Lying, deception and strategic omission : definition and evaluation

Benjamin Icard

THÈSE DE DOCTORAT

de l'Université Paris Sciences et Lettres
PSL University

Préparée à l'École normale supérieure, Paris

Lying, Deception and Strategic Omission
Definition & Evaluation

École doctorale n°158

CERVEAU, COGNITION, COMPORTEMENT

Spécialité SCIENCES COGNITIVES

COMPOSITION DU JURY :

M. Didier BAZALGETTE, Ingénieur
Direction Générale de l'Armement,
Examinateur

M. Denis BONNAY, MCF
Université Paris X / IRePh / IHPST,
Examinateur

M. Hans VAN DITMARSCH, DR
Université de Lorraine / LORIA,
Rapporteur et Président du Jury

M. Paul ÉGRÉ, DR
ENS Ulm / IJN,
Directeur de la thèse

M. Pascal ENGEL, DE
EHESS / CRAL,
Rapporteur

Mme Marie-Jeanne LESOT, MCF
Université Paris VI / LIP6,
Examinatrice

Soutenue par **Benjamin ICARD**
Le **4 Février 2019**

Dirigée par **Paul ÉGRÉ**

ENS ÉCOLE NORMALE SUPÉRIEURE | PSL RESEARCH UNIVERSITY PARIS

Lying, Deception and Strategic Omission

Definition & Evaluation

Benjamin Icard

Lying, Deception and Strategic Omission

Definition & Evaluation

PhD Thesis

Submitted in partial fulfillment of the requirements for the degree
of
Doctor in Cognitive Science
by
Benjamin Icard

Under the supervision
of
Professor Paul Égré

Examination Committee

Supervisor:	M. Paul Égré	École normale supérieure - IJN
Committee:	M. Didier Bazalgette	Direction Générale de l'Armement
	M. Denis Bonnay	Université Paris X - IRePh / IHPST
	M. Hans van Ditmarsch	Université de Lorraine - LORIA
	M. Pascal Engel	EHESS - CRAL
	Mme Marie-Jeanne Lesot	Université Paris VI - LIP6

À mes parents

À Alice

DEC
DÉPARTEMENT
D'ÉTUDES
COGNITIVES

This work was supported by a doctoral scholarship from the Mission pour la Recherche et l'Innovation Scientifique of the Direction Générale de l'Armement (DGA-MRIS). The author also received funding from the Project "*New Ideas in Mathematical Philosophy*", the PSL Research Grant "*Improving Prediction for a Better World*", the ANRs 14-CE30-0010-01, 10-LABX-0087 IEC, 10-IDEX-0001 PSL*. The writing of the dissertation has been helped by the ANR 17-EURE-0017 FrontCog.

Printed and bound at the Département d'Études Cognitives, ENS Ulm.

Acknowledgments

First and foremost, I would like to express my profound gratitude to Paul Égré, my PhD Advisor, for his invaluable input, precious guidance, support, and generosity over the years. I had the great opportunity to meet Paul during my Master's Degree, and since then on many occasions I could realize how lucky I was to work under his supervision. Without his continuing support and wise advice, things would have been much more arduous.

I am very grateful to Hans van Ditmarsch and Pascal Engel for accepting to be *rapporteurs* of my dissertation thesis. I also want to express my sincere gratitude to Marie-Jeanne Lesot, Denis Bonnay and Didier Bazalgette for being *examinateurs* at my defence.

Special thanks go to Denis Bonnay and Hans van Ditmarsch for their very useful feedback and advice during, and between, my Comités de suivi de thèse. I am also thankful to Brent Strickland for his substantial help since the beginning of my empirical investigations on the topic of lying.

I thank Didier Bazalgette for his constant support before and during my PhD, thanks to him I could undertake this thesis with a DGA scholarship and for associating me to applied projects then. I am also particularly grateful to Philippe Capet for his friendship and constant attention concerning topics related to deception in intelligence warfare.

I also thank the following scholars, who dedicated some time and thought to my

work: Alexandru Baltag, Peter van Emde Boas, Raul Fervari, Mathias Girel, Neri Marsili, Sonja Smets in particular.

I am thankful to the audience and commentators in the conferences, workshops and seminars in Paris, Amsterdam, Leiden and Leuven for their interest, and their sometimes challenging inputs.

I would also like to thank Ghislain Atemezing, Reinaldo Bernal, Samira Boujidi, Romain Bourdoncle, Serge Bozon, Bouchra Budel, Géraldine Carranante, Jean-Marie Chevalier, Rose-Hélène Doranges-Daupin, Nathalie Evin-Abitbol, Aurélien Fermo, Miguel Flament, Martin Fortier, Clémentine Fourrier, Guillaume Gadek, Vincent Gaudefroy, Benoit Gaultier, Félix Geoffroy, Janek Guerrini, Anna Giustina, Helena Hachmann, Guillaume Herzog, Dan Hoek, Rojdi Karli, Ekaterina Kubyshkina, David Landais, Baptiste Lanne, Sandra Lasry, Bénédicte Legastelois, Nathalie Marcinek, Hugo Mell, Raphaël Millière, Pierre-Alexandre Miot, Boian Nikiforoff-Kharisanoff Vega, Mélanie Sarzano, Ian Shilito, Andrès Soria-Ruiz, Claire Sourdin, Chloé Tahar, Tristan Thommen, Louis Vayssette, Juliette Vazard, Jacques-Henri Vollet.

Lastly, I would like to thank Alice, my parents and my brother and sister for their continuous support, for their interest and for their care all along these years of intense work.

