Trust in Relevance*

Fabio Paglieri, Cristiano Castelfranchi

Goal-Oriented Agents Lab, Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche, Roma, Italy
{fabio.paglieri,cristiano.castelfranchi}@istc.cnr.it

Abstract. Studies on trust in information sources have mostly focused on whether the source is capable of providing *correct* and *complete* information, thus overlooking another essential aspect of trust: the assessment of *relevance*. Information, even when true, is literally useless, unless it relates meaningfully to the informational needs and practical goals of its recipient. Moreover, relevance, exactly like truth, is not a self-evident feature of information: agents frequently realize whether or not an information was relevant for their purposes only ex post, and this is precisely why relevance, like truth, requires trust. The agent needs to be able to (i) rely on the relevance of the information provided by a source before or without being able to directly verify such relevance, and thus (ii) estimate the quality of sources also based on their ability to consistently deliver relevant (as well as correct and complete) information. In this paper we outline some desiderata for modeling relevance as one of the key features in deciding whether to trust an information source, we analyze its related role in determining belief formation and change, we detail how to assess relevance in order to avoid biases (e.g., giving systematic priority to good news over bad ones), we discuss whether relevance is a subjective or an objective feature of information (and in what sense), and we conclude by suggesting possible ways of implementing and/or formalizing trust in relevance for MAS, based on previous work on trust dynamics.

Keywords: trust, relevance, information sources, beliefs, goals

1. Introduction

Imagine asking a trusted friend to suggest some fancy recipe for a dinner, only to be answered as follows: "Today the weather in Bali is perfect for a day at the beach". Let us stipulate that the weather today is indeed splendid in Bali, and that such information covers all that there is to know about weather conditions there, as far as beach expeditions are concerned. So your friend cannot be faulted for being incorrect or reticent. Yet, you would certainly be disappointed by that answer, and rightly so. In fact, you have just been provided with an information that is not only unrelated to your actual query, but also useless with respect to any of your pragmatic concerns – assuming going to the beach in Bali today is not a feasible option for you, alas! In a nutshell, the problem with that answer is lack of *relevance*. This is crucial to determine the inform-

* AT2012, 15-16 October 2012, Dubrovnik, Croatia.

ational quality of the message, and thus, in retrospection, the overall trustworthiness of its source. In our example, in the absence of some mitigating circumstances (e.g., your friend failed to hear or misunderstood the question), such an irrelevant remark will backfire on the source, giving you reason to be less inclined to rely on your friend for advice in the future.

All of this is self-evident, and thus one would expect to find a substantial body of work on the role of relevance for trust. However, this is not the case: on the contrary, almost all existing theories and models of trust focus on correctness (if source *X* provides information *p*, *p* is true) and completeness (if *p* is true, source *X* provides information *p*) as indicators of source performance, without addressing the problem of relevance. While these factors have often been further analyzed in detail, e.g. distinguishing whether the source is mistaken or reticent due to ignorance (incompetence) or malice (insincerity), similar distinctions have not been raised with respect to relevance. This omission is striking, because there is no doubt that an information is valuable only if it is also relevant, in addition to correct and complete.[1]

Lack of interest for relevance would be partially justified if this was a self-manifesting feature of information: that is, if the relevance of a message was always immediately obvious to its receiver, without doubt or error. If that was the case, there would be nothing to trust or distrust regarding relevance of information, because there would be no knowledge gap and thus no need to rely on someone or something to bridge it: relevance, or lack thereof, would simply be known as soon as a proposition is uttered. Agents would still have to trust sources for being relevant, in order to choose which one to consult *before* being able to directly assess the relevance of their information. But relevance would not be a matter of trust with respect to information.

However, the relevance of an information is not so easy to ascertain – although it might be argued that it is more accessible than truth in most cases. We typically consult sources more knowledgeable than ourselves on the matter at hand, and their superior knowledge often includes also understanding what is important (for us) better than we do ourselves. This is certainly true for all sources in a tutorial position with respect to the receiver (e.g., parents, teachers, advisers, physicians), and it is often true also for sources in general. Besides, the most spectacular example of the need for trust in relevance is provided by information retrieval technologies. When we search something on Google, we never scan more than a tiny fraction of the available results, and we regularly give more weight to those highly ranked by the search engine – which uses, in fact, a relevance algorithm. A recent experimental study with eye tracking techniques (Pan et al. 2007) revealed that college students have substantial trust in the relevance of highly ranked Google results, even when the abstracts provided for these entries were less relevant to their query – that is, they confided in Google to know better than themselves what would satisfy their interests. More gener-

[1] Indeed, relevance, like correctness, is *always* essential to guarantee the quality of information for a receiver, whereas completeness matters only under certain conditions, e.g. an utmost need for precision and exhaustiveness. Moreover, it is plain that relevance is needed to provide an interesting characterization of completeness: in fact, for an information source "being thorough" does not mean randomly mentioning true information, but rather providing all the true information that matters with respect to the speaker's interests – that is, all true *relevant* information.

ally, the need for this kind of trust in relevance should be obvious in any information-intense environment.

