The Semantics of Untrustworthiness

Giuseppe Primiero[1] and Laszlo Kosolosky[2]

[1] Department of Computer Science,
Middlesex University, UK
G.Primiero@mdx.ac.uk
[2] Centre for Logic and Philosophy of Science,
Ghent University, Belgium
laszlo.kosolosky@ugent.be

Abstract. We offer a formal treatment of the semantics of both complete and incomplete mistrustful or distrustful information transmissions. The semantics of such relations is analysed in view of rules that define the behaviour of a receiving agent. We justify this approach in view of human agent communications and secure system design. We further specify some properties of such relations.

KEYWORDS: Trust, Mis- and Distrust, Information Transmission, Expertise, Secure System Design.

1 Introduction

Social epistemology, philosophy of computing, logic, game and network theories, software design are just some of the disciplines that have been struggling with the most elusive and, at the same time, interesting, epistemic concept of trust. Approaches to this notion are diverse in context, techniques and conceptual understanding.[3] Trust has a variety of possible interpretations from a psychological and sociological perspective, see e.g. Rotter (1971); Lewis, Weigert (1985); Shapiro (1987).[4] From an epistemic viewpoint, its definition can be qualified in view of companion notions as those of practical value, testimony, expertise, integrity, and it obviously has a huge relevance on the philosophical debate on knowledge, see e.g. Dalton (2001); Faulkner (2012); De Winter, Kosolosky (2012); Kosolosky (forthcoming); De Winter, Kosolosky (2013); Hardwig (1991); Audi (1997). From a formal viewpoint, part of the debate revolves around the difference in identifying trust as a first-order relation between agents (*'agent A trusts agent B'*) or as a second-order property of relations (*'relation X between agent A and agent B is trustworthy'*) and on that basis to determine the relevant formal structure, see e.g. Castelfranchi (2004); Demolombe (2004); Dastani et al. (2004); Herzig et al. (2010); Kramer er al. (2012); Primiero, Taddeo (2012).

[3] For a genealogical overview of the notion see Simpson (2012).

[4] See also the large list of references about trust in labour organizations available at http://www.ilocarib.org.tt/Promalco_tool/productivity-tools/manual07/m7_12.htm.

From a technical and technological viewpoint, trust is crucial in the context of the design of secure systems in cyberspace. It is often defined on the basis of some basic reputation algorithm and one of the main interests lies in defining relevant propagation methodologies and to constrain problematic properties such as transitivity, see e.g. Beth et al. (1994); Christianson, Harbison (1997); Kamvar et al. (2003); Guha et al. (2004).

Less explored, but certainly as much interesting, is the description of untrustworthy relations, i.e. relations qualified by a *negative assessment of trust*. In this context, the first remarkable condition is a widely spread confusion concerning the difference between distrust and mistrust. Such distinction is in general ignored when the underlying conceptual schemas do not allow for a proper clarification of the related notions (McKinght et al. (2000); Guha et al. (2004); Borgs et al. (2010)). In particular, when trust is identified as a first-order relation, distrust and mistrust relations cannot be understood as directly negative counterparts of the former. In many approaches distrust is reduced to a low or zero degree of trust (gans et al. (2001)), but the definition of distrust as direct negation of trust (Abdul-Rahman,Hailes (1997); Chen,Yeager (2003)) induces ignorance of the subtle semantic distinction with mistrust. In particular this is due to the fact that the negation of the relation $trust(A, B)$ (with A, B agents) as $\neg trust(A, B)$ means that no operation on content vs. epistemic state is possible, i.e. there is no possibility to separate distrust in *any* message from a given agent and distrust in *some* message. More interestingly, this makes it impossible to distinguish between not trusting purposefully false information vs. not trusting accidentally false information. Moreover, many undesired effects occur, among which transitivity obtained by multiplication: $\neg trust(A, B) \wedge \neg trust(B, C) \vdash trust(A, C)$.

In Taddeo (2010) and Taddeo (2010a) trust is defined as a second-order property characterizing first-order relations among agents. First and overall, a trust assessment is not always and not only an assessment of the trustor on the trustee; rather, it is a certain qualification of objective parameters (such as the environment in which this relation holds) and it characterizes a specific task evaluating the likelihood of achieving a specific goal. This evaluation can be obtained by reference to similar cases, for authority arguments, or any other chosen parameter. In order to conceptually capture this fine-grained sort of relation, the relation of trust can be modified from the basic first order $trust(A, B)$ to a second order $trust_A(Action_B)$. In this way, the trustor does not trust directly another agent but it rather trusts the operation that the trustee performs, with the possibility of qualifying *Action* in various ways. For the epistemic context generated by an information channel, trust qualifies the communication $trust_A(InformationTransmission_B)$ between the receiver A and the source of a certain information content B, or in a computational model between a client and a server. In this case one says that the channel including those two terminals and that specific information item is trusted.[5] This understanding of trusted

[5] Notice that while in the following we will be speaking of the actual information transmission from Bob to Alice, the definition of trust as second order relation does not in fact require the first order relation to be actualized: one could perfectly

communication is formalized in Primiero, Taddeo (2012) by a modal type theory which accounts for the two epistemic states involved: verification-terms on propositions for directly known contents, or information items available at both ends of the channel; partial-terms for communicated but not verified (hence, to be trusted) contents, or information items available only at one end and transmitted to the other.[6] Based on a model of trust as a second-order property of relations, one can ground a more fine-grained conceptual analysis of distrust and mistrust. The characterization of these notions will be obtained not as an all-purpose definition, i.e. we shall not attempt to provide definitions of distrust and mistrust as such. We will rather offer a more confined characterization of distrustful and mistrustful relations, in particular for the context of epistemic relations instantiated by channels of information transmission. We will refer in the following to a modular characterization:

- an information transmission that generates uncertainty in the receiver concerning the truthfulness of its content is *untrustworthy*;
- an information transmission that is deemed to convey unintentionally false information is *mistrustful*;
- an information transmission that is deemed to convey intentionally false information is *distrustful*;
- a receiver that assesses an information transmission as mistrustful or distrustful operates on its content accordingly.

Hence, our understanding of untrustworthy transmission will be based on the characterization of the semantic behaviour of the agent who qualifies a channel as distrustful or mistrustful. We are able to offer distinct procedural explanations of such qualifications and will also consider some of the related properties. Notice that our approach does not engage with the either rational or irrational reasons or propagation algorithm that lead an agent or principal to assess the trustworthiness of a channel with respect to an information item and another agent or principal; irrespective of those reasons, we rather consider how the agent behaves in view of such an assessment.

In the following we first characterize untrustworthy transmissions of either complete or incomplete information (Section 2); we then offer a semantic analysis of untrustworthiness (Section 3) and then specifically of distrust (Section 3.2), mistrust (Section 3.1) and cases combining the two (Section 3.3); we conclude by offering an overview of the most relevant properties of untrustworthy transmissions (Section 4).

well consider a property instead of a relation – like ($GoodMathematician_B$) and a second-order property $trust_A(GoodMathematician_B)$, which would actualize in any relation $trust_A(TheoremTransmission_B)$.

[6] It is here important to stress how this notion does not coincide with that of *reliance*, where one terminal *relies* on the other to perform some action without it knowing it or not.

2 Untrustworthy transmissions

Alice and Bob met in London, had a great weekend together and now Alice is preparing to go back to Brussels by train. Bob was given the task to check the timetable and provide that information to Alice.[7]

Bob: *'I checked the timetable, the train to Brussels leaves at 5pm'.*

We consider this information transmission as a first order relation between Bob and Alice: $InformationTransmission(B, A)$. Alice realizes that Bob does not remember that today the Railway Network updates to the Spring time. He is transmitting unintentionally false information, or misinformation.[8] Alice decides to *mistrust* Bob's transmission. Here the proper second order relation is at hand: $Mistrust_A(InformationTransmission(B, A))$. How should she reason in view of her assessment that the communication by Bob is *mistrustful*?

Moreover, Bob is deeply in love with Alice and wants her to miss the train so that she will spend one more day in London.

