Argument and Computation, 2014
Vol. 5, Nos. 2–3, 160–190, http://dx.doi.org/10.1080/19462166.2014.913075

Argument schemes for reasoning about trust[†]

Simon Parsons[a]*, Katie Atkinson[a], Zimi Li[b], Peter McBurney[c], Elizabeth Sklar[a], Munindar Singh[d], Karen Haigh[e], Karl Levitt[f] and Jeff Rowe[f]

[a]*Department of Computer Science, University of Liverpool, Liverpool, UK;* [b]*Department of Computer Science, Graduate Center, City University of New York, New York, NY, USA;* [c]*Department of Informatics, King's College London, London, UK;* [d]*Department of Computer Science, North Carolina State University, Raleigh, NC, USA;* [e]*Raytheon BBN Technologies, Minneapolis, MN, USA;* [f]*Department of Computer Science, University of California, Davis, CA, USA*

(*Received 21 December 2013; accepted 4 April 2014*)

Trust is a natural mechanism by which an autonomous party, an *agent*, can deal with the inherent uncertainty regarding the behaviours of other parties and the uncertainty in the information it shares with those parties. Trust is thus crucial in any decentralised system. This paper builds on recent efforts to use argumentation to reason about trust. Specifically, a set of *schemes* is provided, and abstract patterns of reasoning that apply in multiple situations geared towards trust. Schemes are described in which one agent, A, can establish arguments for trusting another agent, B, directly, as well as schemes that A can use to construct arguments for trusting C, where C is trusted by B. For both sets of schemes, a set of *critical questions* is offered that identify the situations in which these schemes can fail.

Keywords: argument representation; formal models of argumentation; social influence

1. Introduction

Trust can be considered a mechanism for managing the uncertainty that autonomous entities, *agents*, face with respect to the behaviour of entities they interact with and the information supplied by other entities. As a result, trust can play an important role in any decentralised system. As computer systems have become increasingly distributed, and control in those systems has become more decentralised, trust has become steadily more important within computer science (Grandison & Sloman, 2000). There have been studies, for example, on the development of trust in e-commerce (Mui, Moteashemi, & Halberstadt, 2002; Resnick & Zeckhauser, 2002; Yu & Singh, 2002), on mechanisms to determine which sources to trust when faced with multiple conflicting sources (Dong, Berti-Équille, & Srivastava, 2009), and on mechanisms for identifying which individuals to trust based on their past activity (Hang, Wang, & Singh, 2008; Li & Wang, 2010). Trust is especially important from the perspective of autonomous agents and multi-agent systems (Teacy, Chalkiadakis, Rogers, & Jennings, 2008), and as a result we find much work on trust in agent-based systems (Sabater & Sierra, 2005; Wang & Singh, 2006).

Although most of us have an intuitive idea of the meaning of the term 'trust', it is hard to define precisely. As a result of this conceptual slipperiness, there are a number of different definitions of trust in the literature. Sztompka (1999), for example, suggests that:

> Trust is a bet about the future contingent actions of others.

*Corresponding author. Email: s.d.parsons@liverpool.ac.uk

[†]This is a revised and extended version of a paper that appeared at the International Conference on Computational Argumentation (COMMA), 2012 (Parsons et al., 2012).

While McKnight and Chervany (1996), drawing on a range of existing definitions, offer the suggestion that:

> Trust is the extent to which one party is willing to depend on something or somebody in a given situation with a feeling of relative security, even though negative consequences are possible.

Gambetta (1990) states that:

> Trust is the subjective probability by which an individual, A, expects that another individual, B, performs a given action on which [A's] welfare depends.

While Mui et al. (2002) define trust as:

> A subjective expectation an agent has about another's future behaviour based on the history of their encounters.

These definitions, although different, do overlap somewhat. All four definitions given here focus on trust as a mechanism for making predictions about the future actions of individuals. That is, if one individual trusts another, then that first individual can make a (more or less) accurate prediction about what the other will do in the future. One might decide, as Gambetta (1990) does explicitly and Sztompka (1999), does implicitly,[1] that trust can be quantified as a probability. Or one might decide, as Castelfranchi and Falcone (2000) argue, that trust is more complex than mere probability and instead has a rational basis in reasons for beliefs about the future actions of others.

In this paper, we follow Castelfranchi and Falcone (2000) in maintaining that trust should be reason-based:

> Trust as attitude is epistemically rational when it is reason-based. When it is based on well motivated evidences [*sic*] and on good inferences, when its constitutive beliefs are well grounded (their credibility is correctly based on external and internal credible sources); when the evaluation is realistic and the esteem is justified.

Indeed, we go further, suggesting that argumentation, as a mechanism for constructing arguments (reasons) for and against adopting beliefs and pursuing actions, is an appropriate mechanism for reasoning about trust. Several recent approaches examine such reasoning (Parsons, McBurney, & Sklar, 2010; Parsons, Sklar, & McBurney, 2011; Stranders, de Weerdt, & Witteveen, 2008; Villata, Boella, Gabbay, & van der Torre, 2011). In order to develop a comprehensive model of argumentation for reasoning about trust, it is necessary to identify those patterns of argumentation that apply to trust. We make a first attempt to do that in this paper. In particular, we follow Walton, Reed, and Macagno (2008) in identifying argument *schemes*, patterns for constructing arguments, and capturing *critical questions* for those schemes. Critical questions, to paraphrase (Walton et al., 2008), identify assumptions inherent in argument schemes, and so provide a way of capturing the fact that the schemes represent defeasible knowledge. Based on the critical questions and the answers to those questions, we can identify whether applying an argument scheme will give us a sound argument or whether that argument will be fallacious.

We restrict attention in this paper to two classes of schemes for reasoning about trust – schemes concerned with establishing if some individual A can trust another individual B, and schemes for arguing about the propagation of trust. By 'propagation of trust', we mean that the schemes are concerned with whether, if it has been determined that A trusts B (perhaps by applying one of the first kind of scheme) and it has been determined that B trusts C, then it is appropriate that A should trust C.[2] There are other classes of argument scheme that may be applicable to trust, and we discuss some of these without going into too much detail – a full exploration of these other classes of schemes is a topic for another paper. In any domain, the classification of the relevant set of concepts for representing the domain and the relations between them is a hard task, and this is especially so in a complex domain like trust. Given this complexity, the identification of a solid categorisation has value, and this is our justification for enumerating the narrow set of schemes that we present here.

Finally, we should note that our approach to trust is somewhat different to that taken by the research behind the definitions presented above. These tend to the view that trust is an entity – a bet, a probability, or an expectation. As will become clear, we consider trust to be a relation that holds between two entities, the trustor and the trustee. This view is not in conflict with the other definitions – a conditional probability, after all, exactly captures a relationship between two variables. This view is also not novel, since the idea of trust as a relation is common in the literature, not least in the notion of a trust graph, as seen, for example, in Katz and Golbeck (2006). However, what we claim is that our view of trust is more nuanced than that provided by prior work and helps to tease apart the various strands that comprise previous, rather monolithic, notions of trust.

This paper is structured as follows. In Section 2, we present a set of schemes that one agent, A, can use to establish trust in another agent, B. Section 3 then describes a set of critical questions relating to the trust schemes. Section 4 switches attention to what the social network literature calls trust propagation. These schemes also have associated critical questions, and these questions are discussed in Section 5. Section 6 provides general discussion, and Section 7 highlights related work. Section 8 concludes.

2. Argument schemes for trusting

A quick perusal of the literature on trust suggests that there are a number of different ways in which one individual may infer trust in another. McKnight and Chervany (1996) list a number of contexts in which trust may occur, and Jøsang, Gray, and Kinateder (2006) distinguish between *functional* trust, the trust in an individual to carry out some task, and *referral* trust, the trust in an individual's recommendation. If we are using argumentation to reason about trust, then we need to be able to capture these different approaches within an argument since identifying the steps in reaching a conclusion is fundamental to argumentation.

In this section, we collect a number of standard *schemes* or patterns of constructing arguments that relate to trust. In particular, we consider arguments about whether to trust an individual B. While the trust in question could be functional trust or referral trust, the examples we give are all to do with functional trust – they concern whether to trust B to carry out some task. None of these patterns captured by the argument schemes represent deductive reasoning – all may be wrong under some circumstances, and some may be wrong more often than they are right – but all represent forms of reasoning that are *plausible* under some circumstances. Because the schemes are only plausible, there are conditions under which they can fail. For each scheme, we identify these failure conditions in the form of critical questions, in essence asking if the assumptions that underlie the schemes hold in the case we are interested in.

Note that in all the schemes, context is important. B is trusted only in some particular context for some behaviour, and we assume that these schemes are only being applied in that context relating to that behaviour. Thus, as we discuss later, there is a critical question relating to context that applies to every one of the schemes that we consider.

2.1. *The schemes*

We have the following argument schemes. Note that all schemes are stated in positive terms in that the schemes apply, they generate arguments for trusting an individual rather than arguments for not trusting, or for distrusting (we briefly discuss schemes for the latter in Section 6).

Trust from Direct Experience (DE) If *A* has personal experience of *B* and has found *B* to be reliable, then *A* may decide to trust *B*. In this case, and in contrast to the cases below, *A* validates *B* for *A*.

As an example, consider reasoning about whether a restaurant can be trusted to cook good food. When I have visited a restaurant in the past and found that the food is good, then I might trust the restaurant to produce good food in the future. My trust typically increases with the number of visits that I have made when I have had a good meal, although my trust may decrease disproportionately following one bad experience.

This argument scheme is similar to 'Argument from commitment' (Walton et al., 2008, p. 335), where an individual is argued to be committed to some proposition on the evidence that what they said or did shows that they are so committed. In the case of trust, past evidence is reliable behaviour and commitment is typically required to be repeated in order to reinforce trust.

Trust from Indirect Experience (IE) If *A* has observed evidence that leads it to believe that *B*'s behaviour over some period has been reliable, then *A* may decide to trust *B*.

Thus, I might develop a degree of trust in the quality of food served at a restaurant, not because I have visited it, but because I have walked past it on a number of occasions and observed, for example, the fact that the tables are always full, and there are people waiting to get in.

This form of argument is distinguished from the previous case because of the fact that the experience is indirect. In the case of this restaurant example, I have not tried the food; I am just inferring from what I see that it is probably good. Unlike the DE case, my inference about the food could be wrong – the crowds might reflect the fact that it is a fashionable restaurant that serves mediocre food, that it has a famous DJ who regularly spins there, or that the owners are regularly comping[3] indifferent meals for their friends in the hope that it will entice other people to come in.

Trust from Expert Opinion (EO) If *B* is an expert in some domain of competence, then *A* may decide to trust *B*.

In this case, *B* is validated by some entity other than *A*. For example, *B* is a doctor and is validated by the appropriate medical board, or, to continue the restaurant example, *B* is a graduate of the Culinary Institute of America (CIA) and has earned its diploma. This validation is distinct from the previous cases (DE and IE) because the opinion is that of an expert – certified by the medical board or the CIA – rather than a non-expert, *A*. The validation is also backed up by concrete evidence: in the case of our examples, medical school diploma, licensing board certificates, or cooking school diplomas.

In some cases (such as the case where *B* is a chef), *A* is also able to validate *B*'s trustworthiness for themself (by eating *B*'s food) and so augment an argument from expert opinion (based on the CIA diploma) with an argument from direct experience. In other cases, such as the case where *B* is a doctor, it is not clear that *A* can perform any direct validation; and instead, *A* always infers part of their trust in *B* as a result of seeing the certificate hanging on the wall in *B*'s office as an emblem of their expertise.

This argument scheme is clearly related to the 'Argument from expert opinion' from Walton et al. (2008, p. 15), and this relationship is discussed in more detail below.

Trust from Authority (Au) If *B* holds a position in an organisation that exercises powers of authority, then *A* may decide to trust *B*.

We distinguish this case from the previous one (EO) in the sense that here the trustworthiness of *B* stems from the organisation where *B* holds its position, rather than from external certification earned by *B* as an individual. In other words, we see trust from expert opinion arising from what an individual is believed to know in their own right, whereas trust from authority arises from the inherent power and knowledge associated with an individual based on their role in a particular type of organisation, generally an organisation that enforces rules.