Even if trust theory has largely neglected relevance, this notion has received substantial scrutiny in other areas: information theory has not provided any comprehensive account of relevance, as often lamented in the literature (e.g, Bremer & Cohnitz 2004; Floridi 2004, 2008), but specific approaches to communication have focused extensively on it – most notably, in information retrieval (e.g., Cooper 1971; Crestani et al. 1998; Borlund 2003), where relevance is analyzed as an answer to a (possibly underspecified) query, and in pragmatics (e.g., Sperber & Wilson 1995; Wilson & Sperber 2004; Borg 2005), where relevance is defined in terms of how much an utterance contributes to the ongoing dialogical interaction. It is beyond the purpose of this paper to provide a comprehensive review of the vast literature on relevance (but see Floridi 2008 for a brief and highly informative summary, and Cohen 1994 for more extensive discussion): suffice it to say that the problem of trust has never been central in these areas either.

So it seems we have unexplored ground to cover ahead of us. To try and make the best of it, we will proceed as follows: section 2 will detail our definition of relevance, contrast it with how this notion is interpreted in information theory and pragmatics, and address the related but different problem of relevance assessment, to wit, how to compute the relevance of an information; section 3 will explain why relevance is not to be confused with (or reduced to) the utility of an information, to avoid biasing agents towards good news, and because there are features of information that, albeit unrelated with relevance, still impact on its usefulness (e.g., indispensability); section 4 will argue that relevance can be either subjective or objective, depending on whether it is assessed in relation to the agent's goals (what we want) or to some objective interest (what we truly need), to show that this distinction plays an important role in modeling agents dynamics; section 5 will provide some suggestions on how to formalize and implement trust in relevance for multi-agent systems, given all the features highlighted in the rest of the paper and building on previous work in trust dynamics.

2. What relevance is, and how to assess it

It is fairly obvious that relevance is relative to some goal in general sense, e.g. a need, a desire, an intention, an objective, a concern, a passion, a task, and similar. Thus it would seem equally plain that only a grounded, systematic, and analytic theory of (implicit and explicit) goals can provide the basis for a theory of relevance. However, existing models of relevance endured significant pains to *avoid* any explicit reference to goals in their definition of relevance: this is true not only of pragmatics approaches (e.g., Sperber and Wilson notoriously define relevance as maximization of the ratio between benefits and costs in communication), but also of epistemological analysis of relevance. For instance, Floridi defends «a subjectivist interpretation of epistemic relevance […] based on a counterfactual and metatheoretical analysis of the degree of relevance of some semantic information *i* to an informee/agent *a*, as a function of the accuracy of *i* understood as an answer to a query *q*, given the probability that *q* might be asked by *a*» (2008, p. 69). While his approach has many merits, its considerable complexity mostly stems from the self-imposed restriction of not mentioning goals in

the definition of relevance, which forces him to resort to convoluted notions such as (roughly) "how good would I have considered this information as an answer to my query, given the probability that I had posed the query if I had been aware there was something valuable to be known". Whether or not one agrees with this analysis, it is worth wondering whether it would not be much simpler to define relevance in terms of goals, especially since goals seem to be lurking behind these definitions anyway – benefits obviously depend on goals (and the same is true for some costs, e.g. opportunity costs), and something is valuable to know only if it bears on the agent's goals.

Current theories of relevance are "shy about goals" for a simple reason, though: they lack a principled theory to *connect* goals to information. In the absence of that theory, stating that "relevance depends on goals" would of course be moot, because we would not know how to specify such dependence in ways that make relevance assessment tractable. Luckily, such a theory is available: goal-processing, that is, the cognitive process that turns a desire into a full-blown intention (Bratman 1987), has been analyzed as being based on beliefs (Castelfranchi & Paglieri 2007), which in turn are information that the agent considers credible enough to warrant acting upon (Paglieri 2004). So it is possible to define *an information as relevant for a goal if and only if it is a candidate for a belief that supports the processing of that goal* – and the set of goal-supporting beliefs is neither infinite, nor vague, nor context-dependent.

While full details of belief-based goal-processing are provided elsewhere (Castelfranchi & Paglieri 2007), here it suffices to mention all the possible uses that an information can have for the processing of a goal:

- activating the right goal in the right circumstances;
- evaluating whether the goal is already realized, self-realizing, or impossible;
- understanding that the goal depends on us, so we have to decide and act;
- detecting conflicts between active goals and thus prompting to choose;
- choosing between goals based on pros and cons;
- formulating/selecting a plan for the chosen goal, given "know how" and skills;
- performing the chosen action on the basis of the assumed enabling conditions;
- knowing that the goal has been realized and thus stop (or try again, or drop it), and then be satisfied or frustrated;
- understanding the reasons of the outcome (be it success or failure) and thus learn.