Bob: *'I regularly go to St. Pancras, the best way is to take a cab'.*

This is again a regular first order relation $InformationTransmission(B, A)$. Bob is now telling her that the cab would be faster, while he knows the Tube would be. He is transmitting intentionally false information, or disinformation. Alice knows enough of London's traffic jams to guess this is utterly false and she decides to *distrust* Bob's transmission. Here again, a second order relation is at hand: $Distrust_A(InformationTransmission(B, A))$. How should Alice reason in view of her assessment that the communication by Bob is *distrustful*?

A different example stems from system design. Bob is a principal (either human or automated)[9] asking the bank server Alice for a list of the movements on a given bank account number. To this aim, Alice requires Bob to provide a set of identification data: the user's birth date, a password and randomly generated PIN code accompanied by the serial number of the generator. Bob offers back three series of digits:

```
ALICE: Request:BIRTHDATE;PWD;PIN(SOURCE)
BOB: Enter:1103194_;rvcs132RT43;324564-676544(source:343434)
```

The format of the first and second series are not valid, as the password should include at least one symbol and the birth date misses a cipher. Can the server recognize this as a case of unintentionally false data message and ask the client to re-introduce the data? Assume instead that Bob is an attacker who is trying to force Alice's system by inputting the `BIRTHDATE` for a given user, trying to guess the `PWD` and using the attacker's own PIN generator to produce data for `source(KEY)`:

[7] In the following examples, we will use italic fonts when referring to human agents, and typewriter fonts when referring to mechanical principals.

[8] See Floridi (2011).

[9] In computer security, a principal is an entity that can be authenticated in information transactions.

```
ALICE: Request:BIRTHDATE;PWD;PIN(SOURCE)
BOB: Enter:13061955;rtts?672TR21;434367-878799(source:898989)
```

In this case, the tentative attack on Alice's system can be compared to sending intentionally false information, in the sense that it is not a *bona fide* error: Bob is offering data related to the user and combining it with a 'false' source for the `PIN`. As in many systems known to us, after a fixed number of wrong tries by Bob, Alice decides he is an attacker. How should Alice act in view of such assessment? Can we devise a semantics of actions leading to blocking the client?

We start by considering a first-order relation of communication (between Bob and Alice) characterized by second-order relations of mistrust and distrust.

Definition 1 (Information Channel). *An information channel is the physical or virtual relation instantiated by a communication act between a Sender S and a Receiver R such that a transmission of information contents from the former to the latter occurs on it.*

In the following, we shall refer to Bob as the Sender and to Alice as the Receiver. To provide a closer characterization of the relation instantiated by communication between the Sender and the Receiver, we offer a more fine-grained definition of the content transmitted by a communication act over an information channel.

Definition 2 (Complete Information Transmission). *A complete information transmission $\langle \texttt{Metadata}, \mathcal{G} \rangle$ consists of the information metadata functional to a goal $\mathcal{G}$ establishing that an information item A included in $\mathcal{G}$ is valid for the current information channel.*

The `Metadata` element in the transmission can be instantiated by various data, depending on the system design and the target for which the communication happens:

- a pair $\langle \texttt{procedure}, \mathcal{G} \rangle$ will instantiate a system where a verification procedure is required that justifies explicitly the validity of A in $\mathcal{G}$, for example an automated theorem prover;
- a pair $\langle \texttt{source}, \mathcal{G} \rangle$ will instantiate a model of testimony, where the authority of a client is supposed to suffice for the acceptance of the goal statement A *valid*, e.g. by offering the originator of the train schedule, or the code number of the random key generator;
- a pair $\langle \texttt{tags}, \mathcal{G} \rangle$ will instantiate a system where the goal expression is accompanied by identifying tags, relative e.g. to a location or timing ('updated at 4pm'; 'accessing from Brazil');
- a pair $\langle \texttt{user}, \mathcal{G} \rangle$ will instantiate a system where the goal expression is targeted for specific user groups, e.g. a cryptographic message with the mention of those users who have a specific decrypting key, or a scientific explanation of some chemical targeted for farmers (and not for pharmacists).[10]

[10] Such a schema $\langle \texttt{Metadata}, \mathcal{G} \rangle$ seems particularly apt to enrich the dynamics of design of online scientific databases so as to facilitate the selection of appropriate datasets

Our first task is to characterize *untrustworthy* information transmissions. A complete transmission as defined above is trustworthy if its content is assumed to include *correct metadata* and a *valid goal*. We will consider such a transmission error-free and associate it with a certainty state in the Receiver about the content A. Then an information transmission can be considered *untrustworthy* in a first sense if its content is deemed prone to *errors*. Hence, we proceed by characterizing information transmissions with errors:

Definition 3 (Transmission with errors). *A transmission with errors is such because:*

- *it includes incorrect* `Metadata` *relative to an information content A; the* `Metadata` *is then considered non-processable;*[11]
- *it includes an invalid content A; $\mathcal{G}$ declaring validity of A is then considered a non-attainable goal.*

This allows us to design two models of error production, see also Primiero (2013):

1. *wrong informational coupling*: an error in building the pair $\langle \texttt{Metadata}, \mathcal{G} \rangle$, where `Metadata` is inappropriate (i.e. of the wrong type or semantically non-apt to the task),[12] though possibly well-processed and therefore correct, relative to content A in $\mathcal{G}$.
2. *informational malfunctioning*: an execution error by which the processing of `Metadata` is incorrect for $\mathcal{G}$; if executed correctly, `Metadata` is indeed valid for content A in $\mathcal{G}$.

Hence a communication is untrustworthy by errors if it generates uncertainty in the Receiver about the validity of the goal or the correctness of the metadata.

Definition 4 (Untrustworthy Transmission of Complete Information). *An untrustworthy transmission of complete information is a ternary relation holding between the epistemic state of the sender S, the epistemic state of the receiver R and the information $\langle \texttt{Metadata}, \mathcal{G} \rangle$ such that the transmission by S generates uncertainty in R about correctness or validity of the pair $\langle \texttt{Metadata}, \mathcal{G} \rangle$.*

A (complete) transmission is thus deemed untrustworthy (second order) if the Receiver considers possible that the Sender produces an error in her knowledge state about $\langle \texttt{Metadata}, \mathcal{G} \rangle$ (first order relation of communication).

Consider now some slightly modified versions of the above examples.

specific to the purposes of given users. For an analysis of the epistemological issues related to this problem and the connected notion of distributed understanding, see e.g. Leonelli (forthcoming).

[11] The processing of metadata depends on its typology: for `procedure` it will be execution; for `source` it will be reachability; for `tags` it will be checking; for `user` it will be targeting.

[12] As an example, consider the pair $\langle \texttt{geotags}, \mathcal{PWD} \rangle$ for identification on access requests, where geolocalization is superfluous, while `procedure` or `source` would be appropriate.

> Bob: *'The train to Brussels leaves at 5pm. The best way to go to St. Pancras is to take a cab'.*

While this information transmission appears correct, Bob is offering no reason for his claims. As opposed to the above format of complete transmission, something is now missing, namely the metadata `procedure` by which Bob would make his claim valid. This case corresponds to a message of the form $\langle \texttt{procedure\_empty}, \mathcal{G}\rangle$.

Similarly in the second example:

```
ALICE: Request:BIRTHDATE;PWD;PIN(SOURCE)
BOB: Enter:13061955;rtts?672TR21;empty(source:empty)
```

Here the principal is letting the required information about the `PIN` and serial number of the code generator empty. This case corresponds to a message of the form $\langle \texttt{source\_empty}, \mathcal{G}\rangle$. We shall call cases of this form *incomplete transmissions.*[13]

Definition 5 (Incomplete Transmission). *A transmission is incomplete iff:*

- *it misses* `Metadata` *for* $\mathcal{G}$*;*
- *it misses the goal* $\mathcal{G}$ *for which given* `Metadata` *is offered.*

Accordingly, an information transmission can be considered *untrustworthy* in a second sense if its content is *incomplete*, thus inducing again an uncertainty state in the Receiver. We now proceed by characterizing incomplete information transmissions as untrustworthy.