Two further examples help to distinguish between the EO and Au trust schemes. An employee of the Metropolitan Transit Authority (MTA),[4] for example, might well be trusted on questions about subway operation, on the grounds that their job means that they will have reliable and up-to-date information. This is trust from authority. However, if they no longer have any affiliation with the MTA (e.g. because they change jobs), then trust from authority would not be warranted. Although, if our MTA employee retired after 40 years of experience developing train schedules, then they could still be considered a trustworthy expert on train schedules long after they leave their job. Taking another example, consider a university professor who is invited by a reporter to comment on the subject in which they are regarded as an academic expert. The reason that they are trusted to answer questions about their area of academic research is because of their knowledge, certified, for example, by their PhD. However, if this professor also held an administrative role, such as head of postgraduate studies, and was asked questions about the requirements necessary for graduation, then they would be trusted as an authority, not as an expert.

Trust from Reputation (Rep) If *B* has a reputation for being trustworthy, then *A* may decide to trust *B*.

We consider that *A* may either have heard people saying that *B* is trustworthy, or be aware of some aggregate measure of reputation that applies to *B*. To continue with the restaurant example, *A* may hear reports from friends, or may distil the reputation of *B* from a service like Yelp.[5]

We distinguish the idea of an argument based on reputation from an argument based on EO or Au because the recommenders are neither experts nor authorities. If *A*'s sources are experts then, in our view, *A* would be using the expert opinion scheme; similarly, if *A*'s sources represented organisations exercising authority, then *A* would be applying the trust from authority scheme. We also note that an argument from reputation is not the same as propagating trust. If *A* hears that *B* has a good reputation, this is a statement about *B*'s trustworthiness. It, like all the other schemes we consider here, is the establishment of a link between *A* and *B*. (Information derived from reputation may be used in propagating trust between individuals, but the use of the reputation scheme does not imply propagation.[6])

Trust from Moral Nature (MN) If *A* judges that *B* has a good character, then *A* may decide to trust *B*.

Here, *A* is performing some inference about *A* that is grounded not in *A*'s knowledge of *B*'s past behaviour (as with the DE or IE schemes), nor is *A*'s view guided by *B*'s professional expertise, nor by *B*'s position in society. Rather, *A* is making some observation based on some aspects of *B*'s behaviour and inferring trustworthiness from that (Walton et al., 2008, p. 141) classifies this kind of argument as the 'Aristotelian ethotic argument'.

For example, suppose I observe a customer, Elin, at the cash register in a shop. She has realised that the cashier returned too much change to her in a transaction, and she tells the cashier about it and returns the extra change that she received in error. From this observation, I determine that Elin is of sound moral nature. Thus, the MN trust scheme could be applied to any information I receive or interaction I anticipate from Elin in the future.

Trust from Social Standing (SS) If *A* judges that *B* would have too much to lose by lying, then *A* may decide to trust *B*.

Here, *A* performs a kind of expected utility calculation in terms of *B*'s position, asking what *B* could gain by lying and what *B* could lose by being exposed, and deciding that the former is less than the latter. The possible loss is, roughly speaking, related to *B*'s position in society. In a real community, this is, we believe, a large part of what motivates trust between peers (as opposed to between folk who are externally certified) and is one of the reasons that establishing trust is a challenge in the online world, where it is so easy to hide behind anonymity and to create new aliases.

Trust from Majority Behaviour (M) If *A* has found most people in the set from which *B* is drawn to be trustworthy, then *A* may decide to trust *B*.

This is an even less deductive form of trust derivation than most of the above, but it is still one we use. For example, when buying online, many of us are happy to trust our data to merchants we have no specific recommendation about because of our overall good experience with online merchants. In effect, we are generalising from some experiences with some individuals or to all individuals and in a group (possibly while being aware that not all individuals in the group are trustworthy).

Trust from Prudence (Pru) A may decide to trust *B* because it is less risky than not trusting *B*.

The key to the prudence scheme is the assessment of risk involved in trusting *B*. As an example, consider the case where you are running late for an important meeting, but are now lost and seem certain to miss the meeting unless you immediately find the correct route to your destination. This is a situation in which it makes sense to ask for directions from *B*, even though you do not know whether *B* is particularly trustworthy. There is a chance that the directions will be good, and you will get there on time, and taking this risk is better than continuing to blunder around not knowing your way.

Trust from Pragmatism (Pra) A may decide to trust *B* because it (currently) serves *A*'s interests to do so.

The pragmatism scheme only considers *A*'s current interests. If these are served by trusting *B*, then *A* can decide to trust *B*. For example, if *A* can only achieve what it wants to do by trusting *B*, then it may make sense to trust *B*. If I need to get a package to a destination that I cannot possibly reach, then it may make sense to entrust delivery to someone whom I have no other information on just because there is no other way to achieve my goal.[7] If the package is valuable, then this might not be a good scheme to apply, but in less critical situations it may be reasonable. (Imagine asking the taxi driver who took you to the airport to drop a postcard that you had forgotten to post into a postbox so that the addressee gets the card with an appropriate postmark.) Indeed, it may well be a suitable scheme from which to start bootstrapping trust. To go back to our restaurant example, I might decide to trust that a restaurant provides decent food just because I am hungry and need to eat. If it turns out that the food is good, then I may start to trust the restaurant as a purveyor of good food.

Now, while trust from pragmatism depends on *A*'s goal being served by trusting *B*, the scheme makes no reference to *B*'s goals. If the goals of the two agents are aligned, then there is an additional reason for *A* to trust *B*.

Trust from Mutual Goals (MG) A may decide to trust *B* because they (currently) share the same goals.

An example of trust from mutual goals is where one coalition partner trusts another just because the two are part of a coalition, and so are currently working towards the same goals. For example, Naylor (2005) describes the coalition between the US army and the Northern Alliance in Afghanistan in 2001. While the US and the Northern Alliance had the same goal – the toppling of the Taliban – the two sides worked closely together and, from the US perspective, the Northern Alliance could be trusted. However, once the Taliban had collapsed and the goals of the two sides diverged – the US to capture Al Qaida leaders and the Northern Alliance to consolidate power – trust between the two sides diminished. Similarly, two faculty members in a university might support opposing political parties, but because of their shared interest in increased funding for education, they might trust each other to work hard to change education policy to increase that funding.

This form of trust, then, is all about having goals that align. As the last example showed, it is plausible to trust someone even if their beliefs differ significantly from one's own. However, it is also plausible to construct an argument for trusting someone on the grounds that their beliefs are similar to one's own beliefs, as in the next case.

Trust from Mutual Beliefs (MB) A may decide to trust B because B holds the same set of beliefs as A.

A scenario in which mutual belief can lead to trust is the converse of the situation of the two faculty members who had differing political views. If two people share similar political beliefs – about the paramount importance of shrinking the role of government, for example – they may decide that their values are sufficiently closely aligned that each can trust the other to be a reliable partner in helping to elect conservative politicians and in campaigning to restrict women's access to health care. Another version of trust from mutual belief, one that is more related to trust in an agent as an information source than as an actor, is the form of trust that arises if someone tells us some fact that we believe to be true. Because we can verify the statement, it provides us with some evidence that this person is truthful, and so other statements that they make can be trusted to be true.

Talking of evidence that can be verified makes Trust from Mutual Belief sound as if it might be related to Trust from Direct Experience, but this is not the case. DE captures trust that is derived from observing some action. MB captures trust that is derived from finding that another individual shares beliefs. They both might apply in some cases – as, for example, in the case when someone tells us their belief about some objectively verifiable fact like 'I believe it is raining outside'. If we check this, and it is true, then we have direct experience of the individual as a source of information about what is going on outside. Then, because of a shared belief in current precipitation, we might be inclined to trust that same individual's view about how sensible it would be to venture outside (at least in comparison with how much one would trust someone who, unlike us, believes it is not raining).

Trust from Organisation (Org) A may decide to trust B because B is a member of some organisation.

It is not uncommon for individuals to trust others, somewhat indiscriminately, on the basis of the organisation that they belong to. Alumni networks are an example of such organisations, and there are cultures in which A will trust B, a person they have never met, on the basis of some longstanding connection between the family of A and the family of B. (The lack of connection between A and B themselves means that this is not an example of direct or indirect experience.) Indeed, there may be situations in which trust is only ever extended to members of such organisations, to quote Abner Mikva (Kreisler, 1999):

> . . . on the way home from law school one night in 1948, I stopped by the ward headquarters in the ward where I lived. There was a street-front, and the name Timothy O'Sullivan, Ward Committeeman, was painted on the front window. I walked in and I said 'I'd like to volunteer to work for Stevenson and Douglas'. This quintessential Chicago ward committeeman took the cigar out of his mouth and glared at me and said, 'Who sent you?' I said, 'Nobody sent me'. He put the cigar back in his mouth and he said, 'We don't want nobody that nobody sent'. This was the beginning of my political career in Chicago.

In the political context, of course, the source of the trust might be closer to Trust from Mutual Beliefs, though in the world of party machine politics that Mikva was describing, it is loyalty to the organisation (often on the basis of what the organisation can eventually do for you) rather than political belief that is the basis of trust. This reliance on some future benefit is not sufficient to make this example one of mutual goals – the goals could be different since the party worker might not care about getting the candidate elected, they might just care about the sinecure that they will land for helping out.

Note that several of the schemes described here – Trust from Expert Opinion, Trust from Authority, Trust from Moral Nature, and Trust from Social Standing – might be considered to be specific instances of Trust from Majority Behaviour. The reason, after all, that we trust Dr B to advise us on medical matters is that most doctors are trustworthy on medical matters. What distinguishes arguments constructed from the Majority Behaviour scheme from those more specific cases just listed is that the latter are, in our particular cultural milieu, ones that are especially reliable. As we argued in the introduction, part of the issue in developing a representation of any domain is to identify the relevant set of concepts. The fact that we can distinguish these different forms of 'majority behaviour' is a good reason to highlight the schemes.

2.2. *Discussion*

First we should note that these schemes are all concerned with A establishing its own trust in an individual for themself, not deriving trust as a result of reports from others (as mentioned above, establishing schemes for trust propagation is the subject of Section 4). Second, as noted above, all the schemes are stated in the positive, providing reasons for trusting an individual. Here we do not consider schemes for not trusting an individual or for distrusting an individual – as mentioned above, such concerns are briefly considered in Section 6.

The schemes break down into five categories. The first category is one in which A has collected evidence about the trustworthiness of a given individual. This category includes two schemes, direct and indirect experience, relating to whether the evidence that A has collected is direct evidence of B's trustworthiness, or evidence of something from which trustworthiness can be derived.

The second category includes three schemes – expert opinion, authority, and reputation – where B is considered to be trustworthy on some subject, or in some role, not because A has observed them doing this (directly or indirectly) but because there is some validation of B that can, in theory at least, be verified by A. If B is somehow validated as an expert or an authority, A can reasonably trust them. Of course, there are many cases of supposed experts or authority figures being shown to be frauds, but the fact that the argument scheme can lead to misplaced trust is not an argument against the scheme so much as an argument for careful posing of the critical questions. For example, concern about fraudulent experts might lead one to be vigilant in checking their certification. In the reputation scheme, A bases their decision on a distillation of reported evidence from other individuals. Since those individuals are not typically validated as experts or authorities (someone reviewing a restaurant on Yelp might be a chef, but it is hard to establish whether this is the case), what A relies on in the reputation scheme is the 'wisdom of the crowd' (Surowiecki, 2005). In other words, the average of a number of reported experiences is likely close to the experience that *A* will have if *A* is a typical member of the population who are providing the reviews.

It is important to distinguish reputation from reasoning based on *referral trust*. Reputation is established from many individuals and refers to one specific recommendation. Referral trust is trust in an individual's ability to make trustworthy recommendations, and so is established about one individual, by one of the mechanisms discussed here. Referral trust may be established by reputation, that is, *A* may decide to trust *B* to make referrals on the basis of good reports of *B*'s referral ability that are made by *D* and *E*. Note also the distinction between the reputation scheme and the majority scheme. *A* might decide to trust the quality of the food prepared at restaurant *B* because of good reports about *B*. That is an instance of the reputation scheme. Alternatively, *A* may decide to trust the quality of food at *B* because *A* already trusts the quality of the food at all the other restaurants in the neighbourhood of *B*. That is an instance of the majority scheme.