Paris, December 2018

Contents

Introduction

> The Dodo says that the Hatter tells lies. The Hatter says that the March Hare tells lies. The March Hare says that both the Dodo and the Hatter tell lies. Who is telling the truth?
>
> *A puzzle from Lewis Carrol's diary.*[1]

Information integrity is of common concern. In every situation in which communication matters, preserving the quality of information is an imperative need as well as a continuous challenge. Understanding how information flows, identifying patterns, assessing sources, cross-checking information becomes more and more demanding as data become more prevalent in society. Unfortunately, information is more manipulated than ever before. While manipulations of all sorts have always existed, new technologies and social media make lying, deceiving or more subtle forms of disinformation as prevalent in daily life as they were, during the Cold War, in the hidden realm of intelligence warfare. Whether these strategies rely on simple mechanisms or on more convoluted scenarios, they are just as difficult to define as to detect. This thesis proposes to help meet these challenges.

Unsurprisingly, information quality has also been a field of common interest. Information being a multidimensional concept, evaluating and protecting data quality calls for the combination of different methods and tools at the interface

[1] This riddle and other Carrollian puzzles have been compiled and edited by Edward Wakeling [see Wakeling 1992].

between computer science, analytic philosophy and psychology. Broadly speaking, two schools of thought have contributed to define the field of information quality since the 1990s.

The first one, the MIT school, combines complementary approaches, the one *empirical*, the other *theoretical*. The empirical approach consists in using empirical surveys to isolate *dimensions* that are considered crucial by academics, information practioners and consumers with respect to information quality [Wang 1998, Lee *et al.* 2002]. In addition to *accuracy* (or *truth*) of message contents, dimensions of *relevance*, *timeliness* and *completeness* are perceived as being as important as *accuracy* for evaluating the quality of data sets. But this approach based on empirical investigation has been later opposed to a more "*ontological*" approach [Wand & Wang 1996], also named "*theoretical*" [Batini & Scannapieco 2006]. Contrary to the first, the ontological approach looks for theoretical and a prioristic definitions of the dimensions involved in information quality. Crucially, they aim at better understanding how information quality can be *defective*, and even *deceptive*, when one of those dimensions is breached.

The second school of thought, the Italian school, has taken over in this theoretical entreprise by following similar intuitions as Wand & Wang [1996]. Batini & Scannapieco [2006] look at the preservation of information quality as a *fusion operation* with specific issues and constraints. For instance, how can sets of data be merged when they are based on qualitative discrepancies? How to accommodate data that are potentially defective and conflicting with each other? But the leading role in the Italian Group has been played by Floridi who has defined a new field called "*philosophy of information*" following Shannon & Weaver's Information Theory as well as Dretske's influential work on the epistemology of information [Shannon & Weaver 1949, Dretske 1981]. Initial interests on information have been considerably extended to new areas of investigation such as semantic and pragmatic issues, philosophy of science, logic, ethics, etc [see Floridi 2008 2011, Floridi & Illari 2014, for surveys].

As a matter of fact, the analysis of deceitful information has been left aside by the Italian school until works by Luciano Floridi and, more centrally, Don Fallis. Before the new millenium, *misinformation* and a fortiori *disinformation* were

not considered as *information* stricto sensu because semantic data were required to be *true* to count as information *per se* [e.g. Dretske 1981, Grice 1989, Floridi 1996]. From a theoretical perspective, this epistemological debate can be linked to the debate between *subsective* and *non-subsective adjectives* in semantics [Kamp & Partee 1995]. Subsective adjectives are adjectives like *blue* or *good* such that the adjective-noun extension is a subset of the noun extension: a *blue bike* is a *bike* and a *good fisherman* is a *fisherman*. By contrast, non-subsective adjectives are adjectives like *decoy* or *alleged* such that the adjective-noun extension is not necessarily compatible with the denotation of the noun: *is a decoy duck a duck? Is an alleged thief a thief?* More specifically, we can distinguish between two kinds of non-subsective adjectives: *privative non-subsective adjectives* (e.g. *fake*) such that the denotations of the adjective-noun phrase and the noun extension are mutually exclusive (a *fake gun* is not a *gun*), and *plain non-subsective adjectives* (e.g. *alleged*) such that their respective denotations are compatible with each others (an *alleged thief* can be a *thief*). Most authors agree with those various distinctions but some do not [see Pavlick & Callison-Burch 2016, for instance].

Consistent with the dominant view in this semantic debate, Dretske [1983, 57] considers that "*false information, misinformation (...) are not varieties of information — any more than a decoy duck is a duck*". In his *Studies in the Way of Words* published in 1989, Grice makes a similar statement: "*False information is not an inferior kind of information; it just is not information*" Grice [1989, 371]. Since then, however, times have changed. Following Fetzer [2004] and Scarantino & Piccinini [2010] for whom *any* meaningful data is information per se, Fallis [2009a] and Floridi [2011] agree that misinformation and disinformation count as real information. Misinformation is information that is *accidentally* defective while disinformation is information that is *purposefully* defective.

Over the years, epistemologists have devoted efforts to the analysis of deceptive attitudes by combining conceptual analysis with experimental protocols and formal logic. They have offered definitional accounts of lying, deception and disinformation. In doing so, however, they have made choices that inevitably went with ignoring other aspects of deceptive information. My thesis project is motivated by the following observations:

➤ ***Methodological Approaches.*** When deceptive attitudes have been investigated by epistemologists, *theoretical* and *formal accounts* have been favored over *empirical protocols*. But deceptive strategies are social attitudes that are worth investigating from theoretical, formal and empirical perspectives. This thesis aims to combine these different approaches for analyzing cases of deception.

