Towards a Logical Analysis of Misleading and Trust Erosion

Haythem O. Ismail
Department of Engineering Mathematics
Cairo University
Department of Computer Science
German University in Cairo
Cairo, Egypt

Patrick Attia
Department of Computer Science
German University in Cairo
Cairo, Egypt

Abstract

Misleading is, regrettably, an integral part of the commonsense world. Though lying, deception, and similar malignant variants of misleading have been thoroughly investigated in ethics and social psychology, there is a rather slim related literature within the logicist AI tradition. In this paper, we present foundations for a logical theory of general misleading, with an eye on its effect on trust erosion. In particular, we define a bare-bones notion of misleading and identify four dimensions along which we distinguish eighty one variants of misleading. Given this analysis, we suggest that a logical theory of misleading for trust erosion should include an account of belief, desire, intention, and causality. A logical language $\mathcal{L}_M$ is sketched and used to represent the identified assortment of misleading scenarios.

1 Introduction

Lying, deception, and other forms of misleading are, admittedly, part and parcel of the commonsense world. Whether malignant, harmless, good-hearted, or outright altruistic, an instance of misleading does not measure up to the high standards set by a logic-based agent for the reliability of its sources of information. Such an agent can be misled by hostile, lying agents; by cooperative, yet mis-informed agents; or even by fallible perception due to faulty sensors or illusory environments (Ismail and Kasrin 2010).[1] Since commonsense reasoning is driven by observations made through communication or perception, trust in information sources is an important factor for directing belief revision should contradictions arise. Said trust is, at least partially, dependent on the history of misleading of information sources.

Our long-term goal is to develop a theory of logic-based agents which can reason about the erosion and recovery of trust in information sources, where these sources may be other agents or the reasoning agent's own perception processes. We believe that trust erosion in information sources is primarily affected by incidents of misleading (where misleading is construed in a very general sense). Our short-term goal, in this paper, is four-fold: (i) to identify a common core of all varieties of misleading and a limited number of dimensions along which we can distinguish them, (ii) to propose a ranking of varieties of misleading with respect to the extent to which they affect trust erosion, (iii) to pinpoint the necessary ingredients of an ontology for a logic of misleading, and (iv) to develop a logical language for reasoning about misleading scenarios. Anticipating the future coupling of misleading and trust, we are guided in achieving our four goals by how suitable our analysis and constructions are for a theory of trust erosion in information sources.

There is an abundant literature on trust analysis, with contributions from social and managerial psychology (Schweitzer, Hershey, and Bradlow 2006; Elangovan, Auer-Rizzi, and Szabo 2007; Haselhuhn, Schweitzer, and Wood 2010; Levine and Schweitzer 2015, for instance), economics (Cox 2004, for instance), social robotics (Wagner and Robinette 2015), and multi-agent systems and e-commerce (Schillo, Funk, and Rovatsos 2000; Sabater and Sierra 2005, for instance). Most formal theories of trust are probabilistic or game theoretic, but some logicist approaches exist (Demolombe 2009; Herzig et al. 2010; Amgoud and Demolombe 2014; Demolombe 2015; Drawel, Bentahar, and Shashuki 2017). None of the logical theories, however, establishes a link to misleading. Research on lying and deception (but not misleading in general) is also quite varied, drawing interest from psychology and human communication (Buller and Burgoon 1996), economics (Gneezy 2005; Ettinger and Jehiel 2010; Gneezy, Rockenbach, and Serra-Gracia 2013), social robotics (Wagner and Arkin 2011), and is an all-time favorite of philosophy (Mahon 2016).

Within the logicist framework, however, analysis of lying and deception is (to the best of our knowledge) limited to the work of Sakama and colleagues (Sakama, Caminada, and Herzig 2010; Sakama 2011a; 2011b; 2015). While we attempt to base our constructions on the foundations established by them, we do not limit ourselves to the lying and deception varieties of misleading and we keep our analysis motivated by issues of trust erosion.

2 A Bare-Bones Notion of Misleading

Though a lot of work has been done on the analysis of lying and deception, there is, much to our distress, almost no systematic analysis of misleading in general. To identify a bare-bones notion of misleading, we start with what is out

[1] Studies indicate that lying alone is quite pervasive, with an American telling an average of one to two lies every day (DePaulo et al. 1996). Most lies, however, are told by a small percentage of the population (Serota, Levine, and Boster 2010).

there: definitions of lying and deception. First, consider the following adaptation of "the traditional definition of lying" (Mahon 2016):

Cognitive agent S lies if and only if

(l1) S states proposition P to A.

(l2) S believes P to be false.

(l3) A is a cognitive agent.

(l4) S states P to A with the intention that A believes P to be true.