If (and to the extent that) an information satisfies one of these tests and steps in goal-processing, then it is relevant. So relevant means 'usable' for a given goal and its processing (but not necessarily 'useful', see later). However, while the notion of 'use' refers to all possible means and tools that are instrumental to or favorable for a certain outcome, 'relevant' mainly focuses on the epistemic dimension: on information as an instrumental good. 'Relevant' means 'useful to know, to be considered, to be taken into account' (not necessarily in a propositional or even in a conscious way). In other words, *relevance captures the evaluative dimension of information and knowledge* (Miceli & Castelfranchi 1989). Considering information in terms of its relevance implies evaluating it (explicitly or implicitly), assigning it some "*value*" that can only be derived from goals. This evaluation can be explicit, or consist of an implicit precedence given to relevant information by the design of the system, or even be an affective and intuitive 'attraction' for certain information (Castelfranchi 2000). It can be based also on theories about what constitutes useful information, and it can be

modulated by different 'standards' or levels of abstract knowledge characteristic of a given context (Floridi 2008).

A crucial consequence of defining relevance in relation to goals is that it vindicates the following intuition on source assessment: in order to be considered of high quality, a source must not simply provide appropriate answers to my queries, but also *proactively offer relevant information that were never asked*, and yet happens to be important for me. A good source is over-helping (Falcone & Castelfranchi 2001), which in this context means *over-answering*: providing more information than those initially required, or even in the absence of any explicit query. Over-answering is typically meant to provide relevant information beyond what was explicitly requested, possibly because the original query is considered by the source as irrelevant for the agent's real needs (see also section 4).[2] The source (i) recognizes a need for knowledge as a means to some end (or better *ascribes* you such goals, and bets on that), and (ii) decides to collaborate with your end, by providing you the necessary or useful information.

Imagine a tourist in Rome asking at the train station: "How much is a ticket to Naples?". The lady at the ticket counter replies: "It's 10 euros. But they have changed the platform for the next train; it leaves from platform 16 now. You have to hurry, it leaves in 5 minutes". Here the source is not only providing the required information, but also informing the agent of many other facts that she deems important for his (implicit and inferred) goal of "speedily going to Naples by train". Crucially, the source would not be considered as being really helpful if she just offered the *required* information, and failed to offer all the other *relevant* information she knew – even if the tourist did not or could not ask for it. In fact, what is explicitly requested may even be irrelevant, as when the tourist (intending to travel today to Naples) asks "Where I can find a train schedule?", and the answer is "Today the railway personnel is on strike, so there are no trains leaving from this station". And of course the source could also be wrong in ascribing a goal to the agent, thus (unintentionally) providing irrelevant information by over-answering: for instance, if in the previous example the tourist had the goal of leaving the day after tomorrow instead of today, being informed of the ongoing strike would be useless, in spite of the best intentions of the source.

3. Relevance is not utility

In light of the close relationship between goals and relevance, it is tempting to think of the latter as a measure of the *utility of information*, loosely understood as the capacity of that information to foster the agent's goals. In this view, the more an information promotes the agent's current agenda, the greater its relevance for that agent. This intuition is certainly appealing, but ultimately misleading, for two different reasons.

Firstly, equating relevance with the utility of information could make agents exceedingly biased in relevance assessment towards "good news", that is, information that agree with the agent's plans and concerns. For instance, given the goal of "hiking tomorrow", the information "tomorrow will rain" is certainly disappointing for the

[2] This applies only to benevolent sources: a malicious source could of course use over-answering as a strategy to muddle the issue and mislead the other agent. However, we speculate that in human society cooperative over-answering is much more frequent than its malicious counterpart, and that people count on it and assess sources also based on it.

agent, in that it pressures to revise or even abandon the original goal; in contrast, the information "tomorrow will be sunny" is definitely pleasing. Intuition tells us that both information are equally relevant, and yet defining relevance in terms of "goal advancement" would make good news systematically more relevant than bad news.

This bias is not only hard to justify on epistemic grounds, but also potentially harmful for the agent's ability to gather valid information, depending on how relevance interacts with the assessment of other features of information, e.g. its correctness. Imagine relevance determines the *order* in which information is processed: the more relevant an information is, the sooner its credibility is assessed by the agent. Imagine further that information assessment follows some sort of *satisficing* procedure, *à la* Simon (1956): as soon as a candidate information is deemed good enough to warrant belief, no further candidates are considered. These assumptions are in fact very plausible for real-life agents in most situations: now it is easy to see that, in this case, assigning higher relevance to good news can make the agent believe in a positive information, even when there is stronger evidence for a conflicting, less pleasant option – if only one would deign to pay attention to it. Moreover, the bias towards good news would also contaminate the informational ecology where agents live: assuming sources are competing for attention (again, a relatively safe assumption), there would be a strong pressure to communicate only positive information, while passing under silence any bad news. Ultimately this would make not only the individual, but also the whole population blind to a host of relevant (albeit unpleasant) facts. Not to mention the extreme vulnerability to malicious sources, that could use flattery to gain the trust of others and than exploit it for their personal interest.