➤ ***Theoretical Aspects.*** Whenever investigated, *standard cases* (such as *lies*) have received more attention than more *non-standard cases* (such as *misleading inferences* and *strategic omissions*). But non-standard cases are as important as classical cases when studying deceptive strategies. This thesis aims to give a more balanced account of both cases of deception.

➤ ***Practical Interests.*** Epistemologists have been mostly interested in *definitional aspects* of deceptive attitudes. They have generally left *evaluative aspects* to behavioural psychologists working on lie detection or to computer scientists concerned by information security. But definitional and evaluative aspects are two sides of the same coin. Accordingly, this thesis proposes to consider these two perspectives through a comprehensive account of deceptive attitudes.

1. Methodological Approaches

From a *methodological* perspective, epistemologists have focused on aprioristic accounts to understand the dynamics of deceptive attitudes. Based on conceptual analysis, they have tried to isolate (sets of) *necessary* and *sufficient conditions* for capturing these complex attitudes adequately. The definitions they have put forward are called "checklist definitions" [Fillmore 1975] and are assumed to exhaust the meanings of the deceptive attitudes they intend to capture. Soon afterwards, however, checklist accounts were challenged by prototype theory, motivated by Eleanor Rosch's criticisms of explicit definitions, and by semantic and pragmatic results from experiments [see Rosch 1973, Coleman & Kay 1981]. Prototypes do not rely on categorical lists of conditions one should fulfill to be categorised as a *liar* or as a *disinformer* (for instance). They rely on *typical properties*

that are more or less salient when one behaves as a *liar* or as a *disinformer*. In that sense, typical properties are not as restrictive as necessary conditions in checklist accounts.

More recently, logic has entered the definitional scene, particularly logics of knowledge and belief, to see more clearly into this complexity. Static doxastic logics have been used to characterize lying, many aspects of deception, as well as dishonesty [Capet 2006, Sakama *et al.* 2010, Sakama & Caminada 2010, Sakama *et al.* 2014]. Epistemic logicians have introduced specific modal operators in syntax to express the speaker's intention to deceive (in particular). Semantically, these attitudes have been interpreted in basic Kripke models enriched with further accessibility relations to express preferences and intentions. More recently, dynamic epistemic frameworks have also been devised for the logics of lying, truth-telling and bluffing [van Ditmarsch *et al.* 2012, van Ditmarsch 2014], as well as for more exotic cases like self-deception [Sakama 2015] or lies that were *false* but become *true* once announced [Agotnes *et al.* 2016]. These settings have also helped define *types* of epistemic agents depending on their *degree of rationality* or of the *kind of information* they convey.

Concerning rationality-based types, van Benthem & Liu [2004] and Liu [2009] have characterized more *realistic agents* than those usually prescribed by dynamic epistemic settings. They have studied agents with bounded rationality due to limited inferential power, introspection, observation and memory. Concerning information-based types, updates and other revision policies have been proposed to characterize cooperative and non-cooperative agents depending on the information they convey. *Objective* truth-tellers and liars, who respectively convey *true* and *false* information purposefully, have been modelled by Liu [2009] as well as van Ditmarsch [2014], the latter being also interested by bluffers and *subjective* versions of truth-tellers and liars. Then Liu & Wang [2013] gave a syntactic counterpart to those types in a language that authorized self-reference. Their goal was to verify a classic solution to Boolos's version of Raymond Smullyan's Knights and Knaves puzzle.[2] In van Ditmarsch proposal, however, the perspectives of

[2] This puzzle has been so-labelled by Smullyan in his 1978 book on paradoxes and puzzles [Smullyan 1978]. A fictional island hosts two kinds of inhabitants: some who always tell the truth (Knights) and some who always tell falsities (Knaves). A visitor to the island is asked

the deceiver are considered along with the perspectives of the dupes who can either be credulous, skeptical or belief revising agents. This complementary approach was touched upon by Liu [2009] but not systematically investigated until van Ditmarsch [2014].

In this whole theoretical entreprise, experimental epistemology has been generally left aside. In 1981, Coleman & Kay published results supporting prototypical accounts of lying that challenged classical definitions based on necessary and sufficient conditions [Coleman & Kay 1981]. But since then, experimentalists have remained on the fringes of theoretical discussions. It was not until 2013 that Adam J. Arico and Don Fallis released empirical data showing that ordinary English speakers count *bald-faced lies* (viz. *believed-false utterances made without any intention to deceive*) and *proviso lies* (viz. *utterances in which the speakers adds a proviso that undermines any warrant of truth*) as lies in the proper sense [Arico & Fallis 2013]. Then John and Angelo Turri argued in 2015 for a revised (checklist) definition of lying called "objective", in which lying not only implies uttering a content you *subjectively* believe to be false, but also implies that the content itself is *objectively* false [Turri & Turri 2015]. Immediately, though, Wiegmann, Samland and Waldmann responded with new results showing that the Turris' protocols were possibly ill-founded and that the subjective definition was sufficient as it stands [Wiegmann *et al.* 2016].