We can attempt to generalize this definition to one of misleading by considering each condition and either dropping or generalizing it. But, first, consider what it is that we are trying to define. For starters, we cannot just replace "lies" with "misleads", for we are not primarily interested in agent S and their actions, but in agent A—the one being *misled*—and what happens to them and how it affects their trust in S. It will also not do to define what it means for A to be misled. The reason is that "mislead" is an achievement verb (cf. (Vendler 1957)) and we do not want to imply that A is subjected to successful misleading; a *potentially successful* misleading of A is sufficient to shake their trust in S. We propose to replace the clause "Cognitive agent S lies" by "Event E is judged as misleading by cognitive agent A". There are a couple of things to note here. First, we take that which is misleading to be, not an agent nor a statement, but an event. For example, a perception event can be misleading though there is no *misleader* nor is there any form of linguistic communication. Second, you can judge an event to be misleading without being misled by it. For example, a student's untruthful claim to having spent the night working on their dissertation is a misleading event, though their major professor will never be misled by it if they had seen the student at a party the night before. Third, what matters is that A judges the event to be misleading, regardless of what anybody else thinks.

We now turn to conditions **(l1)**–**(l4)**. As already stated, misleading need not involve any form of linguistic communication as mandated by **(l1)**. However, we still need to confine ourselves to misleading events in which some information source S (which is not necessarily an agent) conveys some proposition P. Examples include having a perception with content P, reading a statement of P in a newspaper, and, of course, person S's stating P. For **(l2)**, we have already pointed out that S need not be an agent at all and may, thus, have no beliefs. But even in the prototypical case when S is a person stating P, S may believe P but use it to conversationally implicate another proposition which they do not believe (Adler 1997; Stokke 2016). Thus, a general misleading event involves two propositions: P and the contextually implicated Q. If S is a cognitive agent, misleading occurs if they do not believe Q. This is not necessary, however; misleading may still occur if S believes Q but Q is false. On the other hand, if S is not a cognitive agent, we contend that there cannot be any misleading unless Q is false. Finally, both **(l3)** and **(l4)** may simply be dropped: **(l3)** is presupposed by the left-hand side of our definition and **(l4)** does not make sense if S is not an agent.

Value	Meaning
0	S believes P
?	S believes neither P nor $\neg P$
1	S believes $\neg P$

Table 1: Values and their meanings for **BP**

Hence, we adopt the following bare-bones notion of misleading:

(M) Event E is judged to be misleading by cognitive agent A if and only if:

(m1) E is an event of information source S's (directly) conveying proposition P.

(m2) S's conveying of P together with a common ground C defeasibly imply Q.

(m3) Q is false or S does not believe Q.

There are a couple of points to note about **(m1)** and **(m2)**. It is out of the scope of this paper to provide a general theoretical account of what it means for an event E to be one of an information source S's (directly) conveying a proposition P. The simplest case is when E is the event of a person S's stating P. But other cases include sensor S's producing a signal interpreted as P by the sensing agent, or agent S's performing some action α and thereby conveying the proposition that "S has just performed α." We assume that particular agent theories include statements indicating for some relevant events that they are events of certain information sources conveying certain propositions.

By **(m2)**, we model implicature (Grice 1989) by defeasible implication given some common ground C. Following (Stalnaker 2002), we think of common ground as some proposition which A believes to be common belief (in the sense of (Fagin et al. 1995).) Again, we lay the responsibility of specifying C on particular logical theories that may choose to make use of our notion of misleading. In the example of the deceitful graduate student, the common ground includes the belief that speakers are honest. Thus, together with the student's claim of spending the night working on their dissertation, the common ground implies that they indeed did so. This is defeated, however, by the professor's witnessing the student partying all night.

3 The Many Scenarios of Misleading

We distinguish different types of misleading using four parameters: (i) whether S believes P (**BP**), (ii) whether S intends to deceive A (**ID**), (iii) whether S intends to harm A (**IH**), and (iv) whether being misled has a negative effect on A (**EQ**). Each of these parameters may assume one of three values: 0, ?, and 1. Tables 1 through 4 indicate the conditions represented by each assignment of a value to a parameter. We note the following:

1. **BP**, **ID**, and **IH** can take the values 0 or 1 only if S is a cognitive agent. A value of ? may indicate that S is not a cognitive agent in the first place.