Defenders of utility-based relevance could reply that all these problems arise from a misguided definition of information utility. An information should not be considered useful because it is pleasant (that is, it makes me believe to progress towards my goals), but rather because it *fosters goal achievement as a matter of fact*. Given the goal of "hiking tomorrow", the relevance of the information "tomorrow will be sunny" hinges on whether or not tomorrow will indeed turn out to be a sunny day, regardless of how happy I am today to believe that piece of information. As for bad news, e.g. "tomorrow will rain", their utility should be assessed by considering what negative consequences they help averting, again assuming they turn out to be correct: even if I am not particularly happy of being told today that tomorrow will rain, thus spoiling my hopes for a beautiful hike, this is still good to know, if the alternative would be to stand tomorrow soaking wet on the top of a mountain. So it would seem that, once information utility is defined in terms of objective gain instead of subjective pleasure, considering it as synonymous of relevance is no longer problematic.[3]

However, here it is where our second objection kicks in, because there seems to be factors that affect information utility but not relevance – thus suggesting that the latter cannot be equated to the former, of which at most constitutes only one aspect. Consider an information that is not only useful, but also *indispensable* for a goal *G*: that is, in the absence of that information, *G* cannot be achieved in any other way. An ex-

[3] It is worth noting that this defense of utility-based relevance commits to consider all false information as irrelevant. While there are many who accept this claim, the matter is far from being settled, and proponents of an alethically neutral treatment of information and relevance are not hard to find (e.g., Devlin 1991; Colburn 2000; Fetzer 2004). Severing relevance from utility allows us to take no stance on this ongoing debate.

ample of indispensable information would be the code of a safe, given the goal of "opening the safe". Intuitively, the lack of alternatives makes indispensable information very precious, all other things being equal – that is, it increases the utility of that information. But does indispensability affect also relevance? Apparently, not at all. Imagine an alternative scenario, where the safe can be opened using either one of 1000 different codes. The information on each of these codes is clearly much less valuable than the information on the unique code in the original scenario, but the relevance of it is exactly the same – high or low, depending on how important is the goal of "opening the safe". In other words, while information utility seems to be affected by presence or absence of suitable alternatives, relevance is concerned only with whether or not the information is pertinent to the agent's goal, regardless of how many other information (if any) would be equally pertinent. This asymmetry provides independent reason to be skeptical of utility-based analysis of relevance, no matter how refined. Indeed, the definition of relevance provided in section 2 states that an information is relevant if and only if it is needed at some step of goal-processing, but does not demand that such information fosters or advances in any way the achievement of that goal. As such, our notion of relevance is *not* to be equated with any notion of information utility.

4. Subjective and objective relevance: goals vs. interests

Relevance is without doubt subject-dependent, in that what is relevant for me can well be irrelevant for you, and vice versa. However, subject-dependent does not mean subjective, and in fact it is possible (and, as we shall see, useful) to distinguish between *subjective* and *objective* relevance – both of which are subject-dependent. Subjective relevance refers to the agent's *goals*, that is, anticipatory representations of some state of affair that guide the agent's conduct. For an agent with the goal of "eating hamburgers", the location of the nearest McDonald's is subjectively relevant. Objective relevance, instead, relates to the agent's *interests*, that is, those states of affairs that, once realized, promote the agent's overall well-being, whether or not s/he was (or is) aware of this fact (Conte & Castelfranchi 1995). For an agent overly fond of McDonald's, information on the nutritional properties of hamburgers are objectively relevant to the interest of promoting his/her health, whether or not s/he currently cares about it.

In well-adapted agents, goals tend to align with interests, either because they coincide (an hungry agent has both the goal and the interest of eating) or because goals satisfy interests (the goal of having sex with beautiful partners generally satisfies the interest of reproduction with healthy conspecifics). However, it is neither impossible nor infrequent for agents to have goals that are *not* in their best interest, or even impede it: some youngsters might purposefully spend all their energies in social drinking and online gaming, thus damaging their own interest in having a suitable education and a decent career ahead of them. When this happens, the agent might consider as subjectively irrelevant even information of the utmost (objective) relevance, due to the maladaptive configuration of his/her goals. The youngsters in question, for instance, may honestly fail to see the relevance of their parents' tirade on the importance of doing homework and limiting Internet access.

More generally, even a generally well-behaved agent might temporarily fail to see what is best for him/her, or neglect to consider some alternative means available to an otherwise difficult or impossible end. In these cases, the agent's assessment of (subjective) relevance is "mistaken", more or less severely, not because s/he fails to compute it in relation to goals, but because those goals are either wrong or incomplete, with respect to the agent's objective interests. Imagine I have the goal of going to Rome and I am considering using either the train or the airplane; when I am told that "John will not use his car for the whole week", I might consider this information irrelevant because I fail to see the connection with my current plans – namely, that I could borrow John's car and drive to Rome instead. But of course the information is (objectively) relevant, and always was, whether or not I end up realizing it. What made me err in assessing relevance was the lack of the instrumental sub-goal "driving to Rome" in my plan structure, that was too narrowly constrained to either trains or planes.