Since then, further investigations have followed that argue that the speaker's *intention to deceive* is not necessary for lying [Turri & Turri 2016, Meibauer 2016b, Rutschmann & Wiegmann 2017]. These cases referred as "*bald-faced lies*" are deliberate false utterances that the speaker and the addressee know, or at least believe, to be false.[3] Except for a few philosophers [e.g. Sorensen 2007, Fallis 2009b, Stokke 2013], most disagree that such false utterances are lies, either because they suspect that bald-faced assertions are not real assertions [see Leland 2015, Keiser 2016] or that the speaker has still some intention to deceive in case of bald-faced lies [see Lackey 2013, Dynel 2015, Meibauer 2016a].

to deduce who are the truth-tellers and who are the liars by asking inhabitants a finite set of questions. Boolos [1996] proposed a personal version of this puzzle called "The Hardest Logic Puzzle Ever", and based on a variation due to computer scientist John McCarthy.

[3] For instance, a politician makes a bald-faced lie when he or she claims that they are "*honest*" while having been recently convicted of misusing public money.

Other empirical studies concluded that one can lie by *falsely implicating* [Wiegmann & Willemsen 2017] or by *omitting relevant facts*. In this latter case, Wiegmann & Willemsen [2017] and Wiegmann *et al.* [2017] have tested folk intuitions on two cases of omission. One case is when the dishonest speaker refrains from communicating some piece of information by changing the topic of the conversation. But another case of omission is when the dishonest speaker hides information from the addressee by witholding some relevant part of the truth. They observed that if lay participants consider that the speaker lied in the first case, they do not consider that the he or she did in the second case. Aside from these recent forays, however, the analysis of unreliable attitudes has remained neglected by empirical epistemologists.

Based on this background, my *first goal* in this thesis is to combine conceptual, formal and experimental resources for defining and evaluating deceptive attitudes. Three deceptive attitudes ranging from standard cases (*lying*) to more peripheral ones (*misleading defaults* and *strategic omissions*), will be studied using a multi-methodological approach. Chapter 1 focuses on the definition of lying through conceptual and experimental lenses. Chapter 2 combines conceptual and formal tools to analyze a paradox that concerns *deception by omission* through a misleading default rule. Chapters 3 and 4 propose an integrative framework for evaluating information quality and spotting deceptive attitudes based on the strategic use of vagueness (in particular). In that latter case, my approach will be conceptual and formal but strongly motivated by empirical findings researchers have made.

2. Theoretical Aspects

From a *theoretical* perspective, epistemologists have concentrated their efforts on *standard cases* of unreliable attitudes. Lying, for instance, is one of these paradigmatic cases. Traditionally, a speaker lies to some specific addressee if and only if the speaker makes a dishonest utterance with the intention to deceive the addressee. Broadly speaking, epistemologists have asked whether making an utterance and intending to deceive were necessary conditions for lying *per se*. They have also wondered whether the speaker's utterance must also be false to count

as a lie. In lying, deception concerns the *semantic meaning* of the content uttered by the speaker. In fact, the speaker's utterance aims at making the addressee believe the opposite of the semantic meaning the speaker has in mind.

But there are many other attitudes by which the speaker is unreliable. Deception and disinformation are not based only on the literal meaning of the speaker's utterance. They rely both on semantic and pragmatic features. Deception is classically defined as "*causing a false belief that is known or believed to be false*" by the deceiver [Mahon 2015]. Disinformation is defined as the dissemination of "*misleading information that is intended, or at least foreseen to be misleading*" [Fallis 2009a]. Unlike for lies and disinformation, deceptions are not necessarily made through utterances and are not necessarily intentional [e.g. Demos 1960, Chisholm & Feehan 1977, Adler 1997], even though some philosophers do not agree that there can be non-deliberate deceptions [e.g. Barnes 2007, Carson 2010, Saul 2012]. More crucially, both deception and disinformation can rely on *semantic aspects* (as in *lies*) but also on *pragmatic aspects*, namely on misinterpretations that are intended by the speaker. Pragmatic strategies of this kind are for instance *double bluff*, *presupposition failures* and *false implicatures*. Though being indirect strategies compared to blatant lies, these deceptive mecanisms are considered *classical* for being pervasive in society.

In *double bluff strategies*, the speaker wants the addresse to falsely believe that a content p is false. But instead of simply uttering that *not-p*, the speaker starts out by arousing the addressees' suspicion in order for them to believe that the speaker has misleading intentions (*first bluff*). As a second step, the speaker simply utters that p to them. Since the addressees believe the deceiver to have misleading intentions, they will come to wrongly believe that p is false and will be deceived afterwards (*second bluff*). Double bluff strategies have been studied by Fallis [2014] based on Vincent & Castelfranchi [1981, 764-766] who qualify those strategies as "*pretending to lie*" or "*lying while saying the truth*". An example of double bluff is the following one [based on Fallis 2014]: imagine that the authorities are looking for a thief who is used to stealing with a mate. If the policemen go at the mate's place to ask him or her where the thief is, they will expect the mate to lie for protecting his or her partner. So the mate can tell the policemen the exact truth

since they will probably believe the opposite and be deceived.[4]

Presupposition faking happens when the speakers counterfeit their entitlement to make an utterance they are not entitled to make because appropriate preconditions are missing [e.g. Harder & Kock 1976, Vincent & Castelfranchi 1981, Meibauer 2014]. Preconditions can be missing either because the content of the utterance fails to refer to anything real (i.e. there is not state of affair corresponding to the facts the content describes), or because the speakers themselves are not justified making the utterance they are making in the context. In that case, the speakers lack evidence for being entitled to make their utterances. Presupposition faking can be seen as intended presupposition accommodation when there is, in fact, presupposition failure. In case of presupposition faking, the deceiver deliberately makes a misleading utterance that intends the addressee to be mistaken about his or her preconditions. Let us give an example of presupposition faking inspired by Vincent & Castelfranchi [1981, 763]: imagine that you are organizing a party in Paris and say to a renowned guest: "It is rather a pity that Elizabeth and Philip are at Windsor this week-end", when in fact you are not on first-name terms with the British royal couple. In such a case, you are faking the preconditions that would allow you to make this utterance appropriately.