Value	Meaning
0	S intends to not deceive A
?	S intends to neither deceive A nor to not deceive A
1	S intends to deceive A

Table 2: Values and their meanings for **ID**

Value	Meaning
0	S intends to not harm A
?	S intends to neither harm A nor to not harm A
1	S intends to harm A

Table 3: Values and their meanings for **IH**

2. That **BP** and **ID** are parameters distinguishing varieties of misleading is already a common practice in analyses of lying and deception (Mahon 2016; Sakama, Caminada, and Herzig 2010; Sakama 2011a; 2011b; 2015).
3. **IH** and **EQ** are motivated by our long-term goal of establishing a link between misleading and trust erosion. After conducting a series of interesting studies, Levine and Schweitzer (Levine and Schweitzer 2015) conclude that deception per se does not always harm trust, but selfishness and willingness to harm do. This motivates including something like **IH** as a dimension for classifying misleading scenarios. Moreover, studies show that people are, in general, less forgiving of lies which have more damaging effects on the victim (Gneezy 2005). (Gneezy, Rockenbach, and Serra-Gracia 2013) reports on experiments conducted to identify when people take the decision to lie. One of the findings of the experiments is that the victims' trust in the liers deteriorates more severely if, by following the lie, they lose their monetary payoff. These results suggest the appropriateness of **EQ**.

With our four three-valued parameters, we can distinguish eighty one different scenarios of misleading, **M0**–**M80**. Each scenario is characterized by eight conditions: **m1** through **m4** and one condition from each of the Tables 1 through 4. Symbolically, we can encode the misleading-variants by using the standard ternary encoding of the natural numbers 0–80 over the alphabet $\{0, ?, 1\}$. Table 5 tersely displays the association between the labels (**M**i) and the strings.

Value	Meaning
0	Believing Q has a positive effect on A
?	Believing Q has a neutral effect on A
1	Believing Q has a negative effect on A

Table 4: Values and their meanings for **EQ**

		BP EQ								
		00	**0?**	**01**	**?0**	**??**	**?1**	**10**	**1?**	**11**
IH ID	**00**	0	1	2	3	4	5	6	7	8
	0?	9	10	11	12	13	14	15	16	17
	01	18	19	20	21	22	23	24	25	26
	?0	27	28	29	30	31	32	33	34	35
	??	36	37	38	39	40	41	42	43	44
	?1	45	46	47	48	49	50	51	52	53
	10	54	55	56	57	58	59	60	61	62
	1?	63	64	65	66	67	68	69	70	71
	11	72	73	74	75	76	77	78	79	80

Table 5: The mapping between labels and string encodings of misleading scenarios. Each cell indicates the index i corresponding to label **M**i. The string encoding is constructed by appending the column label to the row label

We rank misleading scenarios along the natural order of the integers 0–80; the higher the number, the more erosive-to-trust the scenario is. Thus, *ceteris paribus*, S's believing P is always better than their having no clue about it, which is always better than their believing it to be false. This is, in fact, the common view in ethics. (For example, see (Saul 2012) who, interestingly, argues against this common view.) Likewise, a positive effect (on A) of a successful misleading is, *ceteris paribus*, always better than a neutral effect, which is always better than a negative effect; this is consistent with the findings of (Gneezy 2005; Gneezy, Rockenbach, and Serra-Gracia 2013). Similarly for the intentions to deceive and harm. Globally, **IH** has the strongest influence on trust erosion, followed by **ID**, followed by **BP**, and finally by **EQ**. That **EQ** comes last makes sense since the consequences of believing Q are generally not under the control of S. On the other hand, **IH** comes first signifying the damaging effect on trust that selfishness and willingness to harm have (Levine and Schweitzer 2015).

Now, it might be suspected that some of the scenarios **M0**–**M80** are not realistic. We have successfully constructed real-life examples of each scenario and we present some of the interesting/exotic ones below.

Example 1. In what follows, we present eleven examples of selected entries from Table 5. All examples are about our two protagonists: the misleading information source Steve (S) and the sharp, trusting agent Ashley (A).

M80 (1111). Steve tells his colleague Ashley that there is no meeting the next day (P) although he believes that there is indeed an important meeting. Steve does so with the intention of deceiving Ashley and of hurting her career at the company as a result of missing the meeting. Believing Steve and missing the meeting, Ashley gets a deduction and a notice.

M79 (111?). The same **M80** scenario above but the meeting gets cancelled and nothing, good or bad, happens to Ashley.

M78 (1110). Same as **M80** but, missing the meeting, Ashley gets the chance to work more on her assigned tasks, produces fantastic results, and ends up getting a raise.

M41 (???1). Steve tells Daphne and Ashley that there is a theory of computation quiz the next day. However, he has no idea whether there is a quiz or not, he just wants Daphne to be nervous and does not care about Ashley who just happens to be there. Believing Steve, Ashley panics, spends the night studying theory of computation, and forgets about the networks quiz which she, consequently, fails.

M39 (???0). Same as **M41** above but it turns out there is a pop quiz in theory of computation the next day. Ashley does great since she spent the night studying.