Furthermore concerning this category of schemes, there is clearly a relation with existing well-known schemes from the informal logic literature. Much has been written about schemes for expert opinion and appeal to authority in general (for example, (Walton, 1997; Walton et al., 2008). The general scheme for expert opinion set out in Walton et al. (2008) expresses that an expert in some subject domain asserts some proposition that can be defeasibly accepted as true:

> Major premise: Source E is an expert in subject domain S containing proposition A.
> Minor premise: E asserts that proposition A is true (false).
> Conclusion: A is true (false).

This scheme is accompanied by six critical questions used to evaluate an argument from expert opinion:

(1) Expertise question: How credible is E as an expert source?
(2) Field question: Is E an expert in the field F that A is in?
(3) Opinion question: What did E assert that implies A?
(4) Trustworthiness question: Is E personally reliable as a source?
(5) Consistency question: Is A consistent with what other experts assert?
(6) Backup evidence question: Is E's assertion based on evidence?

As can be seen from the list above, one of the critical questions accompanying the scheme probes into the matter of the expert's trustworthiness. In attempting to answer this question, our scheme for trust from expert opinion can clearly be deployed. Most plainly, the trust scheme could be used to provide a positive answer to the trustworthiness question of the general scheme above in that the conclusion of the trust scheme confirms that the expert in question has indeed been validated as being a reliable source. For example, if critical question 4 above has been posed against an instantiation of argument from expert opinion concerning some purported medical expert E, then a response might be that E is a doctor who can be trusted as being a reliable source as concluded by the trust from expert opinion scheme. Of course, the conclusion of the trust from expert opinion scheme is itself open to critical questioning, as we discuss later in Section 3. Although this example demonstrates how a specific issue from a more general scheme can readily be employed to probe into the issue and as such suggests some connections between schemes, currently we do not wish to suggest any kind of taxonomic classification of schemes due to the inherent difficulties of adequately covering the many possible ways in which schemes can interact and relate. However, we are open to the possibility of our specialised schemes for trust being used in conjunction with more general ones where issues of trust arise but are not the central concern of the scheme, as is the case of Walton et al.'s scheme for expert opinion.

Returning to our categories of argument schemes, the third contains two schemes, moral nature (MN) and social standing (SS), that are based on *A*'s observations of *B*, but observations that are not directly linked to trust. *A* might use the MN scheme to infer that *B*, whom *A* has observed to be very correct in their dealings with others, will act in a trustworthy way. Similarly, *A* might use the SS scheme to infer that *B*, who is a pillar of the local community, will not default on a loan. In neither case does *A* have any information about *B*'s trustworthiness – that is, *B*'s past behaviour or creditworthiness – but is prepared to use aspects of *B* for which *A* does have information as a proxy for such evidence.

The next category contains four schemes: majority (M), mutual beliefs (MB), mutual goals (MG), and organisation (Org). We can think of the majority scheme as a way of extrapolating any of the previous schemes. If we can show that any of those schemes for deriving trustworthiness apply to a suitably high proportion of a given population for which we do have evidence, we may be happy to infer that some member of the population for which we do not have evidence is also trustworthy. The mutual beliefs, mutual goals, and organisation schemes are grouped with majority because all four of them are about deriving trust in an individual without very detailed knowledge of that individual. The majority scheme does this on the basis of the class the individual falls into, and the organisation scheme does exactly the same, but on the basis of the individual falling into the specific class defined by members of the organisation. The mutual beliefs and mutual goals schemes use knowledge of some aspect of an individual to place them into an appropriate category and infer trust from that.

The final category contains the schemes of pragmatism (Pra) and prudence (Pru). Neither of these makes inferences that have much to do with the individual in whom trust is inferred. Prudence determines trust by assessing the comparative risk of trusting versus not trusting where the reasoning is about the situation not the possibly trustworthy individual. Pragmatism derives trust from a consideration of *A*'s goals and whether they are better served by trusting *B* or not, not on any information about whether *B* is trustworthy in their own right.

Two other schemes we considered, but rejected, as possible schemes are 'precedent', where *A* has trusted *B* before and decides *B* can be trusted again because there were no bad consequences from the previous time(s) *B* was trusted, and 'default', where *A* decides that *B* can be trusted despite having no evidence, perhaps with something in mind like 'tit-for-tat' (Axelrod, 1984)[8] or 'innocent until proven guilty'. We did not list these schemes because we think they are examples of schemes we have already incorporated in our list of schemes rather than new schemes. Precedent is a form of direct experience, and as we shall see, one of the critical questions that applies to direct experience addresses *B*'s past behaviour. We consider default to be a form of the majority scheme. After all, if it was not the case that an agent believed that the majority of individuals it was going to interact with would be trustworthy, then the default scheme would not be rational. (In other words, the default scheme is only rational for populations for which the majority scheme suggests trusting individuals.)

Figure 1 shows a visual characterisation of our argumentation schemes that is based on a conceptual model of trust showing a trustor (A) and a trustee (B) interacting in an environment and in a society. *Pragmatism*, based on self-interest, is an attribute of the trustor and *moral nature* is an attribute of the trustee. *Expert opinion* links the trustee to the environment and *prudence* (viewed as suitability of decision-making) links the trustor to the environment. The trustor has *direct experience* of the trustee and may share *mutual beliefs* and *mutual goals* with them. The trustor also has *indirect experience* of the trustee via the environment and may know the *organisation* that the trustee belongs to. A trustee has *social standing* in the society and may be in a position of *authority* over the trustor. The society includes agents with resemblance to the trustor and trustee, respectively, and supports distinct trust relationships: *reputation* via the former and *majority* via the latter.

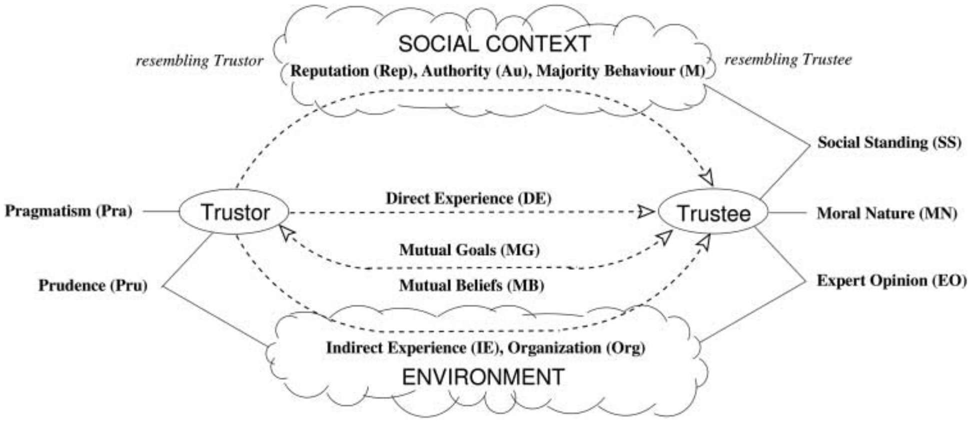

Figure 1. Categorising argumentation schemes for trust.

The above formulation establishes a form of *conceptual completeness* for our schemes. Given the conceptual model of trust, we have schemes that capture all pairwise relationships between the elements of the conceptual model.[9]

3. Critical questions for trusting

As we pointed out above, according to Walton et al. (2008), the role of critical questions is to capture the defeasibility inherent in argumentation schemes. Critical questions are:

> … questions that can be asked (or assumptions that are held) by which a non-deductive argument based on a scheme might be judged to be (or presented as being) good or fallacious. (Walton et al., 2008, p. 15)

We currently think of critical questions in two ways, as prerequisites for an argument scheme to apply and as the basis of possible arguments against the argument made by the scheme. We discuss this dual view more in Section 6.

Here we list critical questions for each of the schemes presented above, each phrased so that if all questions for a given scheme are answered ‘yes’, then the scheme can be used to create a plausible argument. There are also two general questions, which can be posed to any argument scheme that we listed above. These general questions are the following:

T.CQ1 Is this context one in which we have established that *B* is trustworthy?
T.CQ2 Are the negative consequences of misplaced trust sufficiently small that we can discount the possibility that *B* is not trustworthy?

Most of the critical questions relate to *belief* – is it likely that *B* is or is not trustworthy? In many cases, the *utility* of trusting *B* needs to be taken into account as well, and T.CQ2 is one way to begin to capture that. The utility of trusting *B* will presumably be positive if *A* decides to trust *B* and *B* proves to be trustworthy, because *B* will do the action that *A* trusted them to do, and this is presumably something that has utility for *A*. Similarly, if *A* trusts *B* and *B* turns out to be untrustworthy, this is because *B* does not do the action that *A* trusted them to do, and *A* has then not gained the utility it hoped to gain from *B* performing the action, while having paid the opportunity cost of not doing something else that might have gained *A* some benefit. Yet, *A* will also have gained negative experience of *B* to inform further modification of *A*’s trust in *B*.

Having listed these general questions, we turn to scheme-specific questions.

Direct experience There are two critical questions for arguments based on direct experience.

DE.CQ1 Is *B*, the person in whom we are placing our trust, really the person we have experience of?

DE.CQ2 Are we sure that *B* has not been acting in a trustworthy way in order to obtain our trust and then deceive us?

The first of these questions aims to identify the situation in which we have a long series of interactions with *B* and learn to trust them. Then we are presented with a new interaction with someone who claims to be *B* but turns out to be *C*, someone we have never interacted with before (*C*, for example, may have intercepted our communication with *B* and decided to impersonate them, or they may have kidnapped *B*). The second question identifies a 'bait-and-switch' scenario in which *B* has acted as if they were trustworthy in order to later betray our trust. (This is a scenario discussed by Salehi-Abari & White (2010), among others.)

Note that the first question, as we have described it, covers the case where *C* steps in to deceive *A* by pretending to be someone, *B*, that *A* already trusts. The question also covers an alternative scenario where *C* has pretended to be *B* during a series of interactions with *A* so that *A* believes that the person that they trust and have potentially recommended to others, *C*, is actually *B*. This scenario, of course, is the main plot device in the play *Cyrano de Bergerac* (1897), though in that case Cyrano is intending to act in a selfless way.[10] The difference between DE.CQ1 and DE.CQ2 is that the latter is intended to identify situations where *B*, the one building up the trust, was always untrustworthy, but hiding it well. DE.CQ1, in contrast, covers the situation in which *B* really was trustworthy, but another agent, *C*, stepped in and abused that trust.

Indirect experience Again there are two critical questions:

IE.CQ1 Can trust be inferred from the evidence?

IE.CQ2 Is *B*, the person in whom trust is being inferred, really the person who should be trusted?

The first of these addresses the fact that evidence is uncertain and so, to reprise our restaurant example, the fact that there is a large number of people waiting outside a restaurant might not indicate that the food is good (the service might be slow, or the restaurant might be very trendy, or a fire alarm might have caused the diners to evacuate). The second question makes a finer distinction. If the food in the restaurant is good, is this an indication that we should trust food provision by the restaurant owner (so that we can infer that other restaurants owned by the same person are also good), or by the chef (so that if the chef moves to a different kitchen, we should infer that food from that source is good too)?

Expert opinion Here we have several critical questions, the first of which is analogous, though distinct, from the identity question for direct and indirect experience – Is *B* really an expert?

EO.CQ1 Do we have proof that *B* is really an expert?

EO.CQ2 Is *B*'s expertise relevant in this case?

EO.CQ3 Is *B* suitably credible as an expert (Walton et al., 2008, p. 15)?

EO.CQ4 Is *B*'s opinion consistent with that of other experts?

EO.CQ5 Do we know *B* will not benefit as a result of the opinion they stated?

The difference between the first question here, EO.CQ1, and DE.CQ1 is that with DE.CQ1, we are interested in *who B is*, namely whether they are the person we have learnt to trust. With expert opinion, we are vesting all our belief about trustworthiness, which in the direct and indirect experience schemes we have established ourselves, in some certifying body. EO.CQ1 challenges this and is really a two-part challenge: Is the certificate earner really an expert and is the certificate possessor really the certificate earner? EO.CQ2 addresses the fact that *B* might be indisputably an expert, but just not the best expert (*B* may be a doctor, e.g. a dermatologist, but in some very specific medical context, e.g. dermatology; although *B* might have some idea of the best course of action for treating a broken leg, *B* might not be as good an expert as a doctor who specialises in that specific context, e.g. an orthopaedist). EO.CQ3 does not ask whether *B* is the right kind of doctor, but whether we think that *B* is a qualified doctor (maybe based on where they studied), or, as another example, not whether *B* is a lawyer, but whether *B* is an experienced lawyer. EO.CQ4 is a check that *B* does not hold a maverick opinion. EO.CQ5 seeks to question whether *B*'s views have some benefit to *B*: we might not trust a restaurant critic's view of the food at a particular establishment if we knew they were being paid by the restaurant to write their review; though this fact would not necessarily negate *B*'s opinion if we generally trusted *B*'s moral nature.