Another standard case of deception is *false implicatures* [Adler 1997, Fallis 2014]. In false implicatures, the literal meaning of the speakers' utterances are true and may be believed to be true, but the pragmatic interpretations the speakers intend are false. Most authors like Fallis [2009b] and Sorensen [2012] argue that false implicatures are *not lies* because the explicit meaning of the speaker's utterance is not defective although the implicit meaning of the utterance is. In fact, speakers do not believe their utterances to be false and are not lying for this very reason. But according to Adler [1997], Meibauer [2005 2014], Dynel [2011] and more recently Viebahn [2017], untruthful implicatures that are directly intended by the speakers are lies stricto sensu. Ones who conversationally implicate believed-false statements are lying even though the literal statements they made turn out

[4] A famous instance of *double bluff* is given by Sigmund Freud in his book *Jokes and Their Relation to the Unconscious* translated in 1960: "*Two Jews met in a railway carriage at a station in Galicia. 'Where are you going?' asked one. 'To Cracow', was the answer. 'What a liar you are!' broke out the other. 'If you say you're going to Cracow, you want me to believe you're going to Lemberg. But I know that in fact you're going to Cracow. So, why are you lying to me?'.*" [see Freud 1960]. This case of double bluff is analyzed by Fallis [2014].

to be true. False implicature is illustrated by the *Story of the Mate and the Captain* given by Posner [1980] [see also Meibauer 2005 2011]:

> "*A captain and his first mate do not get along well. The mate is a heavy drinker and the captain who never drinks alcohol, wants the situation to stop. So, when the mate is drunk again, the captain writes into the log: 'Today, May* 6^{th} *, the mate is drunk'. When the mate discovers the entry, he is so angry that he looks for revenge. Then, he writes in the log: 'Today, May* 20^{th}*, the captain is not drunk'.*"

The mate's entry is *true* since the captain does not drink any alcohol but it *falsely* suggests that the captain is *usually* drunk, which cannot be possible since he never drinks alcohol. But aside from these standard cases, some non-standard cases also count as instances of deception and disinformation. Strategic omissions and misleading defaults are such non-standard cases. But whether intentional or not, these attitudes have received less attention in the literature than more classical cases.

The few times when strategies of omission have caught the attention, they have been classified as instances of *withholding information*. Bok [1983, 5-6] has charaterized omission as *keeping secrets*: "*to keep a secret from someone (...) is to block information about it or evidence of it from reaching that person, and to do so intentionally*". The speaker keeps a secret s from some addressee if and only if the speaker intentionally and actively withholds the information contained by the secret s from reaching the addressee. In 1988, Scheppele has provided a complementary definition of keeping secrets in which the speaker *no longer acts actively* to keep secret s from the addressee [see Scheppele 1988]. Carson [2010, 57] as well as [Lackey 2013, 240-241] then insisted on the distinction between *keeping secrets* — that consists in *concealing*, or *hiding*, information that is already available, from *withholding information* that consists in non-disclosing information one wants to omit for strategic reasons. As I said, keeping secrets implies that the speaker plays an active role in the deceptive process. On the contrary, withholding information is strict omission: the speaker does not intervene to keep the addressee in the dark about information he or she has. Following Chisholm & Feehan [1977]'s distinction between "*deception by commission*" and "*deception by omission*", Fallis [2014

2018] looks at the speaker's epistemic goals when he or she conceals information or simply withholds it. According to Fallis, the speaker *causes* the addressee to be deceived in the first case while the speaker simply *leaves* the addressee in the dark in the second case. As for concealment, withholding information should be blamed for epistemic reasons, though not as much as for information concealment. In cases of omission, the addressee is left in a worse epistemic state than he or she should have been according to the Gricean principles of cooperation.

Misleading default interpretations are close to *false implicatures*. Misleading defaults consist, for the speaker, in making the addressee infer a presumptive or probable conclusion that is false. But despite being false, this conclusion is a perfectly reasonable interpretation to make. The conclusion the hearer infers is the predictable meaning one would normally infer from the speaker's utterance in natural circumstances. Some have compared defaults to implicatures, more precisely generalized conversational implicatures, — the latter being defined as context-independent pragmatic inferences that are automatic and unconscious [Levinson 1995 2000, Horn 2004]. But the status of defaults with respect to implicatures is unclear and a subject of debates: Levinson [2000] classifies defaults as implicatures per se, Récanati [2004] considers them as a pragmatic enrichment of the output of syntactic processing while Bach [1994] and Horn [2006] adopt an intermediate position in which default meanings are a subpart of what remains implicit in what is explicitly said by the speaker. Debates generally revolve around the cancellability (or defeasibility) of default interpretations, their availability, whether they arise locally (on the basis of a proposition) or globally (during the process of interpretation), and the time it takes for deriving default conclusions (and which is usually considered as being shorter than for other inferences). Default interpretations, and in particular misleading default interpretations, have also been studied in non-monotonic logics as being cases in which the hearer infers a default conclusion which is incorrect based on incomplete beliefs. According to Caminada [2009] and Sakama *et al.* [2010], the speaker also needs to withhold relevant information to make the default interpretation misleading. If the speaker was more informative, he or she would prevent the hearer from reaching a wrong conclusion based on this false default interpretation.