Notice that similar errors could well be fully justified, and even reasonable, if considered in light of our limited cognitive resources. In the example above, I might have had perfectly good reasons to limit my planning efforts to trains and planes: let us say for instance that I do not own a car, my friends who do usually cannot spare it for long periods of time, and renting a car or taking a cab to Rome would be too expensive. Under such circumstances, excluding cars from my initial search for solutions is clearly a sensible policy, and ceases to be so only after being told of the (unexpected) availability of John's car. So there is nothing especially outrageous in the fact that I might at first fail to consider that information relevant for my plans, whereas I could be faulted as obdurate if I continue to ignore it once its relevance becomes apparent, e.g. because my friend patiently highlights the connection to me.[4]

Crucially, agents are in general well aware of their fallible nature, when it comes to assessing relevance. Thus they contemplate the possibility of error, and sometimes even anticipate it. This is why, when faced with an apparently irrelevant information, we often try to "make sense of it" – that is, we *assume* it is relevant, and then try to justify this assumption in light of our goals and background knowledge. Here trust in relevance is again crucial: when the puzzling message comes from a source that we trust to deliver relevant information, we will be extremely thorough in searching for an explanation that will make the message relevant; on the contrary, if the source was suspected from the start to be incapable of providing relevant data, the irrelevant message will just confirm this inability, and no "search for meaning" will ensue. More generally, the main reason why trust is required in handling relevance is precisely because we are not always the best judges of it, and we know it.

On the other hand, the distinction between subjective and objective relevance is also crucial from the standpoint of the information source: should the source take a "tutorial" attitude, that is, tailor its information to the objective interests of the recipient, or should it just efficiently answer explicit requests based on the recipient's goals? What if the source notices a discrepancy between the goals of the recipient (what s/he wants to know) and his/her interests (what s/he should know)? As mentioned, "overhelping" sources address the recipient's interests as the real information needs, rather

[4] Due to length constraints, here we focus only on instances were relevance assessment is *defective*: the agent does not consider as relevant something that actually is. But relevance assessment can also be *deceptive*: agents may consider relevant something that actually is not.

than his/her goals. Designing automated over-helping sources is a crucial challenge for agreement technologies, which also raises complex ethical issues (see Stock & Guerini in press).

5. Steps towards a formal model of trust in relevance

As discussed in section 2, a crucial problem in relevance assessment concerns how to establish exactly what pieces of information are concerned with (and thus relevant for) a given goal. This requires a principled mapping between goals and information, of the kind sketched in some of our previous work (most notably, Castelfranchi 1996, 1997; Castelfranchi & Paglieri 2007), and whether or not such mapping will be easy to implement in any running system is still an open issue. However, this obstacle should not bar progress on other aspects of relevance dynamics. In particular, we do not need to wait upon a fully worked out model of relevance assessment, in order to model trust in the relevance of the source and feedback dynamics from information relevance to source assessment.

Part of the reason why the time is ripe to introduce the topic of relevance in the study of trust is that much of what has been said for trust in general applies also to relevance, *mutatis mutandis*. So it is possible, exerting some care, to take advantage of some off-the-shelf models of trust (for an authoritative review, see Ramchurn et al. 2004) and "plug-in" relevance into them. In this section we will do just that, first adopting Demolombe's conceptual treatment of trust in truth and completeness (for a recent exposition, see Demolombe 2011) and extending it to trust in relevance, and later discussing instead how to expand the model of trust dynamics described by Villata and colleagues (2012) to account for relevance of sources, alongside with competence and sincerity. These two attempts are independent from each other, and are meant to exemplify two different directions to formalize and implement trust in relevance. The first strategy has the merit of offering a rich and nuanced picture of trust in information sources and is viable for formalization in modal logic, although it does not lend itself to immediate implementation and might have some limitations regarding the quantitative treatment of trust (but see Demolombe & Liau 2001; Lorini & Demolombe 2008; Demolombe 2009 for possible solutions). The second approach, on the contrary, offers a quantitative treatment of source qualities and thus facilitate implementation in multi-agent systems, but it also raises issue of feedback distribution – that is, once a source is observed to behave better or worse than expected, how is this performance to be diagnosed, so that it affects differentially each feature of the source? Since pros and cons of these two lines of research complement each other, we believe they are both worthy of being simultaneously and independently pursued.

5.1 Varieties of trust in relevance

In a series of influential papers, Demolombe analyzed various types of trust, or, to be more precise, trust in various features of an information source. Below is a summary of such varieties, taken from a recent paper (Demolombe 2011, p. 15, our emphasis):

Trust in *sincerity*: the truster believes that if he is informed by the trustee about some proposition, then the trustee believes that this proposition is true.

Trust in *competence*: the truster believes that if the trustee believes that some proposition is true, then this proposition is true.

Trust in *vigilance*: the truster believes that if some proposition is true, then the trustee believes that this proposition is true.

Trust in *cooperativity*: the truster believes that if the trustee believes that some proposition is true, then he is informed by the trustee about this proposition.

Trust in *validity*: the truster believes that if he is informed by the trustee about some proposition, then this proposition is true.

Trust in *completeness*: the truster believes that if some proposition is true, then he is informed by the trustee about this proposition.

It is easy to see that validity and completeness can be derived, respectively, from sincerity and competence (validity), and from vigilance and cooperativity (completeness), so we are inclined to disregard them as primitive features of the source. It is also evident that relevance is not included in this (fairly rich) picture of trust in information sources. This is perhaps not surprising, because Demolombe's categories revolve around the notion of truth, whereas relevance concerns the link between information and goals. For the same reason, Demolombe's truth-based concerns are orthogonal to those raised by considerations of relevance. Indeed, it is easy to generate the same four basic "varieties of trust", only with respect to relevance, as follows (labels in parentheses refer to the corresponding category for truth):

Trust in *pertinence* (sincerity): the truster believes that if s/he is informed by the trustee about some proposition, then the trustee believes that this proposition is relevant for him/her.