Authority As we said when we introduced the argument schemes, the argument from expert opinion and the argument from authority are similar. However, the critical questions show some of the differences between them.
Au.CQ1 Is *B* really in a position of authority?
Au.CQ2 Is *B*'s authority relevant in this case?

We assume that expertise can be certified – that is the reason, after all, that doctors hang their medical certificates on their walls, mechanics hang their 'authorised dealer' notifications, and restaurants in New York City are required to display the hygiene rating they were awarded by the NYC Department of Health. Authority, on the other hand, may sometimes be certified (by a uniform) but in other cases may be very hard to prove. The question about relevance is exactly the same as in the expert opinion scheme, and the specific context in which the authority (in this case) is operating is all important.

Reputation Deriving trust from reputation requires that *B* has a good reputation and some assurance that reputation means something for the situation at hand:
Rep.CQ1 Does *B* have a good reputation?
Rep.CQ2 Are we sure that *B*'s reputation has not been manipulated to make it more positive?
Rep.CQ3 Is *B*'s reputation relevant in this case?

Moral nature The questions for this scheme are derived from the critical questions that Walton et al. (2008) give for the ethotic argument (ethotic arguments are those based on the character of the person putting the argument forward):
MN.CQ1 Is *B* a person of good moral character?
MN.CQ2 Is character relevant in this case?
MN.CQ3 Is the degree of trustworthiness being inferred supported by the evidence?

Social standing Social standing is only a guarantee if *B* has significant social standing and there is a mechanism by which standing can be lost. The critical questions address this.
SS.CQ1 Does *B* have any social standing to lose?
SS.CQ2 Does *B* value social standing?
SS.CQ3 Would *B*'s standing be dented by being exposed as untrustworthy?
SS.CQ4 If *B* is untrustworthy, can they be exposed in a meaningful way?

The questions address the following issues. First (SS.CQ1), social standing is only a deterrent if *B*'s peers will see exposure as reflecting badly on *B*. Second (SS.CQ2), if *B* does not care about social standing then nothing can be inferred from the potential loss of it. Third (SS.CQ3), if *B* has already been exposed as untrustworthy, *B* has nothing much to lose, and so standing is no guarantor of trustworthiness. Finally (SS.CQ4), social standing is only a deterrent if it is possible to make the members of *B*'s social circle aware of the loss of trustworthiness. The assumption exposed by this final critical question means that social standing is not necessarily much help as an argument scheme in an online environment where tying reputation to an individual is complicated by the ease of maintaining anonymity and acquiring a new identity.

Majority Since the majority scheme is a form of statistical argument, the need to consider Simpson's paradox[11] forms the basis of a natural critical question M.CQ2:
M.CQ1 Is *B* really in the class of individuals who are trusted?
M.CQ2 Is the class we are considering the most specific class that *B* is a member of?

As an example of M.CQ2, we might be prepared to trust online merchants in general on the basis of the majority scheme, but, on the basis of some bad experiences, might not be prepared to trust online sellers of electronics in particular.

Prudence Since the prudence scheme is about the risk of trusting *B*, the critical questions focus on this aspect:
Pru.CQ1 Is it riskier to not trust *B* than it is to trust *B*?
Pru.CQ2 Is it possible to accurately estimate the risk in trusting and not trusting *B*?
Pru.CQ3 Is there another individual whom we could trust where the risk would be lower than trusting *B*?

Pragmatism The critical questions for the pragmatism scheme focus on the degree to which trusting *B* is in the best interests of the trusting agent:
Pra.CQ1 Does trusting *B* serve our best interests?
Pra.CQ2 Is there another individual whom we could trust such that trusting them would better serve our interests than trusting *B*?

Recall that the pragmatism scheme is intended to capture cases where *A* is in a situation, or believes that they are in a situation, where it is necessary to trust someone, and is trying to identify that someone. Thus the critical questions are less concerned with establishing 'Do I have evidence that suggests it is sensible to trust *B*?' than answering 'If I have to trust someone, is trusting *B* better than trusting other people?'.

Mutual Goals The critical questions for the mutual goals scheme hinge on the relevance of the goals and whether they are shared between *A* and *B*:
MB.CQ1 Are mutual goals relevant in this case?
MB.CQ2 Does *B* really have the same goals as *A*?

Mutual Beliefs The critical questions for mutual beliefs mirror those for mutual goals:
MB.CQ1 Is mutual belief relevant in this case?
MB.CQ2 Does *B* really have the same beliefs as *A*?

Organisation The critical questions for the organisation scheme probe the relevance of the organisation and *B*'s membership of it (as with direct and indirect experience, *B* could be an imposter):
Org.CQ1 Is the organisation relevant in this case?
Org.CQ2 Is the organisation a source of trustable individuals?
Org.CQ3 Is *B* really from the relevant organisation?

The critical questions for MG, MB, and Org are somewhat similar. The first, in each case, is a question about the applicability of the scheme, which is different from questioning context. It is a question about the scheme – is this a situation in which the scheme can be reasonably applied – rather than about the trustor (*B*), since context is a question about the kind of trust being (possibly) placed in *B*. The remaining question(s) then test whether the scheme can be applied in this case, asking whether *B* has the right goals or beliefs to be trusted, or comes from the right organisation to be trusted.

Note that, as mentioned above, all the critical questions discussed in this section are variations on what Walton et al. (2008, p. 93) call the 'trustworthiness question'. In our context, where we are putting trust under the microscope, it makes sense to split the trustworthiness question into these more specific questions tied to specific schemes.

4. Argument schemes for propagation

Under some circumstances, it is possible to perform inference about trust. For example, if *A* trusts *B* and *B* trusts *C*, then it is often reasonable to infer that *A* should trust *C*. This kind of reasoning has been widely studied in the literature on trust. To pick a few examples, Cheng, Govindan, and Mohapatra (2011), Hang, Wang, and Singh (2009), Kuter and Golbeck (2007), and Sun, Yu, Han, and Liu (2005) all provide schemes for taking numerical estimates of the trust that *A* has in *B*, and the trust that *B* has in *C*, and from these computing the numerical measure that one should associate with the trust that *A* has in *C*. At the same time, Francone and Castelfranchi (2010) argue that this kind of propagation is only possible under a very specific set of circumstances. Other schemes for propagation can be found in Guha, Kumar, Raghavan, and Tomkins (2004). In Section 4.1, we describe these propagation schemes as argument schemes. As with the schemes in the previous section, there may be good arguments against these forms of inference, and we capture those in a set of critical questions in Section 5.

4.1. *The schemes*

In the trust literature, inference about trust is often formulated in the context of a *trust graph* – a graph in which nodes are individuals and labelled, directed links between nodes indicate the amount of trust between individuals. We follow this tradition and illustrate each of the forms of inference as a graph as well as a text description.

Direct propagation (DP) This is a common approach to trust propagation that one finds in the literature: if *A* trusts *B* and *B* trusts *C*, then *A* may decide to trust *C* (Figure 2(a)).

As mentioned above, we find this form of propagation discussed in Cheng et al. (2011), Hang et al. (2009), Kuter and Golbeck (2007), and Sun et al. (2005).

Trusting takes place in a specific context, and as Jøsang et al. (2006) point out, in this kind of inference, we sometimes mix functional trust and referral trust. It is often the case, for example, that *B* trusts *C* to do a specific kind of thing, which is functional trust, and *A* infers that *C* can be trusted to do this because *A* trusts *B* to make good recommendations, which is referral trust. In the context of the restaurant recommendation example from Section 2, if I trust Bob as a judge of people, and Bob tells me that Camilla is a good judge of restaurants, I might decide to start judging restaurants by what Camilla says about them. Because my trust in Camilla is restricted to her knowledge of restaurants, I will not necessarily take any suggestions she might make about where to get my car serviced. I might, however, take Camilla's recommendations about where to go for a drink, on the grounds that her knowledge about bars is suitably close to her proven area of expertise (knowledge about restaurants). In contrast, if Bob tells me that Deborah knows a good mechanic, then because I know that Bob can be trusted on referrals, I may decide to take any suggestion that Deborah makes about where to get my car serviced (regardless of whether I know anything about Deborah directly).

In addition, as mentioned above, Francone and Castelfranchi (2010) argue that this kind of propagation – which Francone and Castelfranchi (2010) call *transitivity*[12] – is only reasonable

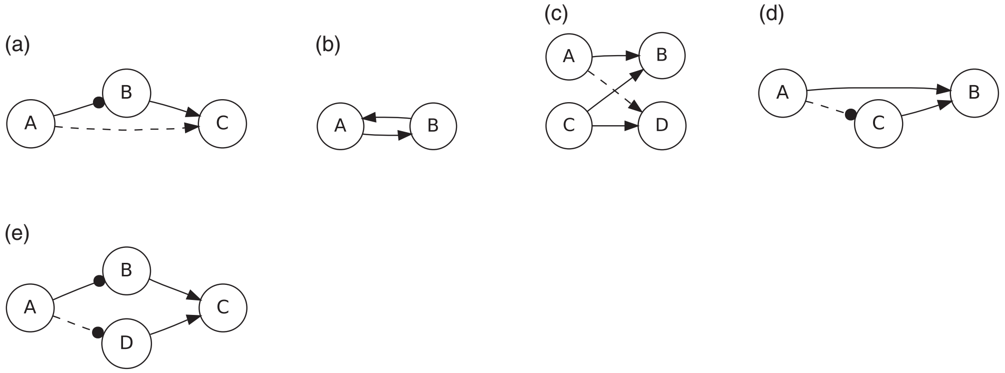

Figure 2. A graphical rendition of trust propagation: (a) direct propagation; (b) reciprocity; (c) co-citation; (d) co-implication; and (e) trust coupling. The solid arcs are the base relations, and the dotted arcs are the inferred relations. Arcs that end with dot indicate referral trust, and arcs that end in an arrowhead indicate functional trust.

under especially restrictive conditions more restrictive than those discussed in Jøsang et al. (2006). In particular, in our terminology, Francone and Castelfranchi (2010) say that if A trusts B in some context, and B trusts C in the same context, and A believes that B trusts C in that context, then A trusts C. The context in which A trusts B and that in which B trusts C are allowed to differ so long as they are close enough. If I trust Bob on the subject of restaurants and Bob trusts Camilla on the subject of bars, I might decide that these two areas of expertise are close enough that I am willing to trust Camilla on the subject of bars. We will revisit the role of context in our discussion of critical questions.

Reciprocity (Rec) Reciprocity says that if A trusts B, then B may decide to trust A (Figure 2(b)).

Reciprocity is suggested as a method of trust inference in Guha et al. (2004) under the name ‘transpose trust’, though it is not found to be a very strong factor in the empirical work in Guha et al. (2004). This is not really surprising if we consider that trust is all about being able to predict the future actions of other agents. An agent that acts somewhat unreliably (thus making it difficult for others to develop trust in it) may have no difficulty learning to trust another agent that acts much more reliably.[13] Context, of course, adds to the reasons that trust does not have to be reciprocal. While I have been building up my trust in Bob’s ability to recommend restaurants, I might not have given him any reason to trust me on the same subject (he might think all my recommendations are lousy, I may not have given him any, and in an extreme case – for example, in the restaurant context, the case in which I read his food blog and never comment on it – Bob may not know who I am). Also, if I am asking for recommendations in the first place, then chances are I do not know myself – which is another reason why Bob may not trust me on this topic. Here, context also plays a role. I might ask Bob about restaurants in a city that I have never visited but one that Bob knows well, which would only cause Bob to doubt my knowledge of restaurants in that particular city. Indeed, my query may make him think that I am a foodie, because I care enough to task for a restaurant recommendation to begin with; which may make Bob trust me to recommend restaurants to him in cities that I know well.