M26 (0111). Steve and Ashley apply for an internship. At the interview they are told that only one person will get the internship and will be notified by e-mail if they get accepted. Steve gets the e-mail, but refrains from saying so to Ashley when she asks to spare her feelings. Consequently, Ashley waits for the e-mail and misses the chance of applying for another great internship.

M24 (0110). Same as **M26** but the internship which Ashley misses the chance of applying for would have been a horrible experience.

M8 (0011). Steve sarcastically tells Ashley that the theory assignment is so easy that he solved it the moment he read it; but he means the exact opposite since the assignment is super difficult. However, Ashley naively believes him, waits till the last minute, and fails to finish the assignment on time.

M7 (001?). Same as **M8** but, although Ashley believes Steve, she starts working early on the assignment anyways.

M6 (0010). Same as **M8** but, after believing Steve and spending the time finishing other important work, the professor realizes that the assignment is too hard and cancels it.

M0 (0000). Here we adapt **M41** above as follows. Steve replaces Ashley and Sam replaces Steve. Further, right after his encounter with Sam and Daphne, Steve meets Ashley and good-heartedly informs her about the theory of computation quiz. Believing Steve, Ashley panics, spends the night studying theory of computation, and forgets about the networks quiz which she, consequently, fails.

□

4 Foundations for a Logic of Misleading

In this section, we lay the foundations for a logic of misleading as per the analysis presented thus far.

4.1 Ontology

Reasoning about misleading, construed after the analysis of Sections 2 and 3, rests upon a rather rich ontology. We take our ontology to at least conform to the following.

1. As mandated by **(M)**, the ontology includes agents, eventualities, and propositions. Agents are distinguished individuals who can have beliefs, intentions, and desires. (More on these below.) We follow (Hobbs 2005) in assuming a category of eventualities which are, intuitively, stretches of time characterized by some proposition's being true (or some state's holding (Ismail 2013).) Propositions are taken at face value, and assumed to be first-class inhabitants of our ontology. This simplifies the language and facilitates quantification over propositions. Such a notion of propositions may be modeled using reified fluents or, more generally, states, as suggested in (Ismail 2013).
2. To accommodate **BP**, **ID**, and **IH**, we follow the standard analysis of belief and intention. Hence, the ontology includes possible worlds, with belief and intention accessibility relations.
3. An account of causality is necessary for reasoning about the effects of misleading, as **IH** and **EQ** mandate. We follow the treatment of causality presented in (Hobbs 2005), which presupposes eventualities and possible worlds.
4. Whether the effect of misleading is positive, negative, or neutral is determined by the *desirability* of that effect. Likewise, an intention to hurt by misleading is an intention that misleading causes an undesirable effect. Hence, for **IH** and **EQ**, our ontology should accommodate a notion of desirability. To that end, we follow the theory of relative desire presented in (Doyle, Shoham, and Wellman 1991). That theory posits a preorder on *models* which are taken to be sets of literals of the logic. We opt for having models as secondary ingredients of our ontology, defined in terms of possible worlds.
5. Since beliefs and intentions, in general, vary over time, we assume a global time-line across all possible worlds.

To summarize, the ontology of misleading includes agents, eventualities, propositions, possible worlds, and a global clock. Moreover, for every agent a, a belief- and an intention- accessibility relation, respectively $\mathcal{R}_a^B$ and $\mathcal{R}_a^I$, are defined: $\mathcal{R}_a^B$ relates pairs of worlds and pairs of times and $\mathcal{R}_a^I$ relates pairs of worlds at a time. (More on this below.) Every world has an associated set of eventualities holding in it (Hobbs 2005); a function $\mathcal{E}$ maps each world w to its associated set $\mathcal{E}(w)$. Finally, a function $\mathcal{M}$ maps a world w to its associated *model*—a subset of $\mathcal{E}(w)$ of the eventualities of some propositional *literals* being true. Here we allude to a particular logical language (like the one presented below) to fix the set of literals. A relative desire relation $\succsim_\alpha$ for each agent α, akin to that of (Doyle, Shoham, and Wellman 1991), preorders the set of models.

4.2 Sketch of a Language

We present a sketch of a logical language $\mathcal{L}_M$ for reasoning about misleading scenarios. $\mathcal{L}_M$ is a first-order language amended with features for defeasible reasoning, symbolized by a connective $\leadsto$. We stay silent about exactly what those features are; we may interpret $\leadsto$ as in (McCarthy 1980), (Reiter 1980), or (Nute 1994), for instance. The vocabulary and informal semantics of $\mathcal{L}_M$ are outlined below.