Trust in *understanding* (competence): the truster believes that if the trustee believes that some proposition is relevant for him/her, then this proposition is relevant for him/her.

Trust in *knowledgeability* (vigilance): the truster believes that if some proposition is relevant for him/her, then the trustee has some information about it.

Trust in *sharing* (cooperativity): the truster believes that if the trustee has information about a relevant proposition, then s/he is informed by the trustee of it.

This shows that (i) the same four relationships identified by Demolombe apply equally well to truth and to relevance, and indeed (ii) the more comprehensive notion of source quality should incorporate also the satisfaction of relevance constraints – that is, a source is of high quality only if its information turns out to be both true and relevant (an expansion of Demolombe's validity), and if it has access to enough true and relevant knowledge (an expansion of Demolombe's completeness). More subtly, while correctness and relevance are both essential and yet independent constraints on source quality, completeness should be more properly re-defined in terms of relevance. Indeed, the requirement of vigilance, as formulated by Demolombe (if some proposition is true, then the trustee believes that this proposition is true), imposes unrealistic demands on sources, and unnecessarily so. It is unlikely for any source, no matter how good it is, to be informed of *all* the facts of the world, and it is also largely

useless, since the agent who is querying that source will not care about most of that information: what does matter, of course, is to be knowledgeable about what is relevant for the agent. Similarly, Demolombe's requirement of cooperativity (if the trustee believes that some proposition is true, then the truster is informed by the trustee about this proposition) actually defines over-cooperation or, more exactly, blind cooperation: nobody would like to be flooded by irrelevant information from a source, only because that source happens to believe them. Once again, what matters is to be told about what the agent cares for, and only about that.

We believe that the requirements of knowledgeability and sharing, as defined above, together suffice to capture a more sensible notion of completeness, insofar as "having information about something" is understood as incorporating the constraint that false information is no information at all (Grice 1989, p. 371).[5] If this is correct, then we should only retain sincerity and competence from Demolombe's list, add our four relevance-based requirements, and then re-define validity as a combination of sincerity, competence, pertinence and understanding, and completeness as the combination of knowledgeability and sharing. Formalizing these notions in modal logic should then be fairly straightforward, following in Demolombe's footsteps, and this would provide a full-blown formal characterization of trust in information sources, including trust in their relevance.

5.2 Relevance, trust and feedback dynamics

In a recent paper, Villata and colleagues (2012) focused on a rather unexplored aspect of trust dynamics: namely, how the quality of an information, once it has been verified, should impact on the assessment of its source (*feedback dynamics*). The issue had already been discussed theoretically by Falcone and Castelfranchi (2004), but this was the first attempt to provide a suitable formalization of it. The model by Villata and colleagues, loosely based on the socio-cognitive theory of trust elaborated by Castelfranchi and Falcone (see their 2010 for a detailed exposition), distinguishes between competence and sincerity as relevant features of the source, and discusses how to integrate them to generate an expectation on the quality of information provided by the source, and how to differentially attribute praise (or blame) to each feature when that information turns out to be better (or worse) than expected.

There is no need to dwell on the technical details of their model, but it is worth noting that (i) once again, relevance was not contemplated as a key dimension, for either information or source, and yet (ii) its integration seems fairly straightforward to achieve. The first thing to note is that falsity and irrelevance are two different failures of information, and as such each of them will produce its own feedback on source assessment. The second aspect worth emphasizing is that, also with respect to irrelevance, it is possible to attribute this shortcoming either to lack of "competence" (in the sense of being unable to understand what the agent really needs) or to lack of honesty

[5] False information about X could still provide information *about something else* (typically, its source): when finance advisers were ensuring investors that sub-primes were a safe investment option, this did not convey any real information about sub-primes, but in hindsight speaks volume about the professional integrity of those advisers. None of this, however, changes the fact that false information about p does not provide any information on p.

(e.g., when the source deliberately misleads the agent, or withholds on purpose some vital information). So here we find the same problem described by Villata and collaborators in relation with truth and falsity of information: how to distribute the feedback between these two dimensions of source quality, i.e. competence and honesty.

Since the kind of competence required to know the truth about the world is very different from the kind of competence needed to divine the agent's informational needs, we believe a good model of information sources should keep these two aspects separate: a source could well be extremely good at understanding what others need to know, and yet poorly informed about such issues, or vice versa. In contrast, honesty (or sincerity, if one prefers that label) seems to be essentially the same feature, whether it applies to truth or relevance: an agent that deliberately lies, misleads, or omits important information is still manifesting the same kind of dishonest attitude, although in different ways (different kinds of "deception"). Thus we propose to extend the model by Villata and colleagues as follows:

- consider three dimensions (instead of two) of source quality: *competence* (whether the source knows the truth), *understanding* (whether the source understands what the agent needs), and *honesty* (whether the source tells what s/he believes to be both true and relevant for the agent);[6]
- consider *two types of feedback*, one related to truth, the other to relevance;
- distribute *truth-based feedback* between competence and honesty, and *relevance-based feedback* between understanding and honesty.