Based on those observations, my *second goal* in this thesis is to study two non-standard cases that are instances of *strategic omissions* and *misleading defaults* to improve our understanding of non-standard cases of deception. The paradox I offer to analyze in Chapter 2 deals with the interpretation one would normally commit after hearing that one will be deceived in some way or another. In that situation, one would usually infer that he or she will be deceived by some deceptive action (*commission*). This is the default conclusion to be reached in common circumstances. But if one is deceived by no action (*omission*), this conclusion is false and one is in fact deceived on the *type* of deception itself.

This thesis will also investigate strategic omissions through the interplay between lying and vagueness (see Chapter 3 and Appendix). In *Lying and Vagueness* published in 2018, Égré and I contrast cases of *half-truths* with cases of *omissions* [see Égré & Icard 2018]. The term *half-truth* has received different senses in the literature but our use coincides with the one of Engel [2016] and is very close to Vincent & Castelfranchi' definition of "*deliberate ambiguity*": "*Given an utterance with two possible interpretations or readings in a given context, one of which is true for [the speaker] A and one of which is false, A may exploit the ambiguity hoping and intending that [the hearer] B understands the false reading*" [Vincent & Castelfranchi 1981, 763]. In that sense, half-truths are pragmatic exploitations of semantic vagueness in which the speaker makes an utterance that is borderline between truth and falsity. By contrast, omissions are instances of (purely) *pragmatic vagueness* whereby speakers are less informative than they should according to Gricean communication principles [Grice 1989]. In our paper, half-truths and omissions are considered as subcases of semantic and pragmatic vagueness, and compared to the standard case of lying.

3. Practical Interests

From a more *practical* perspective, epistemologists have been mostly interested in *defining* deceptive attitudes. They have been less concerned than psychologists and computer scientists with *evaluating* informational strategies in order to *detect* misleading ones. But, in fact, defining and evaluating deceptive attitudes are two sides of the same coin. One attempts to define attitudes in order to detect them

when they materialize. Conversely, detecting deceptive attitudes requires having clear definitions of them to know when they can be ascribed on a fair basis. On the whole, epistemologists have focused mostly on definitions whereas computer scientists have concentrated on evaluation and psychologists, on detection. I briefly review works from psychologists and computer scientists on those issues before presenting my own perspective.

Psychologists have contributed a lot to the developpment of detection devices. The best known technique is the "*polygraph*" also known as "*lie detector*". However, a polygraph does not directly detect lies but the *physiological signals*, or *arousals*, that are caused by telling lies. The questioning procedure that was first used is the "*control test question*" or CQT. In CQT, suspects are asked *control questions* to measure the arousal they provoke when they actually lie. Based on this measure, examiners will be able to detect whether the suspects lie or not when answering to the test questions. In that respect, control questions are deliberately vague in order to force suspects to blatantly lie to them and generate arousal. Then *test questions* are asked that are relevant to the crime under investigation such as "*Did you kill your wife on July* 5^{th}*?*". Contrary to guilty suspects, innocents suspects won't lie at such questions and, therefore, won't show arousal compared to the arousal they have shown at the control questions. In contrast, guilty suspects will generate equal arousal in both cases and give themselves away. The accuracy of CQT, however, has been highly criticized for being based on pseudoscience: *no strong evidence indicates that CQT detects deception at a rate better than chance* [Research Council 2003]. CQT is known to have stress-inducing effects which strongly undermine the accuracy of collected results and may lead to convict innocent suspects.

A more efficient alternative to CQT is known as GKT for "*guilty knowledge test*" [see Lykken 1959 1998]. In this technique, examiners do not ask single "Yes/No" questions (such as "*Did you kill your wife on July* 5^{th}*?*") but series of suggestive questions that address details of the crime only known to the criminal and to the authorities: "*How was your wife killed on July* 5^{th}*? Was she drowned? Was she hit on the head? Was she stabbed? Was she strangled? etc*". The idea is that the correct option amongst the series of questions will provoke *more arousal* in the guilty examinee

than the other options. One of the advantages of GKT is that false positives are controlled by the range of alternatives the examiner provides. Accordingly, polygraphs based on GKT proved to be more accurate than CQT ones but have remained controversial. If GKT have been proved to have a validity above chance level, guilty knowledge tests still perform far below perfection [e.g. Meijer & Verschuere 2015, Meijer *et al.* 2016].

In recent years, other techniques have been developped that rely on verbal and non-verbal cues. Most of these techniques are based on the "*Cognitive Approach to Deception*" according to which lying and deceiving are *more demanding* than truthtelling in terms of cognitive load. Deceptive attitudes take more time because unlike truth-tellers, liars need to fabricate a new story instead of simply rememoring a true one [Vrij *et al.* 2006, Christ *et al.* 2008]. In addition to that, liars should have a clear view of the story they elaborate to avoid contradicting themselves [Suchotzki *et al.* 2017]. This cognitive approach is grounded on empirical data showing that truth-telling does not provoke more brain activity than lying but that lying generates more activity in prefrontal and frontal regions of the brain which are usually activated by complex cognitive tasks [Christ *et al.* 2008, Abe 2009, Ganis & Keenan 2009, Gamer 2011]. Advocates of the cognitive approach have proposed to improve GKT by measuring and comparing the suspects' reaction times to test questions in case of truth-telling and in case of lying [e.g. Seymour & Kerlin 2008, Debey *et al.* 2012, Verschuere *et al.* 2014]. In this thesis, information evaluation won't be studied primarily from the perspective of behavioural psychology but from the one of epistemology. Detection won't be investigated by analyzing the cognitive mechanism involved in deceptive attitudes but by helping figure out the taxonomy of attitudes they correspond to, and the way these attitudes can be ascribed from a procedural perspective. However, it is worth mentioning works from psychologists since they have contributed a lot to improve the detection of deceptive attitudes.