Terms a, possibly subscripted, is an agent variable; e, possibly subscripted, is an eventuality variable; t, possibly subscripted, is a time variable; and w, possibly subscripted, is a possible world variable. The set of fluent/state terms is defined recursively as follows:

1. $P \in \mathbb{P}$ is a fluent constant, where $\mathbb{P}$ is a set of propositional constants disjoint from the rest of the alphabet.
2. p, possibly subscripted, is a fluent variable.
3. $Bel(\alpha, \phi, t)$ is a fluent functional term denoting agent $[\![\alpha]\!]$'s believing fluent $[\![\phi]\!]$ to be true-at-time-$[\![t]\!]$.
4. $Int(\alpha, \phi)$ is a fluent functional term denoting agent $[\![\alpha]\!]$'s intending fluent $[\![\phi]\!]$ to be true.
5. $Conv(\alpha, \phi, t)$ is a fluent functional term denoting agent $[\![\alpha]\!]$'s conveying fluent $[\![\phi]\!]$'s being true-at-t.
6. $cause(\epsilon_1, \epsilon_2)$ is a fluent functional term denoting eventuality $[\![\epsilon_1]\!]$'s causing eventuality $[\![\epsilon_2]\!]$ (Hobbs 2005). This is a fluent term not because causality between event tokens varies over time—it certainly does not—but because we would like such terms to appear as arguments of Bel and Int.
7. $DESIRE(\alpha, \phi)$ is a functional fluent term denoting fluent $[\![\phi]\!]$'s being desirable by agent $[\![\alpha]\!]$ (Doyle, Shoham, and Wellman 1991). This roughly means that, *ceteris paribus*, $[\![\phi]\!]$ is more preferred over $[\![\neg\phi]\!]$.[2]
8. If ϕ and ψ are fluent terms, then so are $\neg\phi$ and $\phi \wedge \psi$. Here we are overloading the sentential connectives.

Fluent terms of the first seven forms and their negations are the literals of the language.

Predicates $\mathcal{L}_M$ has four groups of predicate symbols:

1. $R^B(\alpha, \omega_1, \omega_2, t, t')$ is true if $([\![\omega_1]\!], [\![\omega_2]\!], [\![t]\!], [\![t']\!]) \in \mathcal{R}^B_{[\![\alpha]\!]}$; intuitively, at t' in ω_1, α believes ω_1-at-t to be identical to ω_2-at-t.
2. $R^I(\alpha, \omega_1, \omega_2, t)$ is true if $([\![\omega_1]\!], [\![\omega_2]\!], [\![t]\!]) \in \mathcal{R}^I_{[\![\alpha]\!]}$; intuitively, α's intentions at t in ω_1 are true in ω_2 at some time no earlier than t.
3. $holds(e, w, t)$ is true whenever $[\![e]\!] \in \mathcal{E}([\![w]\!])$ at time $[\![t]\!]$ and $Rexists(e, t)$ is true whenever $[\![e]\!] \in \mathcal{E}(r)$ at time $[\![t]\!]$ where r is the real world (Hobbs 2005).
4. $before(t_1, t_2)$ is true if $[\![t_1]\!]$ precedes $[\![t_2]\!]$ on the global time-line.
5. $Ev(\epsilon, \phi)$ is true whenever $[\![\epsilon]\!]$ is an eventuality of $[\![\phi]\!]$'s being true.

Axioms An $\mathcal{L}_M$ theory contains the following groups of axioms.

1. Appropriate axioms for R^B and R^I. For example, we can borrow the axioms in (Sakama, Caminada, and Herzig 2010) as is, modulo the translation from their modal language to our first-order $\mathcal{L}_M$ and accounting for temporality. These axioms restrict belief to a KD45 modality and intention to a KD modality, with two axioms for interaction between the two modalities.
2. Axioms requiring $before$ to be irreflexive, asymmetric, and transitive.
3. Axioms characterizing $cause$ from (Hobbs 2005).[3]
4. Finally, Ev is characterized by the following axioms which, we believe, are self-explanatory. Henceforth, we write $holds(x, y, z, t)$ as a short hand for $Ev(x, y) \wedge holds(x, z, t)$ and $Rholds(x, y, t)$ as a short hand for $Ev(x, y) \wedge Rexists(x, t)$. Unless otherwise indicated, all variables are universally quantified with widest scope.