Regarding *feedback determination* (how much should source assessment change, given the observed quality of information), we agree with Villata and colleagues that this essentially depends on the mismatch between prior expectations and actual performance of the source: if I expect a source to deliver information of a certain quality (positive or negative, it does not matter), the fact that I receive information of that exact quality adds nothing to my knowledge of the source – it merely confirms my prior assessment (so, no feedback). Conversely, it is when I am surprised (delighted or disappointed) by the quality of information provided from that source that I have reason to revise my assessment, and the greater my surprise, the larger the change in source consideration. We believe that this basic principle applies to the assessment of both truth-based and relevance-based feedback.[7]

---

[6] Villata and colleagues index competence and sincerity (their label for honesty) to content domains. This is not completely satisfactory: domain-indexing works well for competence (a physician is competent about medicines and not about cars, whereas the opposite is true for a mechanic), but it is highly problematic for honesty, since the fact that a source may be deceptive in some cases and not in others does not depend on semantic domains, but rather on personal interests. A suspicious wife will believe his husband to be insincere when he protests his innocence, but she will have no doubt when he confesses his escapades – and yet the domain is exactly the same in both cases. So domain-indexing could be applied also to understanding (it is conceivable that a source is better attuned to my needs regarding some domain rather than others), but should not be adopted for honesty.

[7] A further complication is that, as mentioned before, *sometimes our expectations about the source make us doubt our own assessment of the information*: if a highly trusted source gives me an information that at first sight I judge to be false/irrelevant, I might very well have a second look and try to "make it" true/relevant through interpretative effort. Whether this is a perfectly sensible policy or a dangerous slippery slope towards wishful-thinking is

The issue of *feedback distribution* (what portion of the overall feedback should impact on each features of the source) is much more thorny, especially if one tries (as Villata and colleagues did) to deal with it using fairly limited expressive resources – in particular, without giving the agent any causal attribution theory on source performance, and without endowing agents with any representation of the beliefs and goals of the sources (Falcone & Castelfranchi 2004). When I realize there is something amiss with the information you provided me, I am usually able to figure out a plausible interpretation of what caused such dismal performance: some reasons might actually block feedback dynamics completely (e.g., you gave me wrong information on train schedule because you were not, and could not be, aware of the last-minute strike of railway personnel), while other reasons might indicate how to distribute it (e.g., whether or not I can attribute you some malicious intent or competing interest will be crucial to make me suspect you to be either dishonest or incompetent). Lacking similar resources, one has to resort to rule-of-thumbs for feedback distribution, of the kind proposed by Villata and colleagues – and they acknowledge the limits of this approach, in spite of its merits (e.g., a relative simplicity in agent's architecture).

While it is not our aim to further deal with feedback dynamics, we would like instead to conclude this section by pointing out that only relevance allows to analyze a special type of information failure: *omissions*. As mentioned before, in relation to Demolombe's requirements of vigilance and cooperativity, a source cannot be faulted for not knowing all the facts in the world, or for not mentioning all the things s/he believes to be true. But when a source knows something that is *relevant* for an agent, and yet fails to inform the agent of such fact, then we can accuse the source of omitting something, due to a lack of understanding or honesty. In other words, an omission is not a failure to mention any true fact, but rather failure to communicate relevant (and true) information. Without a proper analysis of relevance, omissions cannot be identified, because one cannot distinguish between the harmless (indeed, beneficial) habit of passing under silence everything that does not matter, and the damaging policy of withholding valuable information, on purpose or by mistake (*reticence*).

6. Conclusions and future work

Our main aim in this paper was to put on the table a crucial and largely unexplored issue in trust theory: so the emphasis was on presenting the problem in all its complexity, rather than on finding suitable ways of solving it. Yet hopefully this first recognition will later prove instrumental to engender mature solutions, inasmuch as problem setting is an essential component of problem solving. In particular, we stressed that (i) relevance requires trust in information sources and is a key feature of their assessment, on a par with competence; (ii) relevance is needed to properly constrain the definition of another key feature of sources, to wit, their completeness; (iii) relevance is based on goals, and a principled theory of the interaction between goals and beliefs

essentially a matter of degree. However, it is worth noting that dealing with similar issues would complicate the simple mismatch-based rule of feedback determination adopted by Villata and colleagues (2012), and a Dempster-Shafer treatment of ignorance gaps might result helpful in that respect – intuitively, the more ignorance I have on my assessment of information quality, the more benefit of the doubt is reasonable to concede to trusted sources.

allows assessing relevance without resorting to convoluted, counter-factual approaches; (iv) relevance contributes to the utility of information, yet should not be identified with it; (v) relevance has both a subjective and an objective side, the former relative to goals, the latter to interests; (vi) several formal and computational models developed for trust in general are easily adapted to capture trust in relevance, and doing so helps refining such models. More generally, we hope that these preliminary considerations have shown that this relatively unexplored line of research has much potential for the theory and technologies of trust, thus providing some encouragement for others to trod the same path further and more systematically.