Co-citation (CC) In co-citation, if A trusts B and C trusts B, and C also trusts D, then A may decide to trust D (Figure 2(c)).

Co-citation is another form of trust inference from Guha et al. (2004), and Guha et al. (2004) assemble empirical evidence that this form of reasoning agrees rather well with people's intuition about trust. (To be more precise, Guha et al., 2004, took numerical information about the trust between A and B, C and B, and C and D, and from this computed a numerical estimate of the trust A has in D. This was found to provide a reasonable match for the trust that experimental subjects said A should have in D.)

This pattern of inference suggests that if A and C both trust the same individual, B, then A may feel that C is a reliable source of information about whom to trust (since they agree on B), and so accepts C's implicit recommendation about D. In terms of the restaurant recommendation example, if I trust Bob to make good restaurant recommendations, and I know that Camilla also trusts Bob, then I might start to think Camilla knows some good sources of restaurant information, so that when she raves about Dave's ability to pick good places to eat, I might be inclined to listen to Dave.

Co-implication (CI) In co-implication, if A and C both trust B, then A may decide to trust C (Figure 2(d)).

The idea of co-implication, which is a novel scheme that we are proposing here, came about by considering how trust values are computed in co-citation, that is, by considering the way that they appear to be computed in Guha et al. (2004). In co-citation, the propagation of values goes from A to B to C to D, with, as described above, the fact that A and C share a common opinion of B leading to A taking C's approval of D as a recommendation of D. Co-implication takes one less step, ending not with the inference of A's trust in D, but rather the explicit inference that A trusts recommendations made by C.

As an example of co-implication, consider that I learn that both Camilla and I trust Bob's views on restaurants, and so I start to believe that Camilla is a good source of referrals because she trusts the same people that I do.

Trust coupling (TC) The name 'trust coupling' comes from Guha et al. (2004), which expresses the form of inference (casting it into our terms) as 'if A trusts B who trusts C, and if D trusts C, then A may decide to trust D' (Figure 2(e)).

We can think of trust coupling as a combination of co-implication and direct propagation. If I trust Bob and Bob trusts Camilla, then I trust Camilla – that is direct propagation. If I trust Camilla and Dave trusts Camilla, then I trust Dave – that is co-implication. Of course, in this description we are playing fast and loose with context. Trust propagation makes sense as long as the context is consistent. If I trust Bob as a recommender of restaurant critics, then I will trust Camilla's restaurant reviews if Bob says they are trustworthy. If Dave agrees with Bob about Camilla's reviewing prowess, then I may think that Dave is also trustable as a recommender of restaurants.

4.2. *Discussion*

All the propagation schemes listed above, except co-implication, which, to the best of our knowledge is stated here for the first time, can be found described in the trust literature. However, elsewhere these schemes are not discussed as argument schemes but in terms of mechanisms for computing numerical trust values for the relations indicated in Figure 2 by dotted lines. Of course, this does not mean that they are universally accepted. Even direct propagation, which is the most widely discussed mechanism, is understood not to be universally applicable. Rather it only applies in a rather restrictive context which we capture by the use of critical questions (see Section 5).

The difference between these argument schemes and the ones in Section 2 is that the latter were all schemes to identify when trust can be established between agents in the absence of any

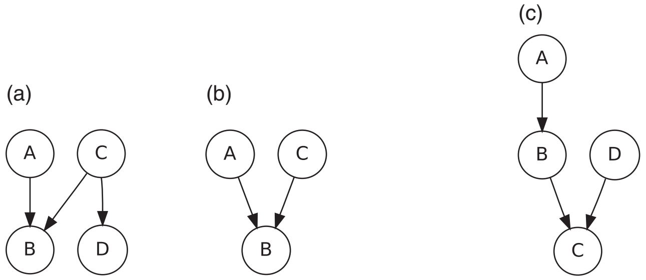

Figure 3. Probabilistic networks in which intercausal reasoning may take place.

information about trust, whereas the propagation schemes are about how to combine existing information about trust relations between agents. In other words, considering trust graphs like those in Figure 2, the argument schemes of Section 2 tell us what solid links exist – which of these basic relationships can be established – while the propagation schemes generate additional connections built on top of the basic relations, providing us with the ability to infer new links.

In addition, when we deal with propagation of trust, we have to be careful with the context of the trust. As we have already mentioned, via Jøsang et al. (2006), propagation involves a combination of trust in an individual with regard to some behaviour – functional trust in that individual – and the trust in an individual to make trustworthy recommendations – referral trust. In direct propagation and co-citation, we infer functional trust. In co-implication and trust coupling we infer referral trust. Figure 2 represents functional and referral trust differently to clarify what is being inferred in each case. Figure 2(a) also clarifies why we do not use the term 'transitivity' for direct propagation: since the relation between A and B is different than that between B and C, the inference of a relation between A and C is a different process than that permitted by the usual mathematical notion of transitivity.

It is worth noting that the form of inference at the heart of co-citation, co-implication, and trust coupling is analogous to *intercausal reasoning* in probabilistic networks (Koller & Friedman, 2009). Intercausal reasoning is perhaps best explained using an example like that in Figure 3(b). If A and C are both causes of B, then *a priori* if either A or C becomes more likely, then so does B. However, if B is known to be the case, then any increase in the probability of A or C cannot increase the probability of B (since that is fixed at 1). Instead, if the probability of A, for example, increases, the probability of C decreases. Thus, to take the classic example (Pearl, 1988), both rain (A) and the operation of the sprinkler (C) make it more likely that the grass is wet (B). However, if I know the grass is wet, then believing that the sprinkler was on makes it less likely that it was rain that made the grass wet.

As described, the relationship between A and C (in Figure 3) is the form of intercausal reasoning known as *explaining away*. Evidence that makes one cause of B more likely provides an explanation for B and simultaneously makes the other cause of B less likely. This is arguably the most common form of intercausal reasoning, but forms in which evidence for one cause makes the other *more* likely are also possible (Druzdzel & Henrion, 1993; Wellman & Henrion, 1991).

If the trust networks in Figure 2(c)–(e) are considered as probabilistic networks, then the corresponding forms of inference make sense in probabilistic terms. In co-implication, Figures 2(d) and 3(b), A's trust in B and C's trust in B together increase A's trust in C. In co-citation, Figures 2(c) and 3(a), the same reason as in co-implication is followed by propagation from C to D. Finally, in trust coupling, Figures 2(e) and 3(c), propagation from A to B precedes the intercausal reasoning of co-implication.

5. Critical questions for propagation

Here we offer some general questions for considering trust propagation schemes. These are, in fact, variations on the questions CQ1 and CQ2 from Section 2. Again it is necessary that the context is applicable and that we take into account the possible negative effects of misplaced trust. However, in Section 2, it was possible to extract these general questions because the pattern of reasoning in the schemes – *A* establishes trust in *B* – was always the same. In trust propagation, however, the individual in which *A* is investing trust will vary from scheme to scheme, so we have to have specific versions of the questions for every scheme.

In addition, since many of the propagation schemes involve referral trust, there is another question that applies to all such trust propagation schemes, namely:

TP.CQ1 Are the individuals who are making referrals in this propagation trustworthy as recommenders?

This is really a meta-question, since the individual who is carrying out the propagation needs to assess trust in the individual who is making the recommendation using exactly the methods that we discussed in Section 2. Thus, in the case of direct propagation, if *A* has been told by *B* that *C* is a good mechanic on the basis of *B*'s direct experience of *C*, not only should the critical questions T.CQ1, T.CQ2, DE.CQ1, and DE.CQ2 from Section 2 be applied to *B*'s trust in *C*'s ability as a mechanic, but *A*'s trust in *B*'s ability to provide a referral should be questioned as well. Thus, if *A*'s trust in *B* is also based on direct experience, the critical questions T.CQ1, T.CQ2, DE.CQ1, and DE.CQ2 should be applied to *A*'s assessment of *B*'s ability to make referrals. Of course, despite the fact that a version of this question applies to every scheme, once again we have to state these questions separately for each scheme.

Direct propagation In the direct propagation scheme, *A* chains its trust in *B* from *B*'s trust in *C*. There are six critical questions that relate to the application of the scheme:

DP.CQ1 Does *A* trust *B*?

DP.CQ2 Does *A* know that *B* trusts *C*?

DP.CQ3 Is *B* trustworthy in the context of referrals? This is the meta-question discussed above – to answer 'yes' it is necessary to answer 'yes' to all the relevant critical questions from Section 3 that apply to the means by which *A* established referral trust in *B*.

DP.CQ4 Is the context of *B*'s trust in *C* a context in which *B* has established that *C* is trustworthy?

DP.CQ5 Are the negative consequences of misplaced trust in *C* sufficiently small that we can discount the possibility that *C* is not trustworthy?

DP.CQ6 If *A*'s referral trust in *B* was established by the application of other propagation schemes, is the overall length of the referral chain lower than the maximum acceptable?

DP.CQ1 and DP.CQ2 check that the relations necessary for the scheme to apply are in place. DP.CQ3 and DP.CQ4 are versions of T.CQ1 and T.CQ2 applicable to the direct propagation scenario. In essence, they make sure that the relations that the scheme operates on – the relation between *A* and *B*, and the relation between *B* and *C* – are in place. DP.CQ3 is worded to reflect the fact that referral trust is just a specific form of trust. DP.CQ5 is the scenario-specific version of T.CQ2, checking that there is not some potentially disastrous consequence of trusting *C* that would overrule the decision to trust them. DP.CQ6 stems from the idea that chains of trust referral eventually become too long to be tenable. The idea is discussed, for example, in Hang et al. (2009) and DP.CQ6 frames this idea as a critical question.

Reciprocity The reciprocity scheme infers *B*'s trust in *A* from *A*'s trust in *B*. There are three critical questions:

Rec.CQ1 Does *B* know that *A* trusts it in the relevant context?
Rec.CQ2 Does *B* have any reason to trust *A* in the same context as *A* trusts *B*?
Rec.CQ3 Are the negative consequences of misplaced trust in *A* sufficiently small that we can discount the possibility that *A* is not trustworthy?

Rec.CQ1 asks whether *B*, the agent from whose perspective the scheme is applied, knows if *A* trusts it – if this is not established then the scheme cannot be applied. Rec.CQ3 is the scenario-specific version of CQ2. Rec.CQ2 is more complex. Reciprocity is one of the propagation schemes in that it does not include referral trust. *A*'s trust in *B* will have been established by one of the schemes in Section 2, and for reciprocity to hold, *B* should be able to establish trust in *A* in the same way. Rec.CQ2 is a meta-question that asks if any of the schemes in Section 2 hold.

Co-citation Co-citation has six critical questions:

CC.CQ1 Does *A* trust *B*?
CC.CQ2 Does *A* know that *C* trusts *B* and *D*?
CC.CQ3 Is the context of *A*'s trust in *B* the same as *C*'s trust in *B*?
CC.CQ4 Is the context of *C*'s trust in *D* the same as the context that *A* is inferring in *D*?
CC.CQ5 Is the context of *A*'s trust in *B* one in which *B* has been found to be trustworthy?
CC.CQ6 Are the negative consequences of misplaced trust in *D* sufficiently small that we can discount the possibility that *D* is not trustworthy?

CC.CQ1 and CC.CQ2 check if all the relations that are required by the scheme are in place. *A* will obviously know whether it trusts *B* (though this may require inference), but, as with DP.CQ2 and Rec.CQ1, the other relations may exist but *A* may be ignorant of them, or they may not exist but *A* may believe that they do.

In co-citation, the inference of trust in *D* makes most sense if the context of all the initial trust relations is the same. If I trust Bob to pick restaurants and Camilla trusts Bob and Dave to pick restaurants, then it is reasonable that I might trust Dave. Some limited difference in context also seems reasonable. If I trust Bob to pick restaurants and Camilla trusts Bob to pick restaurants and Dave to pick bookstores, then I might trust Dave to pick bookstores (on the grounds that Camilla, since we agree on Bob, generally has sound judgement about whom to trust). These aspects are captured by CC.CQ3–CC.CQ5 which instantiate T.CQ1 for this scenario. CC.CQ6 is then the scenario-specific version of T.CQ2.