Definition 6 (Untrustworthy Transmission of Incomplete Information). *An untrustworthy transmission of incomplete information is a ternary relation holding between the epistemic state of the sender S, the epistemic state of the receiver R and the information* $\langle \texttt{Metadata}, \texttt{empty}\rangle$ *or* $\langle \texttt{empty}, \mathcal{G}\rangle$*, such that the transmission by S generates uncertainty in R about the validity or the correctness of any pair* $\langle \texttt{Metadata}, \mathcal{G}\rangle$*.*

An incomplete transmission is hence deemed untrustworthy if the missing information cannot be analytically extracted from the received data. Our task in the following is to define the semantics of such untrustworthy complete or incomplete transmissions.[14]

[13] Converse cases of transmissions including metadata but no goal are also possible to formulate. An example would be a set of building instructions, or deductive steps, that miss a declaration of the building task or the theorem: $\langle \texttt{procedure} : \texttt{R1}, \texttt{R2}, \ldots, \mathcal{G}\texttt{\_empty}\rangle$.

[14] The present definition of untrustworthy transmission implies that mistrust and distrust cannot be defined in absence of communication. In other words, one can judge a source untrustworthy only with respect to a given information transmission. This is an immediate consequence of defining trust as a second order property of first order relations. Nothing, however, prevents from extending this framework in view

3 The Semantics of Untrustworthiness

Incorrectness or incompleteness are thus considered in the following the principal conditions for an untrustworthy transmission. This notion of untrustworthiness is further characterized as inducing uncertainty in the Receiver's epistemic state with respect to the information content of the transmission. Unfortunately, this is only a static characterization of the Receiver's state and it does not specify in any way the consequent course of action of the Sender. On such a basis, the only sensible specification would be to request suspension of any (complete or incomplete) information transmission which is deemed untrustworthy. In other words, Alice could only stop listening to Bob and ignore his messages; and the server could only forbid further attempts at access by the client. This solution appears highly unsatisfactory. Our aim in this section is to offer a more detailed procedural account of the Receiver's epistemic state involved in an untrustworthy transmission, based on an intentional characterization of the Sender's course of action. Once Alice decides that Bob's information transmission is mistrustful, respectively distrustful, how should she reason on the basis of the information she has been given? We shall analyse the semantics of untrustworthiness in view of executable procedural steps when conditions of an untrustworthy complete or incomplete transmission obtain and the intention of the Sender is assessed.

In the context of a semantic theory, *information* $\langle \mathtt{Metadata}, \mathcal{G} \rangle$ is true, meaningful, data. Our characterization of untrustworthy information as incorrect, invalid or incomplete data makes it, by definition, *false information* from the point of view of the receiver. The latter can be wrong in its assessment of the sender, and thus wrongly consider the information as false. But, as the present treatment proceeds from the viewpoint of the receiver's judgment of the source's reliability, for as far as the definition of untrustworthiness is concerned, we are dealing with false information. False information can be further identified in two intensional versions: *misinformation* as unintentionally false information and *disinformation* as intentionally false information.[15] In our model, this property necessarily amounts to an assessment of the Sender's intention by the Receiver, and we will not make any claim about how reasonable such assessment is, nor whether it is correct. Notice that the intentionality assessment by the Receiver does not

of memory-based agents who are able to assess trustworthiness of peers in view of *previous* communications. Notice, however, that even in the present treatment, the notion of (un)trust is not content-bounded, i.e. the formal treatment is not strictly dependent on tokens of information: the definition uses a data/metadata structure such that the assessment of untrustworthiness can be induced only by evaluation of the sender's properties.

[15] For a complete analysis of a theory of strongly semantic information, see Floridi (2011). According to such theory, information cannot be properly false. The problem of the veridicality of information content has been long debated and it is not a settled theoretical issue. For the present purposes, we shall not enter this debate and suffice to say that, in the following, we understand 'false information' as tantamount to false data with meaning. For the introduction of the notion of (un)intentionally false information see also Floridi (2011, p.260).

mean that the model accounts only for conscious beings as Senders. It seems reasonable to say that one way a machine can be said to transmit 'intentionally' false output data is if its program is meant to do precisely that, and that it transmits 'unintentionally' false data if this is only the result of a malfunctioning. Another account of the intentionality of mechanical principals is the one instantiated by our examples above: unintentionally false information is sent by way of *bona fide* mistakes by authorized clients; intentionally false information is sent by purposefully erroneous data intended to deceive another client or server. Now we can characterize channels in view of transmission of intentionally and unintentionally false information:

Definition 7 (Disinformative Channel). *A disinformative channel transmits intentionally false information contents from S to R.*

Definition 8 (Misinformative Channel). *A misinformative channel transmits unintentionally false information contents from S to R.*

Accordingly, we will characterize untrustworthy transmissions as being executed on either a disinformative or on a misinformative channel. Notice that a disinformative channel is not defined by intentional transmission of false information and, accordingly, a misinformative channel is not defined by unintentional transmission of false information. (Un-)intentionality of the transmission is an additional property that does not define the untrustworthiness of the channel.[16]

Our analysis will now specify the semantics of such untrustworthy channels in terms of the possible procedures executable by the Receiver involved in a transmission on either a mistrustful or distrustful channel, by way of specifying the admissible rule steps.

3.1 Mistrust

A clear connection between misinformation and mistrust is formulated as follows:

Definition 9 (Mistrustful Transmission). *An information transmission over a misinformative channel is characterized by a second order property of mistrust.*

To describe the logical behaviour of the Receiver involved in a mistrustful transmission, we relate the second-order property of mistrust to an operation of modal modification. A procedural semantics of modal modification can be informally explained by the use of a modal operator: given a well-defined set P of terms and the full formulation of conditions to be satisfied for a given term to be in P, a modal operator produces the set of 'possibly satisfied' conditions for P,

[16] In fact, we can think of an unintentional transmission of intentionally false information (e.g. the wrongful selection of a REPLY-ALL method in an email communication to transmit a consciously formulated excuse to miss a meeting), as well as an intentional transmission of unintentionally false information (e.g. the correctly addressed email to my boss, where I claim I will be missing the meeting this Friday because of a research workshop in Germany, while I meant in the UK).

i.e. the judgement that for at least one condition one cannot claim whether it is satisfied or refuted. It thus induces the *contingent truth* of the construction of a term in P, pending satisfaction of its condition(s). In this way, the inference to the principle of bivalence, $P \vee \neg P$ remains valid, though not trivial as its definition is reduced to a possibility operator. A classical example of a modally modified expression is 'alleged assassin': it starts by defining an 'assassin' by laying down the conditions $c_1, \ldots, c_n$ for an element of such a set to be construed (what does it mean to be an assassin); then it modifies it by applying the operator 'alleged', which generates for at least one of the listed conditions c_i the modal version $possibly(c_i)$, so that the obtaining of the property 'being an assassin' remains open, depending on c_i's refutation or verification. A similar analysis of the mistrust relation can be offered.[17] Informally, mistrust can be understood as the epistemic operation that, considering a certain content as unintentionally false information, induces the contingent falsity of that content, pending refutation. This should now be applied to our analyses of complete and incomplete information.

Definition 10 (Mistrustful Complete Transmission). *Assume a first-order relation of complete information transmission* $\langle \texttt{Metadata}, \mathcal{G} \rangle$ *between a source* S *and a receiver* R*. The complete transmission so defined is characterized by the second order property of mistrust if* R *thinks that* S *transmits unintentionally false information and so infers that*

1. $\texttt{Metadata}$ *is correct for some* $\mathcal{G}'$*, or*
2. $\mathcal{G}$ *is valid with respect to some* $\texttt{Metadata}'$*, or*
3. *there is a valid pair* $\langle \texttt{Metadata}', \mathcal{G}' \rangle$

The operation induced by this definition simply induces the identification of a different pair $\langle \texttt{Metadata}, \mathcal{G} \rangle$, which the Sender might have intended to transmit. The informal meaning of such variation is that the Receiver assumes false information is being sent unintentionally on the Sender's side. The Receiver is 'prepared' to act accordingly by considering alternative elements in the transmitted pair. There are two possible outcomes for the application of this operation:

- the new pair $\langle \texttt{Metadata}', \mathcal{G}' \rangle$ is formally the result of a subset operation on the original $\langle \texttt{Metadata}, \mathcal{G} \rangle$, i.e. the difference might be an issue of specification, for example in the case of the following pseudo-coded ordering functions on values (m, n), whose specification is contained in $\mathcal{G}$ in the form of the request to build a list of the values the function is applied to:

  $\texttt{procedure} := do\ \ Order(m, n); \mathcal{G} := List(mn)$
  $\texttt{procedure}' := do\ \ Order(n, m);\ \ \mathcal{G}' := List(mn)$

  the ordering $\texttt{procedure}'$ subsects on $\texttt{procedure}$, in this case just by considering the inverted ordering;

[17] See Jespersen, Primiero (2013), also for a brief overview of the literature in formal semantics of modal modifiers.