References

Borg, E.: Intention-based semantics. In: LePore, E., Smith, B. (eds.) The Oxford handbook of philosophy of language, pp. 250-267. Oxford, Clarendon Press (2005)

Borlund, P.: The concept of relevance in IR. Journal of the American Society for Information Science and Technology 54(10), 913–925 (2003)

Bremer, M., Cohnitz, D.: Information and information flow – An introduction. Frankfurt, Ontos Verlag (2004)

Castelfranchi, C.: Reasons: Belief support and goal dynamics. Mathware & Soft Computing 3, 233–247 (1996)

Castelfranchi, C.: Representation and integration of multiple knowledge sources: Issues and questions. In: Cantoni, V., Di Gesù, V., Setti, A., Tegolo, D. (eds.), Human & machine perception: Information fusion. New York, Plenum Press, pp. 235–254 (1997)

Castelfranchi, C.: Affective appraisal versus cognitive evaluation in social emotions and interactions. In: Paiva, A.M. (ed.) Affective Interactions, pp. 76-106. Berlin, Springer (2000)

Castelfranchi, C., Falcone, R.: Trust theory: A socio-cognitive and computational model. London, Wiley (2010)

Castelfranchi, C., Paglieri, F.: The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. Synthese 155, 237–263 (2007)

Cohen, J.: Some steps towards a general theory of relevance. Synthese 101, 171–185 (1994)

Colburn, T.R.: Philosophy and computer science. Armonk, M.E. Sharpe (2000)

Conte, R., Castelfranchi, C.: Social and cognitive action. London, UCL Press (1995)

Cooper, W.S.: A definition of relevance for information retrieval. Information Storage and Retrieval 7, 19–37 (1971)

Crestani, F., Lalmas, M., Van Rijsbergen, C.J., Campbell, I.: Is this document relevant?… Probably: A survey of probabilistic models in information retrieval. ACM Computing Surveys 30(4), 528–552 (1998)

Demolombe, R.: Graded trust. In: Proceedings of Trust in Agent Societies (TRUST 2009), Budapest, AAMAS (2009)

Demolombe, R.: Transitivity and propagation of trust in information sources: An analysis in modal logic. In: Proceedings of Computational Logic in Multi-Agent Systems (CLIMA XII). Berlin, Springer, pp. 13–28 (2011)

Demolombe, R., Liau, C.-J.: A logic of graded trust and belief fusion. In: Proceedings of the 4th Workshop on Deception, Fraud and Trust in Agent Societies (TRUST 2001). Montreal, Canada, pp. 13–25 (2001)

Devlin, K.J.: Logic and information. Cambridge, Cambridge University Press (1991)

Falcone, R., Castelfranchi, C.: The human in the loop of a delegated agent: The theory of adjustable social autonomy. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 31(5), 406–418 (2001)

Falcone, R., Castelfranchi, C.: Trust dynamics: How trust is influenced by direct experiences and by trust itself. In: Proceedings of AAMAS 2004. New York, ACM, pp. 740–747 (2004)

Fetzer, J.H.: Information, misinformation, and disinformation. Minds and Machines 14(2), 223–229 (2004)

Floridi, L.: Information. In: Floridi, L. (ed.), The Blackwell guide to the philosophy of computing and information, pp. 40–61. Oxford/New York, Blackwell (2004)

Floridi, L.: Understanding epistemic relevance. Erkenntnis 69, 69–92 (2008)

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

Lorini, E., Demolombe, R.: From binary trust to graded trust in information sources: A logical perspective. In: Proceedings of Trust in Agent Societies (TRUST 2008). Berlin, Springer, pp. 79–93 (2008)

Miceli, M., Castelfranchi, C.: A cognitive approach to values. Journal for the Theory of Social Behaviour 19(2), 169–193 (1989)

Paglieri, F.: Data-oriented Belief Revision: Towards a unified theory of epistemic processing. In: Proceedings of STAIRS 2004. Amsterdam, IOS Press, pp. 179–190 (2004)

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., Granka, L.: In Google we trust: Users' decisions on rank, position, and relevance. Journal of Computer-Mediated Communication 12, 801–823 (2007)

Ramchurn, S., Huynh, D., Jennings, N.: Trust in multi-agent systems. The Knowledge Engineering Review 19(1), 1–25 (2004)

Simon, H.: Rational choice and the structure of the environment. Psychological Review 63(2), 129–138 (1956)

Sperber, D., Wilson, D.: Relevance: Communication and cognition (2nd ed.). Malden, Basil Blackwell (1995)

Stock, O., Guerini, M.: Investigating ethical issues for persuasive systems. In: Paglieri, F., Tummolini, L., Falcone, R., Miceli, M. (eds.), The goals of cognition. Essays in honor of Cristiano Castelfranchi. College Publications, London (in press)

Villata, S., Paglieri, F., Tettamanzi, A., Falcone, R., da Costa Pereira, C., Castelfranchi, C.: Trusting the messenger and the message. In: Proceedings of Trust in Agent Societies (TRUST 2012). Valencia, AAMAS, pp. 79–93 (2012)

Wilson, D., Sperber, D.: Relevance theory. In: Horn, L.R., Ward, G.L (eds.), The handbook of pragmatics, pp. 607–632. Malden, Blackwell (2004)