Aside from behavioural psychologists, information evaluation has also aroused strong interest amongst computer scientists, especially amongst those working on applications to data security and intelligence processing. They have developped methods for helping practionners make more accurate evaluations of intelligence

messages, namely of the credibility of message contents as well as the reliability of their sources. Rightly conceived, informational messages are linguistic contents, thus having qualitative attributes — *semantic* and/or *pragmatic*, that informational sources deliver with some specific intent — *positive* or *negative*. Since drawing lines can be useful for clarity purposes, a broad distinction can be made between *quantitative* and *qualitative approaches* to information evaluation.[5]

Quantitative approaches look at information evaluation as a *fusion issue*. They are *numerical* and usually based on Zadeh's possibility theory [Zadeh 1978, Dubois & Prade 1990] or on probabilistic reasoning as in Bayesian analysis [Bayes 1763, Jeffreys 1939 1946, Pearl 2014] or as in Dempster-Shafer' theory of evidence [Dempster 1967, Shafer 1976].

Fusion operators based on possibility theory capture *degrees of uncertainty* about epistemic and doxastic attitudes through *possibility* and *necessity measurements*. Different families of fusion operators have been proposed for aggregating imperfect and hetereogenous intelligence data in that case [see Lesot *et al.* 2011 2013]. These operators are defined as *conjunctive* if the resultant score does not go beyond the minimum of initial values, *disjunctive* if the result is greater or equal to the maximum of the arguments, based on a *compromise* if the result is intermediary and, finally, *variable* if the score alternates depending on the initial arguments.

Bayesian analysis has been used to help officers better appreciate the credibility of intelligence messages by using probability degrees instead of verbal quantifiers [e.g. Zlotnick 1972, Fisk 1972, Schweitzer 1978, Schum 1987, Barbieri 2013, Blasch *et al.* 2013]. This proposal was motivated by empirical data showing inconsistencies in individual and collective interpretations of the existing ratings. Probabilities are used to define *prior* credibility ratings for message contents. Then, Bayesian rules are defined to *update* these prior probabilities depending on incoming information one can use to cross-check the message content. But incoming information can also concern the reliability of the message source. In

[5] Even though this distinction is artificial to some extent: quantitative approaches deal with qualitative notions (such as *credibility*, *reliability*, *truth*, *likelihood*, etc.), and most of the qualitative proposals have a quantitative flavour (through the elicitation of degrees for qualitative dimensions). However, this coarse delineation gives a clearer representation of the field of information evaluation [see Rogova & Nimier 2004, Capet & Delavallade 2013, Lesot & Revault d'Allonnes 2017, for surveys].

both cases, however, incoming information helps strike a balance on the prior probability distribution of ratings.

Fusion operators based on the Dempster-Shafer theory aim at helping officers compute the plausibility of some uncertain event, as well as the doxastic attitude to abide by, depending on data obtained from independent sources of various sensor types [e.g. Nimier & Appriou 1995, Nimier 2005, Cholvy 2004 2010, Pichon *et al.* 2012]. Degrees of credibility are expressed by *belief functions* rather than by Bayesian probability distributions. Probabilities encode evidence the officer has for particular messages. But these probabilities are assigned to sets of possible messages representing possible outcomes rather than to single and isolated messages.

Qualitative approaches are *symbolic* and based on *non-classical logics*. But depending on whether they put emphasis on the message content (for rating *credibility*) or on the message source (for rating *reliability*), these approaches split into two strands.

The first strand is *many-valued logic* and consists in giving assessments of the credibility of message contents in semantic framework built on more values than strict *truth* or *falsity*. Contents can receive extra discrete values (as in *three-valued logics*) or values on a continuum from 0 to 1 (as in *fuzzy logics*). In both cases, values are interpreted in an epistemic way: they capture agents' degrees of certainty and uncertainty on information quality. Contrary to existing scales for intelligence evaluation that are based on 6 levels of discrimination, many-valued settings provide message contents with more fine-grained credibility ratings. But the main challenge is to define semantic clauses for the conjunction of message contents that may receive conflicting semantic values. Various combination rules have been proposed in this endeavour [e.g. Akdag *et al.* 1992, Seridi & Akdag 2001, Revault d'Allonnes *et al.* 2007, Revault d'Allonnes & Lesot 2014].

The second strand of qualitative approaches is *modal logic*. Static epistemic operators have been defined to capture the *beliefs*, *desires* and *intentions* of informational sources [e.g. Demolombe & Lorini 2008, Herzig *et al.* 2010]. These operators are then combined to express *profiles* of sources depending on their informational pedigree, that is on their disposition to deliver true messages (*validity*)

and to be maximally informative when they do so (*completeness*) [e.g. Demolombe 2004, Cholvy 2013]. On a higher level, two *types* of unreliable sources have been characterized by Demolombe and Cholvy such as *falsifiers* (who sometimes report *false* information) and *misinformers* (who report *only* false information). Based on these syntactic definitions for sources' profiles and types, axiomatic principles and inference rules are combined to assess new contents and sources through adequate derivations. In this perspective, information evaluation is conceived as a way of ascribing *profiles* and *types* to the sources under investigation.