- the new pair $\langle \texttt{Metadata}', \mathcal{G}' \rangle$ is formally the result of a negation operation on the original $\langle \texttt{Metadata}, \mathcal{G} \rangle$, i.e. $\texttt{Metadata}'$ is equivalent to $\neg\texttt{Metadata}$, and $\mathcal{G}'$ to $\neg\mathcal{G}$, for example in the case

  $\texttt{procedure} := do\ \ Add(m, n);\ \ \mathcal{G} := SUM(mn)$
  $\texttt{procedure}' := do\ \ Sub(m, n);\ \ \mathcal{G}' := DIFF(mn)$

  where the ordering $\texttt{procedure}'$ generates a complement operation of $\texttt{procedure}$.[18]

In the following, we shall use $\circ_S$ and $\circ_R$ to refer to the epistemic states of the sender and the receiver respectively. These can be translated for example in terms of epistemic or doxastic operators in any modal or first-order logic of choice, or dependent typed language. The mistrust and distrust operators are used as additional assumptions in introduction and elimination rules for the negation operator over the content or over the epistemic operator. This gives our language a procedural interpretation, typical for example of proof-theoretical semantics. We shall abbreviate a mistrust property by R over a given information transmission as m_R. The mistrust operator mimics the behavior of the modal modifier in the interpretation offered in (Jespersen, Primiero, 2013, sec.3). The main inferential step induced in the Receiver's state is then formalized as follows:

$$\frac{\circ_S \langle \texttt{Metadata}, \mathcal{G} \rangle \qquad m_R(\circ_S \langle \texttt{Metadata}, \mathcal{G} \rangle)}{\neg \circ_R (\langle \texttt{Metadata}, \mathcal{G} \rangle)}\ \text{Mistrust}$$

The function m_R behaves like a modal modifier, whose meaning is given by an inferential step to a negated state about the $\langle \texttt{Metadata}, \mathcal{G} \rangle$ pair; this in turn means that the Receiver state $\neg \circ_R (\langle \texttt{Metadata}, \mathcal{G} \rangle)$ accounts for a contingent validity of alternative possible elements of the pair, according to one of the following steps:

$$\frac{\neg \circ_R \langle \texttt{Metadata}, \mathcal{G} \rangle}{\circ_R \langle \texttt{Metadata}', \mathcal{G} \rangle} \qquad \frac{\neg \circ_R \langle \texttt{Metadata}, \mathcal{G} \rangle}{\circ_R \langle \texttt{Metadata}, \mathcal{G}' \rangle} \qquad \frac{\neg \circ_R \langle \texttt{Metadata}, \mathcal{G} \rangle}{\circ_R \langle \texttt{Metadata}', \mathcal{G}' \rangle}$$

The logical negation at work in these cases is indeed modal but not privative, in that it does not distribute directly over the pair, it rather applies to the epistemic state of the receiver, which in turn generates possible alternatives. Which of the three cases above is effectively induced from the m operator is the result of an assessment that might be quantitatively or contextually resolved by the Receiver. In this sense, the present analysis accounts for first-time only transmissions and it does not define any mistrust propagation procedure; a similar remark will hold for the definition of distrust.

Let us reconsider our examples.

Bob: *'I checked the timetable, the train to Brussels leaves at 5pm'*.

[18] The characteristic behaviour of a modal modification operation is precisely that of oscillating between a subsective relation and a privative one: an 'alleged assassin' is an assassin (hence subsecting on the set of assassins by inducing one that is also suspected to be one) or is not (hence inducing the privative case).

How does Alice reason by implementing one of the rules for modal modification? Assuming, for example, that she believes Bob indeed checked the timetable, but this was just before an update was due, her best course of action is to consider the possibility that the train might not leave at 5pm.

> Alice: *'Bob checked before the update. The train might leave at some other time'.*

In the Server-Client example:

```
ALICE: Request:BIRTHDATE;PWD;PIN(SOURCE)
BOB: Enter:1103194_;rvcs132RT43;324564-676544(source:343434)
```

Assuming the Server can recognize a missing cypher in the first entry and a missing symbol in the second entry for an otherwise structurally correct message, its best course of action would be to assume the request is authentic (i.e. it is not an attack and does not require a plausible deniability reaction) and offer a second try.

```
ALICE: Modify: Incorrect entry BIRTHDATE;PWD. Retry
```

We now proceed with the appropriate counterpart for incomplete information.

Definition 11 (Mistrustful Incomplete Transmission). *Assume a first order relation of incomplete information transmission between a source S and a receiver R. The incomplete transmission so defined is characterized by a second order property of mistrust if R, when informed by S that*

1. *either $\langle \texttt{Metadata}, \texttt{empty} \rangle$, i.e. correct metadata but no goal is provided;*
2. *or $\langle \texttt{empty}, \mathcal{G} \rangle$, i.e. a goal is valid but no metadata is provided;*

thinks that S transmits unintentionally incomplete information, as S might hold either a valid $\mathcal{G}$ or a correct `Metadata`*, infers that*

1. *there might be a corresponding valid goal $\mathcal{G}$ for the transmitted metadata; or*
2. *there might be correct* `Metadata` *for the transmitted goal.*

While mistrust on a complete transmission induces content change request (modify), the meaning of a mistrust state in view of incomplete information simply amounts to content completion (request). The consideration that the Sender only unintentionally transmits incomplete information leads the Receiver to establish either the possible validity of some goal or the correctness for some procedure. So the initial step is the Receiver assessing the incomplete transmission to be unintentional:

$$\frac{\circ_S \langle \texttt{Metadata}, \texttt{empty} \rangle \qquad m_R(\circ_S \langle \texttt{Metadata}, \texttt{empty} \rangle)}{\neg \circ_R (\langle \texttt{Metadata}, \texttt{empty} \rangle)}$$

$$\frac{\circ_S\langle \texttt{empty}, \mathcal{G}\rangle \qquad m_R(\circ_S\langle \texttt{empty}, \mathcal{G}\rangle)}{\neg \circ_R (\langle \texttt{empty}, \mathcal{G}\rangle)}$$

In turn, the Receiver's reaction dictated by mistrust can be mimicked by the following inferential steps:

$$\frac{\neg \circ_R (\langle \texttt{Metadata}, \texttt{empty}\rangle)}{\circ_R\langle \texttt{Metadata}, \exists\mathcal{G}\rangle} \qquad \frac{\neg \circ_R (\langle \texttt{empty}, \mathcal{G}\rangle)}{\circ_R\langle \exists\texttt{Metadata}, \mathcal{G}\rangle}$$

Let us see how this applies to our examples.

Bob: *'The train to Brussels leaves at 5pm'.*

Here Bob is giving again Alice some goal information, neglecting the procedural aspect, the 'how' he knows. Here Alice can just assume that the information might be unintentionally false.

Alice: *'I do not know whether Bob has checked. I should check, then I know when the train leaves.'.*

In the Server-Client example:

```
ALICE: Request:BIRTHDATE;PWD;PIN(SOURCE)
BOB: Enter:11031946;rvcs?132RT43;empty(source:empty)
```

Assuming the Server recognizes the missing source on the random generated code, in an otherwise structurally correct message, its best course of action would be to assume the request is authentic (i.e. it is not an attack and does not require a plausible deniability reaction) and offer to complete the data.

```
ALICE: Request:Source empty. Complete.
```

3.2 Distrust

In the present section we offer a focused analysis of either complete or incomplete transmissions over disinformative channels. The connection between disinformation and distrust is formulated as follows:

Definition 12 (Distrustful Transmission). *A disinformative channel is an information transmission characterized by a second order property of distrust.*

How to describe the logical behaviour of the Receiver involved in a distrustful transmission? Our approach consists in analysing the second-order property of distrust as an operation of *privative modification* on the content of the transmission. Privative modification for procedural semantics can be defined as a specific kind of subsective operation: given a well-defined set P, it produces the set of functions from elements $p \in P$ to elements of the complement set $\neg P$. By looking at such functions, one considers subsective predications over the set P that induce the complement set. A classical example of a privatively modified case is the expression ‘fake banknote’: it starts from the set of elements that share the property of ‘being a banknote’; then it modifies it by applying the operator ‘fake’, which generates the set of non-banknotes (without actually including everything else, like horses and pens).[19] The logical behaviour of the Receiver of a distrustful transmission can be similarly formulated. Informally, distrust can be seen as the epistemic operation that, considering a certain content A as intentionally false information, induces the complement content $\neg A$, without this inducing B's and C's.