Co-implication Co-implication has five critical questions:

CI.CQ1 Does *A* trust *B*?
CI.CQ2 Does *A* know that *C* trusts *B*?
CI.CQ3 Is *B* trustworthy within the context of referrals?
CI.CQ4 Is the context of *A*'s trust in *B* the same as *C*'s trust in *B*?
CI.CQ5 Are the negative consequences of misplaced trust in *C* sufficiently small that we can discount the possibility that *C* is not trustworthy?

As with previous schemes, CI.CQ1 and CI.CQ2 establish the existence of the required trust relations, CI.CQ3 and CI.CQ4 capture CQ1 for this scenario, and CI.CQ5 captures T.CQ2.

Trust coupling Trust coupling has six critical questions which combine those for direct propagation and co-implication.

TC.CQ1 Does *A* trust *B*?
TC.CQ2 Does *A* know that *B* and *D* trust *C*?

TC.CQ3 Is B trustworthy in the context of referrals?

TC.CQ4 Do B and D trust C in the same context?

TC.CQ5 Is the context of B and D's trust in C a context in which B and D have established that C is trustworthy?

TC.CQ6 Are the negative consequences of misplaced trust in D sufficiently small that we can discount the possibility that D is not trustworthy?

In the critical questions for trust coupling, TC.CQ1 and TC.CQ2 check that the necessary relations exist, TC.CQ3, TC.CQ4, and TC.CQ5 together pose the relevant version of CQ1 for this scenario, and TC.CQ6 poses the relevant version of T.CQ2.

As already noted, one important difference between the critical questions for propagation and the critical questions for developing trust between agents is that in the propagation cases, the structure of the schemes differ; whereas the trust development cases are all about A developing trust in B. This is the reason that we cannot write general critical questions that apply to all schemes, and instead have to have scheme-specific versions of general questions like 'Are all the necessary relations in place?'.

Another, perhaps more subtle, difference is in the form of information tested by the critical questions. In the trust development schemes, the questions were all about what A knows. Depending on that knowledge, A may infer trust in B or not. In the propagation schemes, the agent who is inferring trust (A in all cases other than reciprocity where it is B) not only has to examine what it knows, but also its knowledge of other agents' trust. In addition, many of the schemes involve relying on another agent's ability to make referrals, and these schemes involve[14] two different trust contexts – in an application of the DP scheme, for example, A trusts B in the context of referrals, and B trusts C to, for example, repair a car.

6. Discussion

It is relatively well-established that argument schemes can be interpreted as schemes for generating arguments, for example, as in Atkinson, Bench-Capon, and McBurney (2006). Thus, our schemes for trust identify situations in which trust might be inferred. If A knows that B is in a position of authority, for example, then A can construct an argument that A should trust B. It is less clear how critical questions should be interpreted. This is recognised in work such as Wyner (2012), which discusses the formal semantics of critical questions and advocates the expression of critical questions as auxiliary premises of schemes that would be best made explicit.

For our purposes we can see two possible options. One is that critical questions are a form of prerequisite for the application of an argument scheme. Since argument schemes are defeasible, this interpretation of the critical questions gives them a role rather like the prerequisites in a default rule in default logic (Reiter, 1980). That is, when the critical questions are false, the argument scheme does not hold, and no inference can be made using it. When the critical questions are true, the argument scheme can be applied, but there is no guarantee that the inference will not subsequently be over-turned. Under this interpretation, the inference that A should trust authority B should only be made if it can be verified that B is really in a position of authority, and that B's authority is relevant in the case at hand. Otherwise, no conclusion should be drawn about B's trustworthiness. In some cases, a more gradual version of this interpretation might be applicable. In this case, the more critical questions are true, the more likely the argument scheme is to hold; and the more critical questions are false, the more likely the scheme is to not hold.

The second option for interpreting critical questions is to take the questions as the basis for additional arguments against the one put forward by the scheme. In this case, the critical questions do not prevent the application of the scheme, or even necessarily overcome the argument that is generated by the scheme, rather the critical questions just provide one or more additional arguments that are to be considered. Under this interpretation, A is free to conclude that authority B is trustworthy however the critical questions are answered; but if, for example, B's authority is not found to be relevant, then this will constitute an argument against B being trustworthy that must be weighed against the argument for B being trustworthy in some process for determining the acceptability of all the relevant arguments being considered by A. A's overall view of B's trustworthiness will then depend on the weights and interactions of all the relevant arguments.

As it stands, all the schemes we have discussed are all related to trust as opposed to *distrust*. That is, all are about reasons for having a positive attitude about another agent, having reasons to think what they say is true, and what they say they will do will actually happen – either generating trust in an individual or propagating that trust – with the critical questions capturing the cases where these schemes do not apply. To fully specify reasoning with trust, we would need to consider two additional kinds of scheme.

The first relates to *distrust*. This is a topic that has been the subject of much less research than trust, but there is still a body of work that we can rely on to guide us. In particular, McKnight and Chervany (2001) summarise some of the literature and provide some useful guidelines. In particular, they discuss the relationship between trust and distrust, suggesting that distrust is the negative of trust. If trust in agent B is a belief that B is reliable, then a lack of trust in B is a lack of belief that B is reliable. Distrust in B is a belief that B is not reliable, so that things that B says are true will turn out to be false, and things that B says that they will do, turn out not to be done. We can easily imagine malicious reasons for B to (not) do these things, because B wishes us harm. But, equally, the distrust may just be because B does not care. A lack of trust in B does not suggest distrust – we may have no opinion about B's trustworthiness – and, equally, a lack of distrust does not imply the existence of trust.

McKnight and Chervany suggest that one deals with trust and distrust as essentially separate – rejecting the idea that trust and distrust lie on the same scale, for example, with trust as value 1, distrust as value -1, and 0 indicating a lack of trust. Rather, trust and distrust are reasoned about separately, with certain individuals the subject of arguments for being trusted and distrusted. The only relationship between trust and distrust, then, is that the same individual cannot be both trusted and distrusted in the same context. For example, it is not possible to both trust that one's teenage child will arrive on time and distrust that they will arrive on time, but it is possible to trust that she will cook great food when she promises to make dinner and distrust that she will arrive on time.

Thinking of distrust as the negative of trust, we can easily imagine schemes for reasoning about the existence of distrust in B that are variants on the schemes we have already discussed. Direct experience says that if I have had lousy food in a restaurant on several occasions then I will distrust the restaurants ability to produce good food. Equally, I might distrust a politician's position on women's rights because of an application of the organisation scheme to the knowledge that they are a member of a right-wing political organisation that has a track record of attacking support for working mothers, and it seems plausible that all the schemes we have discussed could be adapted in this way. Moreover, mere distrust, even when proven, in another individual is not necessarily a reason not to deal with them. The Internet protocol TCP, for instance, is designed to circumvent the known unreliability of packet delivery over unreliable Internet connections, and there may be equally pragmatic ways to make use of untrustworthy individuals. We leave the full exploration of argument schemes for distrust for another paper.

The second additional kind of scheme is one that bears on decisions about trust and distrust without directly being an argument for trusting or distrusting. For example, consider the use of *ad hominem* arguments (Walton, 1998), in this context. *Ad hominem* arguments are arguments of the general form 'B is a bad person, so anything provided by B is not good'. In some cases such arguments are fallacious – the argument that Bill Clinton's healthcare proposals were bad because he had an extra-marital affair is usually regarded as a fallacious argument – but in the case of trust, certain kinds of *ad hominem* arguments can be convincing. An *ad hominem*-derived scheme related to trust might be one in which if A has direct experience of B in which B was not reliable, then A might decide not to trust B. This scheme is one for A not trusting B, and we can imagine a scenario in which this scheme is used to construct an argument for not trusting B that attacks (in the usual sense of the term in argumentation semantics, Baroni, Caminada, & Giacomin, 2011) an argument for trusting B that was established by one of the schemes we listed above. The line between such schemes and those for trust and distrust may be narrow since we can imagine a stronger scheme that says if A has direct experience of B in which B did something to actively injure A's interests, then there should be an argument that supports A's decision to distrust B (where distrust is interpreted in the sense discussed above).

7. Related work

The work presented in this section draws on two specific topics, argumentation schemes and trust, on which a significant amount of literature exists. While this paper is the first to take a comprehensive approach to combining the two areas, we focus here on the existing work from the two separate areas.

7.1. *Related work on trust*

As discussed in Section 1, varying definitions of trust have been offered in the literature. We are concerned with the issue of reasoning about trust from a specifically computational perspective. It has long been recognised in the multi-agent systems literature, for example, Castelfranchi and Falcone (2000) and Ramchurn, Huynh, and Jennings (2004), that the autonomy afforded to intelligent agents brings with it uncertainty regarding whom can be trusted or not within multi-agent interactions. There are many different ways in which trust can be classified. For example, trust can be directed towards individuals (such as an interaction partner) or an overall system (such as an auction governed by rules) (Erriquez, 2012; Ramchurn et al., 2004). Trust information may derive from direct interaction with others, indirect interactions, e.g. from reputation information available, or from some third party authority (Huynh, Jennings, & Shadbolt, 2006). As can be seen in the schemes and critical questions we have presented in this paper, we cover these different facets of trust.

Other distinctions such as *basic trust* (modelling a general disposition based on an agent's experiences), *general trust* (directed at a particular agent but without considering a particular situation), and *situational trust* (directed towards a particular agent in a particular situation) have also been set out in the literature (Marsh, 1994). Indeed, Marsh (1994) served as one of the first and most prominent accounts of computational modelling of trust and, as such, provides a detailed formal model of trust. In contrast, our schemes are intended to provide patterns of reasoning that can easily be applied in multiple situations and are broad enough to cover the different scenarios in which reasoning about trust is of importance; though as we showed in Section 2, we can account for some of these distinctions in our general characterisation of our schemes.

More recent work has made the link between trust and argument-based representations. Work by several of the authors has suggested that argumentation, with its ability to capture the reasons for its conclusions and the data from which those conclusions were inferred, is a natural formalism for reasoning about trust (Parsons et al., 2010). Further work in this line (Parsons et al., 2011; Tang, Cai, McBurney, Sklar, & Parsons, 2012) has provided logic-based representations to reason about trust, and it is shown how this can be combined with reasoning about beliefs. Although the formal systems presented in Parsons et al. (2011) and Tang et al. (2012) are incomplete, they make progress towards showing how trust can be represented as a form of argumentation and the work has led to a prototype reasoning system (Tang, Cai, Sklar, & Parsons, 2011) which has continued to evolve (Parsons, Sklar, Salvitt, Wall, & Li, 2013).

There is a growing interest in the use of argumentation to handle trust beyond that described above. For example, Villata et al. (2011) use meta-level argumentation to capture the trust an agent has in an information source, while Stranders et al. (2008) use a fuzzy approach, and Matt, Morge, and Toni (2010) suggest how arguments may be combined with statistical data to augment existing trust models. The scheme-based representation we have presented in this paper tackles the issue of trust from a more flexible perspective than any of the previous approaches to trust argumentation that we have mentioned here, since the definition of a set of schemes captures different nuances of the notion of trust that are conflated in other work, teasing out different aspects of trust rather than treating trust as a monolithic concept.

Two further approaches take a more abstract view. In Erriquez (2012) and Erriquez, van der Hoek, and Wooldridge (2011), a framework is set out that takes inspiration from Dung's abstract argumentation frameworks (Dung, 1995) whereby a graph theoretic model is used to capture the *distrust* relations within a society, and this model is used to formulate notions of mutually trusting coalitions. In Dung's framework, nodes in a graph represent arguments, but in Erriquez (2012) and Erriquez et al. (2011) they denote agents, and in Dung's framework edges denote a binary attack relation, whereas in Erriquez (2012) and Erriquez et al. (2011) they represent a distrust relation between agents that is used in determining which coalitions are acceptable. Similar in scope is Harwood, Clark, and Jacob (2010) which uses argumentation to decide on trustworthiness, building links between agents, labelling them with 'trust' or 'distrust', and applying something like Dung-style semantics to decide whom to trust. While these accounts of trust are inspired by well-understood representations from computational argumentation, they remain at the abstract level, without delving as deeply into the specifics of arguments about trust as is done in the scheme-based approach that we have presented.