AEv1. $\exists e[Ev(e, p)]$

AEv2. $\exists e[holds(e, Bel(a, p, t), w, t')] \Leftrightarrow \forall w_1[R^B(a, w, w_1, t, t') \Rightarrow \exists e_1[holds(e_1, p, w_1, t)]]$

AEv3. $\exists e[holds(e, Int(a, p), w, t)] \Leftrightarrow \forall w_1[R^I(a, w, w_1, t) \Rightarrow \exists e_1, t_1[holds(e_1, p, w_1, t_1) \wedge \neg before(t_1, t)]]$

AEV4. $\exists e, t[holds(e, cause(e_1, e_2), w, t)] \Leftrightarrow \forall t \exists e[holds(e, cause(e_1, e_2), w, t)]$

AEv5. $\exists e[holds(e, \neg p, w, t)] \Leftrightarrow \neg\exists e[holds(e, p, w, t)]$

AEv6. $\exists e[holds(e, p_1 \wedge p_2, w, t)] \Leftrightarrow \exists e_1, e_2[holds(e_1, p_1, w, t) \wedge holds(e_2, p_2, w, t)]$

4.3 Formalizing Misleading Scenarios

As indicated in Section 3, each of **M0–M80** is a conjunction of **(M)** together with four statements, one from each of Tables 1–4. Thus, it suffices to formalize **(M)** together with the twelve statements in the tables. The representation of scenario **M**i with encoding $\eta\delta\beta\varepsilon$ has the following general form

$$\begin{array}{l}
\exists a, p_1, p_2, t, t', t''[\\
\quad Rholds(E, Conv(a, p_1, t''), t) \wedge\\
\quad [Rholds(E, Conv(a, p_1, t''), t) \wedge RC(A, t)\\
\qquad\qquad \rightsquigarrow \exists e_1[Rholds(e_1, p_2, t')]] \wedge\\
\quad \exists e_2[(Rholds(e_2, \neg p_2, t') \vee Rholds(e_2, \neg Bel(a, p_2, t'), t)\\
\qquad\qquad] \wedge \Phi(\eta, \delta, \beta, \varepsilon)]
\end{array}$$

Here E is a place holder for the eventuality judged as misleading by agent A and $RC(A, t)$ stands for whatever A takes to be common ground in the real world at time t. $\Phi(\eta, \delta, \beta, \varepsilon) = IH(\eta) \wedge ID(\delta) \wedge BP(\beta) \wedge EQ(\varepsilon)$ represents the conjunction of statements corresponding to η, δ, β and ε from Tables 1–4, respectively.

$$\begin{array}{l}
I\,H(\eta) =_{\mathrm{def}} \exists p_3, e_1, e_2, e_3, e_4[\\
\quad Rholds(e_1, \Theta(\eta), t) \wedge\\
\quad Ev(e_2, Bel(A, p_2, t')) \wedge Ev(e_3, p_3) \wedge\\
\quad Rholds(e_4, Bel(a, cause(e_2, e_3)\\
\qquad\qquad\qquad \wedge DESIRE(A, \neg p_3), t'), t)]
\end{array}$$

[2] A note for readers familiar with (Doyle, Shoham, and Wellman 1991): Since the semantics of desire in that beautiful theory is based on models, which we do not have, we replace a model m with a possible world w. In the formal machinery, each mention of m is replaced by $\mathcal{M}(w)$.

[3] Again, this requires some adjustment. Hobbs (Hobbs 2005) takes worlds to be sets of eventualities. Thus, a world w in our ontology does not correspond to a world in Hobbs's; the set $\mathcal{E}(w)$ does.

η	$\Theta(\eta)$
0	$Int(a, \neg cause(E, e_2))$
?	$\neg\Theta(0) \wedge \neg\Theta(1)$
1	$Int(a, cause(E, e_2))$

Table 6: Different forms of $\Theta(\eta)$

δ	$\Delta(\delta)$
0	$Int(a, \neg Bel(A, p_2, t'))$
?	$\neg\Delta(0) \wedge \neg\Delta(1)$
1	$Int(a, Bel(A, p_2, t'))$

Table 7: Different forms of $\Delta(\delta)$

Table 6 shows the different forms of η. $IH(\eta)$ says that information source a has (or lacks) some intention regarding E's causing A to believe p_2, as indicated by η, and that this intention (or lack thereof) is simultaneous with E. Further, A's believing p_2 is believed by a to have an undesirable effect on A. In what follows, $\Delta(\delta)$, $\Psi(\beta)$, and $\Upsilon(\varepsilon)$ are as tabulated in Tables 7–9.

$$\begin{aligned}
ID(\delta) =_{\text{def}} \exists e_1, e_2, e_3[&\\
&Rholds(e_1, \Delta(\delta), t) \wedge\\
&Ev(e_2, Bel(A, p_2, t')) \wedge\\
&Rholds(e_3, Bel(a, cause(E, e_2) \wedge \neg p_2, t'), t)]
\end{aligned}$$

$$BP(\beta) =_{\text{def}} \exists e[Rholds(e, \Psi(\beta), t)]$$

$$\begin{aligned}
EQ(\varepsilon) =_{\text{def}} \exists e_1, e_2[&\\
&Rholds(e_1, cause(E, e_2), t) \wedge\\
&Ev(e_2, Bel(A, p_2, t')) \wedge\\
&\Upsilon(\varepsilon)
\end{aligned}$$