Definition 13 (Distrustful Complete Transmission). *Assume a first order relation of complete information transmission $\langle \texttt{Metadata}, \mathcal{G}\rangle$ between a source S and a receiver R. The complete transmission so defined is characterized by a second-order property of distrust if R, when informed by S that $\langle \texttt{Metadata}, \mathcal{G}\rangle$, thinks that S transmits intentionally false information and infers that*

1. $\texttt{Metadata}$ *is correct for* $\neg\mathcal{G}$*, or*
2. $\mathcal{G}$ *is valid with respect to* $\neg\texttt{Metadata}$*, or*
3. *there is a valid pair* $\langle \neg\texttt{Metadata}, \neg\mathcal{G}\rangle$

We shall also abbreviate a distrust property by R as d_R. The distrust operator mimics the behavior of the privative modifier in the interpretation offered in (Primiero, Jespersen, 2010, sec.3).The main inferential step induced in the Receiver's state is then formalized as follows:

$$\frac{\circ_S\langle \texttt{Metadata}, \mathcal{G}\rangle \qquad d_R(\circ_S\langle \texttt{Metadata}, \mathcal{G}\rangle)}{\circ_R\neg\langle \texttt{Metadata}, \mathcal{G}\rangle}\text{ Distrust}$$

The function d_R behaves like a privative modifier, whose meaning is given by an inferential step to the complement of the set generated by the $\langle \texttt{Metadata}, \mathcal{G}\rangle$ pair. Hence, in view of such operation, the meaning of the Receiver state $\circ_R\neg\langle \texttt{Metadata}, \mathcal{G}\rangle$ is further explained by one of the following steps:

$$\frac{\circ_R\neg\langle \texttt{Metadata}, \mathcal{G}\rangle}{\circ_R\langle \neg\texttt{Metadata}, \mathcal{G}\rangle} \qquad \frac{\circ_R\neg\langle \texttt{Metadata}, \mathcal{G}\rangle}{\circ_R\langle \texttt{Metadata}, \neg\mathcal{G}\rangle} \qquad \frac{\circ_R\neg\langle \texttt{Metadata}, \mathcal{G}\rangle}{\circ_R\langle \neg\texttt{Metadata}, \neg\mathcal{G}\rangle}$$

The procedural explanation of distrust in view of complete information transmission reduces to a pair of rules: negation introduction on content and negation distribution over the content pair. Which of the three cases above is effectively induced from the d operator is a question of assessment that might be quantitatively resolved depending on the number of previous cases of distrust involving the given S and R: one can then devise a scale that maps the lower level of trust in the Sender to the more complex case of privative modification (by establishing e.g. that negating one element in the pair is less distrustful than negating both, and that negating the goal is more distrustful than negating metadata); or the assessment might be a matter of contextual or purely contentual evaluation.

Let us go back to our examples.

[19] See Primiero, Jespersen (2010), also for a brief overview of the literature in formal semantics for privative modifiers.

> Bob: *'I regularly go to St. Pancras, the best way is to take a cab'.*

How does Alice reason by implementing one of the rules for privative modification? Assuming for example that she does indeed know that Bob is acquainted with travelling to St. Pancras, her best course of action is to deny the validity of his goal statement.

> Alice: *'Bob regularly goes to St. Pancras, he knows the best way is not the cab'.*

In the Server-Client example:

```
ALICE: Request:BIRTHDATE;PWD;PIN(SOURCE)
BOB: Enter:13061955;rtts?672TR21;434367-878799(source:898989)
```

Assuming the Server recognizes the mismatch between the data on `BIRTHDATE;PWD` and `SOURCE`, it might assume it is a random number generator attempting an attack and so require a plausible deniability reaction:

```
ALICE: Access denied. Further attempts denied.
```

Let us now consider the notion of distrust in view of intentionally incomplete transmissions.

Definition 14 (Distrustful Incomplete Transmission). *Assume a first order relation of incomplete information transmission between a source S and a receiver R. The incomplete transmission so defined is characterized by a second order property of distrust if R, when informed by S that*

1. *either* $\langle \texttt{Metadata}, \texttt{empty} \rangle$*, i.e. metadata is correct but no goal is provided;*
2. *or* $\langle \texttt{empty}, \mathcal{G} \rangle$*, i.e. goal is valid but no metadata is provided;*

thinks that S transmits intentionally incomplete information, as S does not hold either a valid $\mathcal{G}$ or a correct `Metadata` *and infers that*

1. `Metadata` *should not be considered correct; or*
2. *goal $\mathcal{G}$ should not be considered valid;*

The meaning of a distrust state in view of incomplete information simply amounts to disregarding the transmission considered. The informal idea is that the Receiver, upon reception of an incomplete message, assumes that the Sender is not even rightfully transmitting what he knows, maybe only making up his mind (and not even being able to do so completely) and in the worst case scenario performing an attempt to attack without appropriate privileges. Though the semantics of this case is slightly more complex to analyse, we can provide an explanation which actually reduces to the previous format of distrust for complete information transmission. The possible inference steps need to be formulated in view of an appropriate understanding of the pairs $\langle \texttt{Metadata}, \emptyset \rangle$ and $\langle \emptyset, \mathcal{G} \rangle$. We start with the negation introduction operation as defined above for complete transmissions:

$$\frac{\circ_S\langle \mathtt{Metadata}, \mathtt{empty}\rangle \qquad d_R(\circ_S\langle \mathtt{Metadata}, \mathtt{empty}\rangle)}{\circ_R\neg(\langle \mathtt{Metadata}, \mathtt{empty}\rangle))}$$

$$\frac{\circ_S\langle \mathtt{empty}, \mathcal{G}\rangle \qquad d_R(\circ_S\langle \mathtt{empty}, \mathcal{G}\rangle}{\circ_R\neg(\langle \mathtt{empty}, \mathcal{G}\rangle)}$$

Incompleteness of the transmission by missing metadata or a missing goal is to be ascribed to a voluntary act of the Sender. Then, according to our distrust operator, in the first case the Receiver refuses to assert validity for any goal and, in the second case, refuses to assert correctness for any metadata provided. In turn, the reaction dictated by distrust can be mimicked by the following inferential steps:

$$\frac{\circ_R\neg(\langle \mathtt{Metadata}, \mathtt{empty}\rangle)) \qquad \circ_R(\langle \neg\mathtt{Metadata}, \forall\mathcal{G}(\neg\mathcal{G})\rangle)}{\circ_R\langle \neg\mathtt{Metadata}, \neg\mathcal{G}\rangle}$$

$$\frac{\circ_R\neg(\langle \mathtt{empty}, \mathcal{G}\rangle) \qquad \circ_R(\langle \forall\mathtt{Metadata}(\neg\mathtt{Metadata}), \neg\mathcal{G}\rangle)}{\circ_R\langle \neg\mathtt{Metadata}, \neg\mathcal{G}\rangle}$$

Notice that in this case we do use negation introduction but not full distribution: we simply consider the empty element as meaning that no element is available and let distribute negation only over the element which actually occurs in the pair. This reduces to the last case of distrust for complete information, i.e. full information disregard.

Back again to our examples.

> Bob: *'The best way to St. Pancras is to take a cab'*.

Here Bob is giving Alice some goal information, neglecting the procedural aspect, the 'how' he knows. What is Alice's best course of action when assessing that Bob is sending intentionally false incomplete information?

> Alice: *'He gives no reason, I should trust none. The best way to the station is not the cab'*.