7.2. *Related work on argumentation schemes*

We now turn to discuss the literature related to our chosen form of representation, argumentation schemes. Argument schemes were born out of the literature on informal logic, and the most influential work in this area, at least as far as work on computational argument is concerned, is the work of Walton et al. (2008), which we have cited extensively above. This book provides a solid introduction to the use of argument schemes and catalogues a large number of them. We consider these schemes to be very general in that they are not fitted to a specific domain. As a result, they have wide applicability, but do not capture all the subtleties or peculiarities of specific domains. As a result, as we have pointed out above, some of the schemes and associated critical questions in this paper can be viewed as sub-classes of the schemes and critical questions from Walton et al. (2008). We envisage that our schemes could be used in conjunction with the existing more general ones as a way to probe into issues of trust that may arise as part of the critical questioning.

There are also a number of works which, like this paper, identify argument schemes for specific domains. For example, Bench-Capon and Prakken (2010) and Wyner and Bench-Capon (2007) consider argument schemes for legal reasoning, Reed and Walton (2005) discuss argument schemes for agent communication, Ouerdane, Maudet, and Tsoukias (2008) consider argument schemes for decision support, and Tolchinsky, Modgil, Cortes, and Sanchez-Marre (2006) look at argument schemes for deliberation in the sense of Walton and Krabbe (1995), that is, the process by which several entities reach a combined plan for action. In this same line of work, Prakken (2005) presents the case for using argument schemes as an alternative to logic as a means of knowledge representation (again focusing on the legal domain).

More recently, researchers have become interested in transforming argument schemes into computational versions to enable them to be used in systems that perform automated reasoning. For example, some of the authors of this paper developed a scheme for practical reasoning (Atkinson et al., 2006) that has been widely used to facilitate reasoning about what to do in a number of different problem scenarios. Schemes concerning witness testimony and expert opinion have also been used in computational argument (see, for example, Gordon, Prakken, & Walton, 2007), and there have been several attempts to capture argument schemes from legal reasoning in various forms of logic (Bex, Prakken, Reed, & Walton, 2003; Prakken, 2010; Verheij, 2003). The reason that such schemes have attracted the attention of researchers in computational argumentation is that their defeasible nature makes them naturally attractive to the non-deductive form of reasoning that many people are interested in capturing in argument-based approaches. Transferring the natural language schemes into a computational account poses challenges that cover a number of considerations.

Firstly, the adequacy of any representation concerning the different elements that the schemes comprise (such as facts, subjective judgements, causal theories, and so on) needs to be demonstrated. This point has been recognised in work such as Wyner, Atkinson, and Bench-Capon (2012) where a functional language for a computational analysis of schemes is set out. Further, it is shown in Wyner et al. (2012) how argumentation schemes expressed in the functional language given can be systematically related to one another, which is an issue that arises when it is desirable to use multiple schemes in any one setting.

The issue of representing the interconnections between schemes has also been tackled through specifications for a World Wide Argument Web (Rahwan, Zablith, & Reed, 2007) that uses Walton's theory of argumentation schemes (Walton et al., 2008). As such, an ontology is presented to enable the representation of networks of arguments on the Semantic Web.

Another consideration when making the schemes computational is the treatment of critical questions. As noted earlier, according to Walton et al. (2008), the role of critical questions is how the defeasibility inherent in argumentation schemes is recognised. The critical questions are intended to be employed as part of a dialogue whereby the answers to the critical questions can provide a reason for rejecting the conclusion of the scheme. In the computational account of the practical reasoning argumentation that is given in Atkinson et al. (2006), the questions are used to identify attacks that can be made on an argument derived from instantiating the scheme whereby the attack arises due to an assertion made in answering a critical question. For example, consider a critical question such as 'Does the action have the stated consequences?'. When this question can be answered in the negative and an assertion can be made that the action has consequences different to those stated in the instantiation of the scheme, then this can constitute an attack on the argument instantiating the scheme.

8. Conclusion

This paper has taken a first step towards identifying argument schemes for reasoning about trust. The overall aim behind deriving these schemes is to provide a computational mechanism for establishing arguments about trustworthiness from a description of some scenario that does not itself include any explicit information about trustworthiness. We have identified 13 general schemes that allow an individual to establish trust in another, and a further five schemes that allow trust to be propagated between individuals. For each of these schemes, we have also identified a set of critical questions. The purpose of the critical questions is to identify cases in which the schemes might not apply – if the answer to any critical question is negative, then the scheme may not apply (or may be used to derive a lower level of trust).

A natural question to ask of such a set of schemes is whether it and the associated set of questions is exhaustive. This, of course, is hard to establish, and we believe that a full set of schemes will only emerge over time. This is certainly the case in the work of Walton, whose schemes for arguments on the basis of expert opinion have continued to develop, for example, from those listed in Walton et al. (2008) to those in Gordon et al. (2007). In terms of the exhaustiveness of the critical questions, we take a more pragmatic approach. Our current work focuses on formalising these schemes, as in Bench-Capon and Atkinson (2010), as a precursor to being able to build them into a tool for reasoning about trust (Parsons et al., 2013). As we progress with the formalisation, we can ensure that questions are associated with every predicate and object in the set of formal schemes, making sure that, as desired, every aspect of a given scheme can be tested against the scenario in which it might be applied to check that it is valid.

As already suggested, Bench-Capon and Atkinson (2010) provide one approach to formalising argument schemes. Hunter (2008), though not directly writing about argument schemes, provides another approach which uses meta-level argumentation from Wooldridge, Parsons, and McBurney (2005) (though he cites a different version) to write down logical statements about what arguments are acceptable, given the person putting them forward. This approach can be extended, as in Sklar, Parsons, and Singh (2013), to make similar statements about which arguments are acceptable given a combination of the information used in the argument, and the individual who supplied that information. The patterns of reasoning formalised in Sklar et al. (2013), then, are very close to critical questions, and a generalisation of these patterns is a central issue in our current work.

This formalisation, then, is one line of future work, and we have already hinted at a second – to extend the set of schemes to include schemes for distrusting an individual along the lines discussed in Section 6. We anticipate that these schemes will include both ‘negative’ versions of the schemes (in the spirit of Walton’s similar ‘positive’ and ‘negative’ versions of schemes), which seem to be appropriate for capturing distrust, and schemes that are used to attack another agent’s claim or position, and other schemes that do not directly concern trust but are related to it, such as witness testimony. Another line of future work will be to explore the connections, if any, between the five propagation schemes and modal logics of relations: the diagrams of Figure 2 bear a striking similarity to those in Popkorn (1994) illustrating the correspondence theory between modal logic and the algebra of relations. Finally, a major part of our future work will involve the formal specification of our schemes and their associated critical questions to enable them to be made fully computational.

Acknowledgements

The authors would like to thank the anonymous reviewers of previous versions of this paper. The comments of these reviewers were very helpful in improving the paper. We would also like to thank the participants at the 2012 Conference on Computational Argumentation (COMMA), especially Trevor Bench-Capon and Tom Gordon, for their comments on an earlier version of this paper.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory, the National Science Foundation, the National Security Agency, or the US Government. The US Government is authorised to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

Funding

This research was funded under Army Research Laboratory Cooperative Agreement Number W911NF-09-2-0053 by the National Science Foundation [Grant #1117761] and by the National

Security Agency under the Science of Security Lablet grant (SoSL). Additional funding was provided by a University of Liverpool Research Fellowship and by a Fulbright-King's College London Scholar Award.

Notes

[1] Subjective probability having a natural interpretation as a propensity to make bets at particular odds (Jaynes, 2003, p. 655).

[2] As we shall see, this is just one possible pattern of trust propagation.

[3] 'Comping' is food industry slang for providing food and drink at no cost, typically to friends or family of the providers. The *Urban Dictionary*, http://www.urbandictionary.com, suggests the term is derived from 'complimentary'.

[4] The MTA is the body that runs the public transportation system in New York City.

[5] http://www.yelp.com.

[6] Recall that propagation is when A takes the step that says A trusts B and combines it with information that B trusts C to infer something about the relation between A and C. Reputation, in contrast, is a mechanism by which A may establish trust in B from information given to A by D and E.

[7] This is not necessarily a very strong argument for trusting, and the cinephile reader may recognise in this example a minor plot device from the classic 1973 movie *The Sting* (http://www.imdb.com/title/tt0070735/) where *(spoiler alert!)* the failure of the scheme to predict trust was important.

[8] Where it can be a winning strategy to start out trusting others to be cooperative.

[9] There may, of course, be other schemes that capture aspects of the same relationships. Indeed, this paper contains a number of schemes that were not present in Parsons et al. (2012), even though that set of schemes was also conceptually complete in the same way as this one is.

[10] In the play *(spoiler alert!)*, Cyrano is in love with his cousin Roxanne, but believes that his appearance means that Roxanne can never love him. Roxanne, meanwhile, has a crush on Christian, one of Cyrano's comrades in the Gascon Cadets, and Christian is also in love with Roxanne. Christian, who is unable to express his love for Roxanne, enlists Cyrano to help him, and Cyrano composes letters and speeches that Christian can use to woo Roxanne. It is as a result of these communications, ostensibly from Christian, but actually from Cyrano, that Roxanne falls in love with and marries Christian. Much later, once Christian has died a heroic death, the deception is revealed, and Roxanne realises that it is Cyrano that she has loved all along. Sadly Cyrano is dying by the time that this occurs. One might, of course, argue about Cyrano's motives. Perhaps he is acting nobly because he thinks that his love for Roxanne can never be requited. Perhaps he is just using any means he can to be able to interact intimately with her.

[11] Formally, Simpson's paradox Blyth (1972) is that

$$\Pr(A|B) < \Pr(A|\neg B)$$

but

$$\Pr(A|B, C) \geq \Pr(A|\neg B, C),$$
$$\Pr(A|B, \neg C) \geq \Pr(A|\neg B, \neg C).$$

Less formally, to take the widely quoted example, Simpson's paradox is how the baseball player David Justice can have a higher batting average than Derek Jeter for each year in the period 1995–1997, and yet have a lower batting average over the three years combined (Pavlides & Perlman, 2009; Ross, 2004). To see how this can be the case, it is necessary to know that a batting average in baseball is the number of 'safe hits' that a player attains in some period divided by the number of 'at-bats', that is, the number of chances to have a safe hit. Justice's batting averages are 0.253, 0.321, and 0.329 for the three years, and 0.298 when aggregated over the three years, Jeter's are 0.250, 0.314, 0.291, and 0.300, respectively. The paradox arises because the number of at-bats varies from year to year. Justice had many at-bats in

1995 and 1997, but many fewer in 1996 as the result of a shoulder injury. Jeter had only a few at-bats in 1995, when he was a rookie, and many more in 1996 and 1997.

12. We prefer the term ‘direct propagation’ since if one considers the context in which trust is being applied, we can view this as inferring a functional trust relation between A and C on the basis of a functional trust relation between B and C and a referral trust relation between A and B. We discuss this more below.
13. As parents of teenagers will know, the fact that your children fail to do the chores that you ask them to do will not stop them from relying on you to do things like providing them with regular meals.
14. More correctly these schemes *typically* involve two different trust contexts since it is perfectly possible that the context of the functional trust that is being considered is the ability to make referrals, so that the entire propagation scheme is relating to referral trust.

References

Atkinson, K., Bench-Capon, T.J.M., & McBurney, P. (2006). Computational representation of practical argument. *Synthese, 152*(2), 157–206.

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

Baroni, P., Caminada, M., & Giacomin, M. (2011). An introduction to argumentation semantics. *The Knowledge Engineering Review*, *26*(4), 365–410.

Bench-Capon, T., & Atkinson, K. (2010). Argumentation schemes: From informal logic to computational models. In C. Reed & C. Tindale (Eds.), *Dialectics, dialogue and argumentation: An examination of Douglas Walton’s theories of reasoning and argument* (pp. 103–114). London: College.

Bench-Capon, T.J.M., & Prakken, H. (2010). Using argument schemes for hypothetical reasoning in law. *Artificial Intelligence and Law, 18*, 153–174.

Bex, F.J., Prakken, H., Reed, C., & Walton, D.N. (2003). Towards a formal account of reasoning about evidence: Argumentation schemes and generalisations. *Artificial Intelligence and Law, 11*, 125–165.

Blyth, C.R. (1972). On Simpson’s paradox and the sure-thing principle. *Journal of the American Statistical Association, 67*(338), 364–366.

Castelfranchi, C., & Falcone, R. (2000, January). *Trust is much more than subjective probability: Mental components and sources of trust*. Proceedings of the 33rd Hawaii international conference on system science, Maui, Hawai’i, IEEE Computer Society.

