

Millikan, R. G. (2004). *Varieties of Meaning*. Cambridge, Mass.: MIT Press.

Millikan, R. G. (2005). *Language: a biological model*. Oxford: Oxford University Press.

Mitchell, P., Robinson, E. J., Nye, R., & Isaacs, J. E. (1997). When speech conflicts with seeing: Young children's understanding of informational priority. *Journal of Experimental Child Psychology, 64*(2), 276-294.

Origgi, G., & Sperber, D. (2000). Evolution, communication and the proper function of language. In P. Carruthers & A. Chamberlain (Eds.), *Evolution and the Human Mind: language, modularity and social cognition* (pp. 140-169). Cambridge: Cambridge University Press.

Recanati, F. (2002). Does linguistic communication rest on inference? *Mind and Language, 17*, 105-126.

Robinson, E. J., Mitchell, P., & Nye, R. (1995). Young children's treating of utterances as unreliable sources of knowledge. *Journal of Child Language, 22*, 663-685.

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology, 32*(5), 880-892.

Saul, J. M. (2002). What is said and psychological reality; Grice's project and Relevance theorists' criticisms. *Linguistics and Philosophy, 25*, 347-372.

Schwarz, N. (1994). *Judgment in a social context: Biases, shortcomings, and the logic of conversation*. San Diego, CA, US: Academic Press.

Schwarz, N., Strack, F., Hilton, D., & Naderer, G. (1991). Base rates, representativeness, and the logic of conversation: The contextual relevance of irrelevant information. *Social Cognition, 9*(1), 67-84.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., et al. (2010). Epistemic vigilance. *Mind and Language, 25*(4), 359-393.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and Cognition* (Second edition ed.). Oxford: Blackwell.

Todorov, A., & Willis, J. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592-598.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185*, 1124-1131.

Vanderbilt, K. E., Liu, D., & Heyman, G. D. (in press). The development of distrust. *Child Development*.

Whitcombe, E. L., & Robinson, E. J. (2000). Children's decisions about what to believe and their ability to report the source of their belief. *Cognitive Development, 15*(3), 329-346.

Zahavi, A., & Zahavi, A. (1997). *The Handicap Principle. A missing piece of Darwin's puzzle*. New York: Oxford University Press.