In the Server-Client example:

```
ALICE: Request:BIRTHDATE;PWD;PIN(SOURCE)
BOB: Enter:fffgggrrttt;323232rere;434367-878799(source:empty)
```

Assuming the Server recognizes the fully unstructured *and* incomplete message, it might assume it is a random number generator attempting an attack and so reject the given information and deny access:

```
ALICE: DATE:invalid; PWD:invalid: KEY:invalid; SOURCE:empty.
Access denied. Further attempts denied.
```

3.3 Mixed conditions

A variant case is when a composed message is assessed to be partly intentionally false, and partly unintentionally so. An example would be the following complete information transmission by S to R:

> $\texttt{procedure}_1$: I checked the timetable;
> $\mathcal{G}_1$: the train to London leaves at 5pm.
> $\texttt{procedure}_2$: I regularly go to the station;
> $\mathcal{G}_2$: the best way is to take a cab.

Assume that R assesses that both contents are false, but the first is unintentionally so, because R believes S does not know the timetable was updated few minutes ago; while the second is intentionally so, as it is known to R that S knows the metro is faster. The appropriate inference is of the following form

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\circ_S \langle \texttt{procedure}_1, \mathcal{G}_1 \wedge \texttt{procedure}_2, \mathcal{G}_2 \rangle}{\circ_S \langle \texttt{procedure}_1, \mathcal{G}_1 \rangle} \quad m_R(\circ_S \langle \texttt{procedure}_1, \mathcal{G}_1 \rangle)}{\neg \circ_R (\langle \texttt{procedure}_1, \mathcal{G}_1 \rangle)}}{\circ_R \langle \texttt{procedure}_1, \mathcal{G'}_1 \rangle} \quad \circ_S \langle \texttt{procedure}_1, \mathcal{G}_1 \wedge \texttt{procedure}_2, \mathcal{G}_2 \rangle}{\circ_S \langle \texttt{procedure}_2, \mathcal{G}_2 \rangle} \quad d_R(\circ_S \langle \texttt{procedure}_2, \mathcal{G}_2 \rangle)}{\circ_R \neg(\langle \texttt{procedure}_2, \mathcal{G}_2 \rangle)}}{\circ_R \langle \texttt{procedure}_2, \neg\mathcal{G}_2 \rangle}}{\circ_R \langle \texttt{procedure}_1, \mathcal{G'}_1 \wedge \texttt{procedure}_2, \neg\mathcal{G}_2 \rangle}$$

In this example, the mistrust and distrust operator induce respectively modification and negation over the goal only. Similar constructions can be offered for the cases of incomplete information transmissions.

4 Some properties of untrustworthy transmissions

As our analysis of distrust and mistrust builds on the model of trust as second order property characterizing the first order property of information transmission, in the following we will consider properties of untrustworthy transmissions by comparison with properties of trusted communications. Such properties rely on the formal analysis by modal frames given in Primiero, Taddeo (2012).

4.1 Reflexive untrustworthiness

In Primiero, Taddeo (2012), trusted communications are obtained by combining in one language a weak truth predicate with a strong truth predicate: by the latter, contents directly verified by an agent are claimed true; by the former, contents for which a provability condition is not available for the agent are declared non-refuted. A starting rule is defined that corresponds to reflexivity of the trust relation, see Primiero, Taddeo (2012, Lemma 2): for every proposition A that is non-falsified and for which the weak truth predicate holds, there is an agent who assumes A by trust; this agent can, in particular, be the same whose state accommodates an assumption on A. This can be further clarified out of the formalism as saying that agents can trust themselves on contents for which they have no refutation available. Can a similar reflexivity property be defined for untrustworthy contents? Or in other words: given a content for which an agent can assume A, can he proceed by untrusting A (thus, inducing either $\neg \circ A$ for mistrust or $\circ\neg A$ for distrust)? Satisfaction of this property would mean that for any agent $\circ_A$, the operations

$$\frac{\circ_A\langle\texttt{Metadata},\mathcal{G}\rangle \qquad m_A(\circ_A\langle\texttt{Metadata},\mathcal{G}\rangle)}{\neg \circ_A (\langle\texttt{Metadata},\mathcal{G}\rangle)}\text{ Mistrust}$$

$$\frac{\circ_A\langle\texttt{Metadata},\mathcal{G}\rangle \qquad d_A(\circ_A\langle\texttt{Metadata},\mathcal{G}\rangle)}{\circ_A\neg\langle\texttt{Metadata},\mathcal{G}\rangle}\text{ Distrust}$$

are admissible. Suppose that given $\circ_A\langle\texttt{Metadata},\mathcal{G}\rangle$ then $\neg \circ_A (\langle\texttt{Metadata},\mathcal{G}\rangle)$; according to the interpretation given in Section 3.1, this means that in some extension of the epistemic state A considers either $\texttt{Metadata}'$ or $\mathcal{G}'$ or both true. In case of $\circ_A\neg\langle\texttt{Metadata},\mathcal{G}\rangle$, it means A considers $\neg\texttt{Metadata}$ or $\neg\mathcal{G}$ or both true. This semantics is thus clearly non-monotonic in view of the fact that the content A is only weakly accepted (i.e. in the form of a possibility operator); in particular, it should be possible to declare as false or as contingently false a content that has been so far accepted as contingently true. It seems thus reasonable that untrustworthy transmissions can be reflexive if the underlying semantics accommodates a non-monotonic transition between epistemic state.[20]

4.2 Limited transitive untrustworthiness

A second provable property of trust relations in Primiero, Taddeo (2012, Lemma 3) is backward ordered transitivity: if a content A can be held true by agent k trusting agent j, and j holds A true by trusting i, then k trusts i on A.[21] Untrustworthy transmissions do not seem to relate so easily Senders and Receivers.

One aspect of untrustworthy *complete* transmissions is that they are typically non-transitively iterated when the d_R or m_R operator is applied uniformly at all passages, i.e. it is entirely explicit what the untrustworthiness is all about. To show why this property holds for *distrustful complete* transmissions, the following reasoning should suffice:[22]

[20] Notice that this means that a standard upwards monotonic intuitionistic semantics would, for example, not be feasible for this purpose. This is the reason why the model for trust introduced in Primiero, Taddeo (2012), though intuitively based on a verificationist semantics, extends the standard setting with a weak truth predicate based on missing refutations. A full calculus for the latter is given in Primiero (2012).

[21] Notice how this is the case for our treatment of trust as a second-order property characterizing first-order relations of information transmission on specified contents. Trust defined as first-order relation requires a restriction on transitivity, what is called in the literature *promiscuous trust*.

[22] In what follows we abbreviate `Metadata` with simply `M` for simplicity of reading.

> Assume $\circ_i(\langle \mathtt{M}, \mathcal{G}\rangle)$ and $d_j(\circ_i(\langle \mathtt{M}, \mathcal{G}\rangle)$ about $\mathtt{M}$; then $\circ_j(\langle \neg\mathtt{M}, \mathcal{G}\rangle)$. If $d_k(\circ_j(\langle \neg\mathtt{M}, \mathcal{G}\rangle))$ about $\neg\mathtt{M}$, then $\circ_k(\langle \neg\neg\mathtt{M}, \mathcal{G}\rangle)$; hence, k could not distrust i: $d_k(\circ_i(\langle \mathtt{M}, \mathcal{G}\rangle))$ would not hold about $\mathtt{M}$. Hence transitivity fails. Similarly, if d_R would apply to $\mathcal{G}$ or to both elements of the transmitted pair.

On the other hand, the same reasoning does not hold in general for *mistrustful complete* transmissions:

> Assume $\circ_i(\langle \mathtt{M}, \mathcal{G}\rangle)$ and $m_j(\circ_i(\langle \mathtt{M}, \mathcal{G}\rangle))$ about $\mathtt{M}$, then $\circ_k(\langle \mathtt{M}', \mathcal{G}\rangle)$; and if $m_k(\circ_j(\langle \mathtt{M}', \mathcal{G}\rangle))$ about $\mathtt{M}'$, then $\circ_j(\langle \mathtt{M}'', \mathcal{G}\rangle)$; hence, $m_k(\circ_i(\langle \mathtt{M}, \mathcal{G}\rangle))$ might still hold, if $\mathtt{M}''$ would not reduce to $\mathtt{M}$. Similarly, if m_R would apply to $\mathcal{G}$ or to both elements of the transmitted pair.