Example 1. As an illustration, we formalize case **M80** of Example 1 from Section 3. For readability, some variables have been replaced by more mnemonic (Skolem) constants.

$$\begin{aligned}
&Rholds(E, Conv(Steve, \neg Meeting, t'), t) \wedge\\
&Rholds(E, Conv(Steve, \neg Meeting, t'), t)\\
&\qquad \rightsquigarrow \exists e_1[Rholds(e_1, \neg Meeting, t')] \wedge\\
&Rholds(e_2, Bel(Steve, Meeting, t'), t) \wedge\\
&Rholds(e_3, Int(Steve, cause(E, e_4)), t) \wedge\\
&Ev(e_4, Bel(Ashley, \neg Meeting, t')) \wedge Ev(e_5, Deduct) \wedge\\
&Rholds(e_6, Bel(Steve, cause(e_4, e_5) \wedge\\
&\qquad DESIRE(Ashley, \neg Deduct), t'), t) \wedge\\
&Rholds(e_7, Int(Steve, Bel(Ashley, \neg Meeting, t')), t) \wedge\\
&Rholds(e_8, Bel(Steve, cause(E, e_4)), t) \wedge\\
&Rholds(e_9, DESIRE(Ashley, \neg Deduct), t') \wedge\\
&Rholds(e_{10}, cause(E, e_4), t) \wedge\\
&Rholds(e_{11}, cause(e_4, e_5), t')
\end{aligned}$$

□

β	$\Psi(\beta)$
0	$Bel(a, p_1, t'')$
?	$\neg\Psi(0) \wedge \neg\Psi(1)$
1	$Bel(a, \neg p_1, t'')$

Table 8: Different forms of $\Psi(\beta)$

ε	$\Upsilon(\varepsilon)$
0	$\forall p_3, e_3, e_4, t''[Rholds(e_3, Cause(e_2, e_4), t) \wedge Rholds(e_4, p3, t'') \Rightarrow$ $\exists e_5 Rholds(e_5, DESIRE(A, p_3), t'')$
?	$\neg\Upsilon(0) \wedge \neg\Upsilon(1)$
1	$\exists p_3, e_3, e_4, e_5, t''[Rholds(e_3, Cause(e_2, e_4), t) \wedge Rholds(e_4, p3, t'')$ $\wedge Rholds(e_5, DESIRE(A, \neg p_3), t'')]$

Table 9: Different forms of $\Upsilon(\varepsilon)$

5 Conclusion

We presented an account of misleading as a catalyst to trust erosion in information sources. A suitable bare-bones notion of what it means for an agent to judge an eventuality as misleading is the corner stone of our account. According to it, all misleading scenarios involve an information source's conveying some proposition which, given what the agent takes to be common ground, defeasibly imply another proposition that is either false or not believed to be true by the information source. We have identified eighty one variants of misleading as generated by four three-valued parameters: whether the source believes what they convey, whether they intend to deceive the agent, whether they intend to harm the agent, and whether misleading results in an undesirable effect to the agent. If this analysis is correct, a logical theory of misleading for trust erosion necessarily includes theories of belief, desire, intention, and causality. We have sketched a first-order language $\mathcal{L}_M$ to represent scenarios of misleading.

Future research can go in at least three fruitful directions. First, we need to go to the lab and conduct various experiments on human subjects to validate the details of our analysis of misleading. Second, a more thorough investigation of $\mathcal{L}_M$ and its properties is called for. Finally, we should turn to our long-term goal and introduce an account of trust erosion to $\mathcal{L}_M$.

References

Adler, J. E. 1997. Lying, deceiving, or falsely implicating. *Journal of Philosophy* 94:435–452.

Amgoud, L., and Demolombe, R. 2014. An argumentation-based approach for reasoning about trust in information sources. *Argument & Computation* 5(2-3):191–215.

Buller, D. B., and Burgoon, J. K. 1996. Interpersonal deception theory. *Communication Theory* 6(3):203–242.

Cox, J. C. 2004. How to identify trust and reciprocity. *Games and Economic Behavior* 46:260–281.

Demolombe, R. 2009. Graded trust. *In Proceedings of the Workshop on Trust in Agent Societies (TRUST'09)* 1–12.

Demolombe, R. 2015. Analytical decomposition of trust in terms of mental and social attitudes. In *The Cognitive Foundations of Group Attitudes and Social Interaction*. Springer. 59–74.