Cheng, N., Govindan, K., & Mohapatra, P. (2011, April). *Rendezvous-based trust propagation to enhance distributed network security*. Proceedings of the 30th IEEE international conference on computer communications (INFOCOM), Shanghai, China.

Dong, X.L., Berti-Équille, L., & Srivastava, D. (2009, August). *Integrating conflicting data: The role of source dependence*. Proceedings of the 35th international conference on very large databases, Lyon, France.

Druzdzel, M.J., & Henrion, M. (1993). Intercausal reasoning with uninstantiated ancestor nodes. In *Proceedings of the 9th conference on uncertainty in artificial intelligence*, San Mateo, CA (pp. 317–325). Morgan Kaufmann.

Dung, P.M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games. *Artificial Intelligence, 77*, 321–357.

Erriquez, E. (2012). *Computational models of trust* (PhD thesis). Department of Computer Science, University of Liverpool, Liverpool.

Erriquez, E., van der Hoek, W., & Wooldridge, M. (2011). An abstract framework for reasoning about trust. In *Proceedings of the 10th international conference on autonomous agents and multiagent systems (AAMAS 2011)*, Taipei (pp. 1085–1086).

Francone, R., & Castelfranchi, C. (2010, September). *Transitivity in trust: A discussed property*. Proceedings of the Undicesimo Workshop Nazionale ‘Dagli Oggetti agli Agenti’, Rimini, Italy.

Gambetta, D. (1990). Can we trust them? In D. Gambetta (Ed.), *Trust: Making and breaking cooperative relations* (pp. 213–238). Oxford: Blackwell.

Gordon, T.F., Prakken, H., & Walton, D. (2007). The Carneades model of argument and burden of proof. *Artificial Intelligence, 171*(10–11), 875–896.

Grandison, T., & Sloman, M. (2000). A survey of trust in Internet applications. *IEEE Communications Surveys and Tutorials, 4*(4), 2–16.

Guha, R., Kumar, R., Raghavan, P., & Tomkins, A. (2004). *Propagation of trust and distrust*. Proceedings of the 13th international conference on the World Wide Web, New York.

Hang, C.-W., Wang, Y., & Singh, M.P. (2008). *An adaptive probabilistic trust model and its evaluation*. Proceedings of the 7th international conference on autonomous agents and multiagent systems, Estoril, Portugal.

Hang, C.-W., Wang, Y., & Singh, M.P. (2009). *Operators for propagating trust and their evaluation in social networks*. Proceedings of the 8th international conference on autonomous agents and multiagent systems, Budapest, Hungary.

Harwood, W.T., Clark, J.A., & Jacob, J.L. (2010). Networks of trust and distrust: Towards logical reputation systems. In D. M. Gabbay and L. van der Torre (Eds.), *Logics in security*. Copenhagen.

Hunter, A. (2008, July). *Reasoning about the appropriateness of proponents for arguments*. Proceedings of the 23rd AAAI conference on artificial intelligence, Chicago, IL.

Huynh, T.D., Jennings, N.R., & Shadbolt, N.R. (2006). An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems, 13*(2), 119–154.

Jaynes, E.T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Jøsang, A., Gray, E., & Kinateder, M. (2006). Simplification and analysis of transitive trust networks. *Web Intelligence and Agent Systems, 4*(2), 139–161.

Katz, Y., & Golbeck, J. (2006). *Social network-based trust in prioritized default logic*. Proceedings of the 21st national conference on artificial intelligence, Boston, MA.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: MIT Press.

Kreisler, H. (1999). Politics, values, and the separation of powers: Conversation with Abner Jay Mikva. April 12. Part of *Conversations with history*, Institute of International Studies, University of California, Berkeley, CA. Retrieved from http://globetrotter.berkeley.edu/people/Mikva/mikva-con0.html

Kuter, Y., & Golbeck, J. (2007). SUNNY*: A new algorithm for trust inference in social networks using probabilistic confidence models*. Proceedings of the 22nd national conference on artificial intelligence, Vancouver.

Li, L., & Wang, Y. (2010). *Subjective trust inference in composite services*. Proceedings of the 24th AAAI conference on artificial intelligence, Atlanta, GA.

Marsh, S.P. (1994). *Formalising trust as a computational concept* (PhD thesis). Department of Computing Science and Mathematics, University of Stirling, Stirling.

Matt, P.-A., Morge, M., & Toni, F. (2010, May). *Combining statistics and arguments to compute trust*. Proceedings of the 9th international conference on autonomous agents and multiagents systems, Toronto, Canada.

McKnight, D.H., & Chervany, N.L. (1996). *The meanings of trust* (Working Paper 96-04). Carlson School of Management, University of Minnesota, MN.

McKnight, D.H., & Chervany, N.L. (2001). Trust and distrust definitions: One bite at a time. In R. Falcone, M. Singh, & Y.-H. Tan (Eds.), *Trust in cyber-societies: Integrating the human and artificial perspectives* (pp. 27–54). Lecture Notes in Computer Science 2246. Berlin: Springer-Verlag.

Mui, L., Moteashemi, M., & Halberstadt, A. (2002). *A computational model of trust and reputation*. Proceedings of the 35th Hawai'i international conference on system sciences, Hawai'i.

Naylor, S. (2005). *Not a good day day to die: The untold story of operation Anaconda*. New York: Berkley Caliber Books.

Ouerdane, W., Maudet, N., & Tsoukias, A. (2008, September). *Argument schemes and critical questions for decision aiding process*. Proceedings of the second international conference on computational models of argument, Toulouse, France.

Parsons, S., Atkinson, K., Haigh, K., Levitt, K., McBurney, P., Rowe, J., …, Sklar, E. (2012). *Argument schemes for reasoning about trust*. Proceedings of the 4th international conference on computational models of argument, Vienna, Austria.

Parsons, S., McBurney, P., & Sklar, E. (2010, May). *Reasoning about trust using argumentation: A position paper*. Proceedings of the workshop on argumentation in multiagent systems, Toronto, Canada.

Parsons, S., Sklar, E., & McBurney, P. (2011). *Using argumentation to reason with and about trust*. Proceedings of the 8th international workshop on argumentation in multiagent systems, Taipei, Taiwan.

Parsons, S., Sklar, E., Salvitt, J., Wall, H., & Li, Z. (2013). *Argtrust: Decision making with information from sources of varying trustworthiness (demonstration)*. Proceedings of the 12th international conference on autonomous agents and multiagent systems, Minneapolis, MN.

Pavlides, M., & Perlman, M. (2009). How likely is Simpson's paradox. *The American Statistician, 63*, 226–233.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Popkorn, S. (1994). *First steps in modal logic*. Cambridge: Cambridge University Press.

Prakken, H. (2005). AI & law, logic and argument schemes. *Argumentation, 19*, 303–320.

Prakken, H. (2010). On the nature of argument schemes. In C.A. Reed & C. Tindale (Eds.), *Dialectics, dialogue and argumentation. An examination of Douglas Walton's theories of reasoning and argument* (pp. 167–185). London: College.

Rahwan, I., Zablith, F., & Reed, C. (2007, July). Laying the foundations for a World Wide Argument Web. *Artificial Intelligence, 171*(10–15), 897–921.

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

Reed, C., & Walton, D. (2005). Towards a formal and implemented model of argumentation schemes in agent communication. *Autonomous Agents and Multi-Agent Systems*, *11*(2), 173–188.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence, 13*, 81–132.

Resnick, P., & Zeckhauser, R. (2002). Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. In M.R. Baye (ed.), *The economics of the Internet and e-commerce* (pp. 127–157). Amsterdam: Elsevier Science.

Ross, K. (2004). *A mathematician at the ballpark: Odds and probabilities for baseball fans*. New York: Pi Press.

Rostand, E. (1897). *Cyrano de Bergerac*. Play.

Sabater, J., & Sierra, C. (2005, September). Review on computational trust and reputation models. *AI Review, 23*(1), 33–60.

Salehi-Abari, A., & White, T. (2010). *Trust models and con-man agents: From mathematical to empirical analysis*. Proceedings of the 24th AAAI conference on artificial intelligence, Atlanta, GA.

Sklar, E., Parsons, S., & Singh, M. (2013). *Towards an argumentation-based model of social interaction*. Proceedings of the 9th workshop on argumentation in multiagent systems, Minneapolis, MN.

Stranders, R., de Weerdt, M., & Witteveen, C. (2008). Fuzzy argumentation for trust. In F. Sadri & K. Satoh (Eds.), *Proceedings of the eighth workshop on computational logic in multi-agent systems* (pp. 214–230). Lecture Notes in Computer Science 5056. Berlin: Springer Verlag.

Sun, Y., Yu, W., Han, Z., & Liu, K.J.R. (2005). Trust modeling and evaluation in ad hoc networks. In *Proceedings of the 48th annual IEEE global communications conference*, St Louis (pp. 1862–1867).

Surowiecki, J. (2005). *The wisdom of crowds*. New York: Anchor Books.

Sztompka, P. (1999). *Trust: A sociological theory*. Cambridge: Cambridge University Press.

Tang, Y., Cai, K., McBurney, P., Sklar, E., & Parsons, S. (2012). Using argumentation to reason about trust and belief. *Journal of Logic and Computation, 22*(5), 979–1018.

Tang, Y., Cai, K., Sklar, E., & Parsons, S. (2011, November). *A prototype system for argumentation-based reasoning about trust*. Proceedings of the 9th European workshop on multiagent systems, Maastricht, The Netherlands.

Teacy, W.T.L., Chalkiadakis, G., Rogers, A., & Jennings, N.R. (2008). *Sequential decision making with untrustworthy service providers*. Proceedings of the 7th international conference on autonomous agents and multiagent systems, Estoril, Portugal.

Tolchinsky, P., Modgil, S., Cortes, U., & Sanchez-Marre, M. (2006). CBR and argument schemes for collaborative decision making. In *Proceedings of the first international conference on computational models of argument*, Liverpool, UK (pp. 71–82).

Verheij, B. (2003). Dialectical argumentation with argumentation schemes: An approach to legal logic. *Artificial Intelligence and Law, 11*, 167–195.

Villata, S., Boella, G., Gabbay, D.M., & van der Torre, L. (2011). *Arguing about the trustworthiness of the information sources*. Proceedings of the European conference on symbolic and quantitative approaches to reasoning and uncertainty, Belfast, UK.

Walton, D. (1997). *Appeal to expert opinion: Arguments from authority*. University Park, PA: The Pennsylvania State University Press.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. Cambridge: Cambridge University Press.

Walton, D.N. (1998). *Ad hominem arguments*. Tuscaloosa, AL: University of Alabama Press.

Walton, D.N., & Krabbe, E.C.W. (1995). *Commitment in dialogue: Basic concepts of interpersonal reasoning*. Albany, NY: State University of New York Press.

Wang, Y., & Singh, M.P. (2006). *Trust representation and aggregation in a distributed agent system*. Proceedings of the 21st national conference on artificial intelligence, Boston, MA.

Wellman, M.P., & Henrion, M. (1991). Qualitative intercausal relations, or explaining 'explaining away'. In *Proceedings of the 2nd international conference on principles of knowledge representation and reasoning*, San Mateo, CA (pp. 535–546). Morgan Kaufmann.

Wooldridge, M.J., Parsons, S., & McBurney, P. (2005, July). *The meta-logic of arguments*. Proceedings of the 4th international conference on autonomous agents and multi-agent systems, Utrecht, The Netherlands.

Wyner, A., & Bench-Capon, T.J.M. (2007). Argument schemes for legal case-based reasoning. In *Proceedings of the 20th international conference on legal knowledge and information systems (JURIX)*, Leiden, The Netherlands (pp. 139–149).

Wyner, A.Z. (2012). Questions, arguments, & natural language semantics. In *Proceedings of the 12th workshop on computational models of natural argument*, Montpelier, France (pp. 16–020).

Wyner, A.Z., Atkinson, K., & Bench-Capon, T. (2012). Towards a formal language for argumentation schemes. In *Proceedings of the 9th international workshop on argumentation in multi-agent systems*, St Paul, MN (pp. 203–222).

Yu, B., & Singh, M. (2002). Distributed reputation management for electronic commerce. *Computational Intelligence, 18*(4), 535–349.