Finally, transitivity is not in general applicable to untrustworthy *incomplete* transmissions. Let us start with *distrustful incomplete* transmissions.

> If $\circ_i(\langle \mathtt{M}, \mathtt{empty}\rangle)$ and $d_j(\circ_i(\langle \mathtt{M}, \mathtt{empty}\rangle))$ about $\mathtt{M}$, then $\circ_j(\langle \neg\mathtt{M}, \neg\mathcal{G}\rangle)$; and if $d_k(\circ_j(\langle \neg\mathtt{M}, \neg\mathcal{G}\rangle))$, then we are treating a case of complete transmission. Assume that d_k is now about $\neg\mathtt{M}$, then $\circ_k(\langle \neg\neg\mathtt{M}, \neg\mathcal{G}\rangle)$; hence, $d_k(\circ_i(\langle \mathtt{M}, \mathtt{empty}\rangle))$ does not hold about $\mathtt{M}$, but would in view of $\circ_k(\neg\mathcal{G})$. This means that for an incomplete transmission, distrust may iterate monotonically among senders and receivers, depending on the qualification of the second distrustful transmission. Similarly, if d_R would apply to $\mathcal{G}$, or to both elements of the transmitted pair or if the transmission would be of the form $\langle \mathtt{empty}, \mathcal{G}\rangle$.

For *mistrustful incomplete* transmissions, a similar case can be presented:

> Assume $\circ_i(\langle \mathtt{M}, \mathtt{empty}\rangle)$ and $m_j(\circ_i(\langle \mathtt{M}, \mathtt{empty}\rangle))$ about the empty goal, then $\circ_j(\langle \mathtt{M}, \mathcal{G}\rangle)$; and if $m_k(\circ_j(\langle \mathtt{M}, \mathcal{G}\rangle))$ we are again with a complete transmission. Assume that m_k is about $\mathtt{M}$, then $\circ_k(\langle \mathtt{M}', \mathcal{G}\rangle)$; hence, $m_k(\circ_i(\langle \mathtt{M}, \mathcal{G}\rangle))$ still holds if $\mathtt{M}'$ does not reduce to $\mathtt{M}$ (and it does not hold if the reduction does). Similarly, if m_R would apply to $\mathcal{G}$ or to both elements of the transmitted pair. This means that an incomplete transmission may or may not iterate monotonically (depending on reducibility of the completing elements selected).

4.3 Symmetric untrustworthiness

Symmetry fails for trusted communications (Primiero, Taddeo (2012, Lemma 4)): if A is held true by agent j trusting agent i, it cannot be the case that A holds true for agent i by trusting agent j. Similarly, untrustworthy relations are not symmetric. [23] For the case of distrustful complete transmissions:

[23] In the following, we will be using two obvious simplifications: that no double-games are in place, and that the untrustworthiness assessments are public.

> Assume $\circ_i(\langle \texttt{M}, \mathcal{G} \rangle)$ and $d_j(\circ_i(\langle \texttt{M}, \mathcal{G} \rangle))$ about M, then $\circ_j(\langle \neg\texttt{M}, \mathcal{G} \rangle)$; i can still distrust j as he holds M in his initial message.[24]

Similarly for mistrust:

> Assume $\circ_i(\langle \texttt{M}, \mathcal{G} \rangle)$ and $m_j(\circ_i(\langle \texttt{M}, \mathcal{G} \rangle))$ about M, then $\circ_j(\langle \texttt{M}', \mathcal{G} \rangle)$; then i can mistrust j in any case and only distrust j in case $\texttt{M}'$ reduces to $\neg\texttt{M}$.

So it seems that in the case of complete information transmission, distrust and mistrust can be symmetric. Let us consider the case of untrustworthy incomplete transmission.

> Assume $\circ_i(\langle \texttt{M}, \texttt{empty} \rangle)$ and $d_j(\circ_i(\langle \texttt{M}, \texttt{empty} \rangle))$, then $\circ_j(\langle \neg\texttt{M}, \neg\mathcal{G} \rangle)$; i can still distrust j in view of M, as he holds the opposite; in view of $\neg\mathcal{G}$, if i holds some $\mathcal{G}$ he can distrust j; if i holds truly empty, he cannot distrust j on $\neg\mathcal{G}$ as for that he would need to hold $\mathcal{G}$, which he does not. To the purpose of reflexivity, is distrust on M sufficient.

For incomplete mistrustful transmission:

> Assume $\circ_i(\langle \texttt{M}, \texttt{empty} \rangle)$ and $m_j(\circ_i(\langle \texttt{M}, \texttt{empty} \rangle))$, then $\circ_j(\langle \texttt{M}, \mathcal{G} \rangle)$; M is irrelevant in this case; if i holds truly empty, then he will mistrust j on $\mathcal{G}$ and if i holds some $\mathcal{G}'$ there are three cases: either $\mathcal{G}'$ reduces to $\mathcal{G}$, then no untrustworthiness is at stake; or it reduces to some $\mathcal{G}''$ different than $\mathcal{G}$, then he will mistrust j; or it reduces to $\neg\mathcal{G}$ and then i distrusts j. This means that an incomplete transmission may or may not be symmetric (depending on reducibility and on elements selected).

4.4 Expert untrustworthiness

Our model shows that distrust operations only apply on the basis of a sufficient degree of expertise on the Receiver's part. This is possible only in view of the characterization of trust as second-order property, hence on the basis of the underlying relation of communication transmission. The procedural explanation of a distrust operation as a privative operator that negates partly or completely the content of the transmission, requires at least sufficient competence on the Receiver side about both metadata and goal. In a mistrust operation, however, in which the procedural explanation in terms of a modal operator induces contingency on metadata and goal, the Receiver's expertise is not required: modal modification admits anything from possibility to judgement suspension. Thus, according to this model, a layperson with no specific competence is but able to mistrust an expert; on the other hand, only an expert (to some sufficient degree) can distrust the content of a transmission. Different types of expertise have been

[24] A more powerful framework allowing us to express the reasons for untrust assessments, would make it possible to formulate the reactions of i to j's response to the initial message. In this way, i's further assessment could be dictated on the basis of the correctness of j's one. We leave this to further research.

identified by, among others, Collins, Evans (2002, p.254). Applying their threefold distinction between 'no expertise', 'interactional expertise' and 'contributory expertise' to our account entails that a layperson who has 'no expertise' is unable to distrust an expert. Enough expertise to interact with scientists from the field and carry out analysis (interactional expert) and/or enough expertise to contribute to the actual field of science (contributory expertise) is required. Put briefly: a layperson is unable to distrust an expert unless he or she can at least be classified as an interactional expert. Although a clearly skeptical result, this approach does differ from radical skepticism as offered by Frances (2005) and Brewer (1998) as it shows in what instances a layperson is and is not able to justifiably adjudicate expert testimony.[25] A model in which (ir)rational reasons behind assigning trustworthiness to relations are taken into account, could possibly offer a different analysis. In particular, we can imagine a layperson qualifying transmissions as distrustful on the basis of previous experiences with an expert's trustworthiness.

In the application to (secure) systems design, this property means that the possibility to design a control system able to initialize different responses on the basis of an assessment of the intentionality of false information received from a client relies crucially on the *expertise* of the system with respect to the expected input. Of course, any such system would have to match the input of any given requesting client against the design criteria (e.g. the number of digits/symbols present in `PWD`, the format of `BIRTHDATE`, the structural correctness of the `PIN` string). This is not enough to define expertise as to discern between intentional attacks and unintentionally mistaken entries. In our example, we have mentioned some possible design criteria for defining this kind of expertise: for example, the requirement that the server Alice overdrives the *structurally correct* `BIRTHDATE` and `PWD` entries against a *mismatching* `SOURCE` entry, i.e. where the latter is not coherent with the expected one in view of the user associated with the previous two entries. In this case, the system is evaluating the latter condition as more relevant to the first two, and hence assessing a distrust action rather than a new attempt request. Combinations of such conditions might lead to a better and more efficient system design of secure systems in view of trust and untrust assessments.