DePaulo, B. M.; Kashy, D. A.; Kirkendol, S. E.; and Wyer, M. M. 1996. Lying in everyday life. *Journal of Personality and Social Psychology* 70(5):979–995.

Doyle, J.; Shoham, Y.; and Wellman, M. P. 1991. A logic of relative desire. In Ras, Z. W., and Zemankova, M., eds., *Methodologies for Intelligent Systems: Proceedings of the 6th International Symposium, ISMIS '91*. Berlin, Heidelberg: Springer. 16–31.

Drawel, N.; Bentahar, J.; and Shashuki, E. 2017. Reasoning about trust and time in a system of agents. *Procedia Computer Science* 109c:632–639.

Elangovan, A.; Auer-Rizzi, W.; and Szabo, E. 2007. Why don't I trust you now? An attributional approach to erosion of trust. *Journal of Managerial Psychology* 22(1):4–24.

Ettinger, D., and Jehiel, P. 2010. A theory of deception. *Microeconomics* 2(1):1–20.

Fagin, R.; Halpern, J.; Moses, Y.; and Vardi, M. 1995. *Reasoning about Knowledge*. Cambridge, Massachusetts: The MIT Press.

Gneezy, U.; Rockenbach, B.; and Serra-Gracia, M. 2013. Measuring lying aversion. In *Journal of Economic Behavior & Organization*, volume 93. 293–300.

Gneezy, U. 2005. Deception: The role of consequences. In *The American Economic Review*, volume 95. 384–394.

Grice, P. 1989. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Haselhuhn, M. P.; Schweitzer, M. E.; and Wood, A. M. 2010. How implicit beliefs influence trust recovery. *Psychological Science* 21(5):645–648.

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

Hobbs, J. R. 2005. Toward a useful concept of causality for lexical semantics. *Journal of Semantics* 22(2):181–209.

Ismail, H. O., and Kasrin, N. S. 2010. Focused belief revision as a model of fallible relevance-sensitive perception. In *Proceedings of the 33rd German AI Conference (KI 2010)*. Springer Verlag.

Ismail, H. O. 2013. Stability in a commonsense ontology of states. In *Proceedings of the Eleventh International Symposium on Logical Formalization of Commonsense Reasoning (COMMONSENSE 2013)*.

Levine, E. E., and Schweitzer, M. E. 2015. Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes* 126:88–106.

Mahon, J. E. 2016. The definition of lying and deception. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

McCarthy, J. 1980. Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence* 13(1–2):27–39.

Nute, D. 1994. Defeasible logic. In *Handbook of logic in artificial intelligence and logic programming, volume 3: Nonmonotonic reasoning and uncertain reasoning*. New York, NY: Oxford University Press. 353–395.

Reiter, R. 1980. A logic for default reasoning. *Artificial Intelligence* 13(1–2):81–132.

Sabater, J., and Sierra, C. 2005. Review on computational trust and reputation models. *Artificial Intelligence Review* 24(1):33–60.

Sakama, C.; Caminada, M.; and Herzig, A. 2010. A logical account of lying. *Proceedings of the 12th European Conference on Logics in Artificial Intelligence (JELIA 2010)* 286–299.

Sakama, C. 2011a. Dishonest reasoning by abduction. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-11)*, 1063–1068.

Sakama, C. 2011b. Logical definitions of lying. In *Proceedings of the 14th International Workshop on Trust in Agent Societies (TRUST11)*.

Sakama, C. 2015. A formal account of deception. In *Proceedings of the AAAI Fall 2015 Symposium on Deceptive and Counter-Deceptive Machines*, 34–41.

Saul, J. 2012. Just go ahead and lie. *Analysis* 72:3–9.

Schillo, M.; Funk, P.; and Rovatsos, M. 2000. Using trust for detecting deceitful agents in artificial societies. *Applied Artificial Intelligence* 14(8):825–848.

Schweitzer, M. E.; Hershey, J. C.; and Bradlow, E. T. 2006. Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes* 101:1–19.

Serota, K. B.; Levine, T. R.; and Boster, F. J. 2010. The prevalence of lying in America: Three studies of self-reported lies. *Human Communication Research* 36:2–25.

Stalnaker, R. 2002. Common ground. *Linguistics and Philosophy* 25:701–721.

Stokke, A. 2016. Lying and misleading in discourse. *The Philosophical Review* 125(1):83–134.

Vendler, Z. 1957. Verbs and times. *The Philosophical Review* 66(2):143–160.

Wagner, A. R., and Arkin, R. C. 2011. Acting deceptively: Providing robots with the capacity for deception. *International Journal of Social Robotics* 3:5–26.

Wagner, A. R., and Robinette, P. 2015. Towards robots that trust: Human subject validation of the situational conditions for trust. *Interaction Studies* 16(1):89–117.