5 Conclusions

We have presented an analysis of channels qualified by the transmission of either complete or incomplete intentionally and unintentionally false information. We have shown how such qualifications induce an appropriate understanding of the notions of distrust and mistrust respectively. We offered a treatment of distrustfully and mistrustfully qualified transmissions in view of a procedural approach

[25] Other, even more optimistic social epistemological attempts to deal with the problem of adjudicating between rival experts are due to Goldman (2001) and Haack (2004). However as suggested by Miller (forthcoming), they do not significantly enhance the layperson's epistemic arsenal.

that defines the former by a privative and the latter by a modal modification on contents. We have explored how basic properties of reflexivity, transitivity and symmetry behave for such channels. Applications are in expertise and secure systems design. From here, we intend to develop a formal treatment of distrust and mistrust operations for multi-agent and distributed systems, the study of propagation relations, their properties and their identification in view of error conditions and limited information availability. Moreover, we have considered how our model of untrustworthiness operates in expertise contexts. From here, we intend to offer a practical treatment of distrust and mistrust operations, in the sense of critically examining the conditions under which experts and laymen interact in real social contexts, a theoretical analysis of the criteria for better secure systems design and how these relations are and should be informed by (un)trustworthiness.

Acknowledgments

The first author wishes to thank the participants to the Fifth Workshop on the Philosophy of Information held at the University of Hertfordshire where this paper was first presented, for useful comments and discussions. The second author wishes to thank Jan de Winter for comments on previous versions of the paper. Both authors would like to thank the reviewers for their comments that helped improve the manuscript.

Bibliography

A. Abdul-Rahman, S. Hailes. A distributed trust model. In *New Security Paradigms Workshop*, Cumbria, UK, September 1997:4860.

R. Audi. The place of testimony in the fabric of justification and knowledge. *American Philosophical Quarterly*, 34:405–422, 1997.

T. Beth, M. Borcherding, and B. Klein. Valuation of trust in open networks. In *Proceedings of the Third European Symposium on Research in Computer Security*, ESORICS '94, pages 3–18, London, UK, UK, 1994. Springer-Verlag.

C. Borgs, J. Chayes, A.T. Kalai, A. Malekian, and M. Tennenholtz. A novel approach to propagating distrust. In *Proceedings of the 6th international conference on Internet and network economics*, WINE'10, pages 87–105, Berlin, Heidelberg, 2010. Springer-Verlag.

S. Brewer. Scientific expert testimony and intellectual due process. *The Yale Law Journal*, 107(6):1535–1681, 1998.

C. Castelfranchi. *Trust Mediation in Knowledge Management and Sharing*, volume 2995 of *Lectures Notes in Computer Science*, pages 304–318. Springer Verlag, 2004.

R. Chen, W. Yeager. Poblano: A distributed trust model for peer-to-peer networks. Technical report, Sun Microsystems, Santa Clara, CA, USA, February 2003.

B. Christianson and W. Harbison. Why isn't trust transitive? In *Security Protocols 4*, volume 1189 of *Lecture Notes in Computer Science*, pages 171–176. Springer, 1997.

H.M. Collins and R. Evans. The third wave of science studies: studies of expertise and experience. *Social Studies of Science*, 32(2):235–296, 2002.

R. Dalton. Peers under pressure. *Nature*, 413:102–104, 2001.

M. Dastani, A. Herzig, J. Hulstijn, and L. Van Der Torre. Inferring trust. In *Proceedings of the Fifth Workshop on Computational Logic in Multi-agent Systems (CLIMA V)*, volume 3487 of *Lecture Notes in Artificial Intelligence*, pages 144–160. Springer Verlag, 2004.

R. Demolombe. *Reasoning about Trust: a formal logic framework*, volume 2995 of *Lectures Notes in Computer Science*, pages 291–303. Springer Verlag, 2004.

P. Faulkner. The practical rationality of trust. *Synthese*, pages 1–15, 2012.

L. Floridi. *The Philosophy of Information.* Oxford University Press, Oxford, 2011.

B. Frances. *Skepticism comes alive.* Oxford Univ Press, 2005.

G. Gans, M. Jarke, S. Kethers, G. Lakemeyer. Modeling the impact of trust and distrust in agent networks. In *Proceedings of the Third International Bi-Conference Workshop on Agent-oriented Information Systems.* Montreal, Canada, May 2001.

A. Goldman. Experts: which ones should you trust? *Philosophy and Phenomenological Research*, 63(1):85–111, 2001.

R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 403–412, New York, NY, USA, 2004. ACM.

S. Haack. Truth and justice, inquiry and advocacy, science and law. *Ratio Iuris: An International Journal of Jurisprudence and Philosophy of Law*, 17(1):15–26, 2004.

J. Hardwig. The role of trust in knowledge. *The Journal of Philosophy*, 88:693–708, 1991.

A. Herzig, E. Lorini, J. F. Hübner, and L. Vercouter. A logic of trust and reputation. *Logic Journal of the IGPL*, 18(1):214–244, 2010.

B. Jespersen and G. Primiero. Alleged assassins: Realist and constructivist semantics for modal modification. In G. Bezhanishvili et al., editor, *TbiLLC 2011*, volume 7758 of *Lecture Notes in Computer Science*. Springer Verlag, 2013.

S.D. Kamvar, H. Garcia-Molina and M.T. Schlosser. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 640–651, New York, NY, USA, 2003. ACM.

L. Kosolosky. 'Peer review is melting our glaciers': Exploring how and why the intergovernmental panel on climate change (ipcc) went astray. *Journal for General Philosophy of Science*, Special Issue on Climate Change, to appear.

S. Kramer, R. Goré, and E. Okamoto. Computer-aided decision-making with trust relations and trust domains (cryptographic applications). *Journal of Logic and Computation*, 2012.

S. Leonelli. Data interpretation in the digital age. *Perspectives on Science*, forthcoming.

J.D Lewis and A.J. Weigert. Trust as a social reality. *Social Forces*, 63:967–985, 1985.

D. H. McKnight and N. L. Chervany. Trust and distrust definitions: One bite at a time. In Rino Falcone, Munindar P. Singh, and Yao-Hua Tan, editors, *Trust in Cyber-societies, Integrating the Human and Artificial Perspectives* , volume 2246 of *Lecture Notes in Computer Science*, pages 27–54. Springer, 2000.

B. Miller. Scientific consensus and expert testimony in courts: lessons from the bendectin litigation. In A. Froeyman, L. Kosolosky, J. van Bouwel (eds.), *Foundations of Science*, Special Issue on 'Science vs. Society: Social Epistemology meets the Philosophy of the Humanities', forthcoming.

G. Primiero. A contextual type theory with judgemental modalities for reasoning from open assumptions *Logique & Analyse*, vol. 220, pp.579-600, 2012.

G. Primiero. A taxonomy of errors for information systems. *Minds & Machines*, 2013. DOI:10.1007/s11023-013-9307-5.

G. Primiero and B. Jespersen. Two Kinds of Procedural Semantics for Privative Modification. volume 6284 of *Lecture Notes in Artificial Intelligence*, pages 252–71, Berlin, Germany, 2010. Springer Verlag.

G. Primiero and M. Taddeo. A modal type theory for formalizing trusted communications. *Journal of Applied Logic*, 10:92–114, 2012.

J. B Rotter. Generalized expectancies for interpersonal trust. *American Psychologist*, 26:443–452, 1971.

S.P. Shapiro. The social control of impersonal trust. *American Journal of Sociology*, 93(3):623–658, November 1987.

T. Simpson. What is trust. *Pacific Philosophical Quarterly*, 93(4):55–569, 2012.

M. Taddeo. An information-based solution for the puzzle of testimony and trust. *Social Epistemology*, 24(4):285–299, 2010.

M. Taddeo. Modelling Trust in Artificial Agents, a First Step toward the Analysis of e-Trust. *Minds & Machines*, 20(2):243–257, 2010.

J. De Winter and L. Kosolosky. The epistemic integrity of scientific research. *Science and Engineering Ethics*, pages 1–18, 2012.

J. De Winter and L. Kosolosky. The epistemic integrity of NASA practices in the space shuttle program. *Accountability in Research*, 20(2):72–92, 2013.