Quantitative and qualitative approaches have played a prominent role in devising innovative methods for information evaluation. Modal logic proposals are more recent but can be explained by the influence of quantitative approaches that already dealt with qualitative notions in need of higher specification (*credibility*, *reliability*, *likelihood*, *confidence*, etc.). From those observations, my *third goal* in this thesis is to propose a modal approach in numerical belief revision [Aucher 2004, van Ditmarsch 2005, van Ditmarsch & Labuschagne 2007], which aims to bridge the gap between qualitative and quantitative approaches on information evaluation, and to combine definitional and evaluative perspectives on deceptive attitudes. In Chapter 4, I define a plausibility setting in which *prior* distributions of credibility degrees are defined for intelligence messages based on the evidence officers have for, or against, their contents. That being done, distributions are *updated* depending on the reliability of their sources.

4. General Outline

This thesis is entitled "Lying, Deception and Strategic Omission: *Definition & Evaluation*" and aims at improving our understanding of some deceptive attitudes by combining the methodological approaches, theoretical aspects, and practical interests I have presented. The dissertation is composed of four chapters and one appendix. Each chapter is self-contained: the notions I intend to analyze in those chapters, as well as the experimental and logical materials I use to do so, are explained in due course and reminded when necessary. But although they are independent from each others, the chapters and appendix are integrated in a

deliberate order.

In Chapter 1 entitled *Two Definitions of Lying*, I combine conceptual resources with experimental protocols to see more clearly into the definition of lying. People's understanding of the verb "lie" seems to alternate between a subjective understanding insisting mostly on the speaker's intention-to-deceive, and an objective understanding insisting on the falsity of the agent's utterance as being a necessary condition for lying. A problematic and disputed case is when an agent intends to say something false, but ends up saying something actually true. Did the agent lie in those cases? Turri & Turri [2015] answered negatively, but Wiegmann *et al.* [2016] responded positively.

Based on two experiments,[6] I argue that the subjective account offers a better explanation of people's understanding of "lie" than the objective one. I test lay people's intuitions on various predicates derived from the root verb "lie" (viz. "*lied successfully*", "*lied*", "*liar*", etc.), the aim of which being to frame people's intuitions on both definitions, and to show that they favor the subjective definition in the critical condition. But although I do support this claim, I also discuss the mechanisms by which the objective definition can be retrieved from the subjective one when the speaker's utterance turns out to be false.

While the goal of Chapter 1 is to clarify the definition of lying, Chapter 2 looks at a more unusual case of deception in which the speaker deceives the addressee through a misleading default inference. This chapter entitled *The Surprise Deception Paradox* combines conceptual resources with formal modelling to analyze a paradox on *deception by omission* that was formulated by Smullyan [1978].

In this paradox, a sly speaker makes an announcement that triggers a default inference leading a vulnerable addressee to falsely believe that he will be deceived by some action (*deception by commission*) when in fact, he won't be deceived by any action (*deception by omission*). A paradox arises when the addressee starts reasoning about the deception he has been preyed to: on the one hand, if he wasn't deceived, then he didn't get what he expected (because he expected to be deceived after the speaker's announcement), and hence he was actually deceived. But on

[6] Supervised by Paul Égré and Brent Strickland at the Institut Jean Nicod in Paris.

the other hand, if he was deceived, then he exactly did get what he expected, and hence he was not "deceived". However, the paradox dissipates when the deceiver explains to the addressee the type of deception he has been preyed to. Then comes a state of surprise: the addressee is surprised to realize that he failed to expect the deception by omission that would actually happen.

This chapter uses dynamic belief revision theory [Baltag & Smets 2006, van Benthem 2007, Baltag & Smets 2008b] to investigate on those theoretical and paradoxical issues concerning *deception by omission*. I argue that the speaker's intended default can be reinterpreted through a belief update referred as "*radical upgrade*" in the literature. Modelling this rule helps understand the various stages of the speaker's deceptive plan as well as the addresse's resulting surprise. One interesting aspect of this paradox is that it raises similar issues to another famous epistemic puzzle known as the *Surprise Examination Paradox* [e.g. O'Connor 1948, Scriven 1951, Shaw 1958]. In both cases, surprise ensues because the protagonist fails to be a perfect reasoner and thus, to reach adequate conclusions about the world. But another interesting aspect in the surprise deception case is that the speaker's announcement of deception is literally true (*the speaker will deceive the addressee*), but is pragmatically misleading (*the addressee won't be deceived the way he assumes he will be*). For this reason, misleading defaults can be compared with classical instances of deception such as *double bluff strategies*, *presupposition faking* or *false implicatures*, but should be classified as non-standard cases of deception for being less pervasive than the latter.

While Chapters 1 and 2 investigate *standard* and *non-standard cases* of deception, Chapter 3 and 4 integrate these *definitional aspects* with the *evaluative perspective* on deception. Entitled *The Definition of Intelligence Messages*, Chapter 3 focuses on the scale that is commonly used for evaluating information in the intelligence domain [see STANAG-2511 2003, DIA-2 2010]. This scale is based on two evaluative dimensions for assessing message contents and message sources, namely the *credibility* of contents and the *reliability* of sources. But these dimensions have been critized for leading officers to conflate objective facts with subjective interpretations when they evaluate messages. In Chapter 3, I argue that this confusion follows from the subjective dimensions of credibility and reliability that leave in

the background objective dimensions of truth and honesty.

Chapter 3 sheds light on these objective dimensions to show that a taxonomy of messages can be derived which integrates *standard* and *non-standard cases* of deception. I propose a 3×3 matrix that partitions the scale into nine categories based on the *truth* of contents and *honesty* of sources. I distinguish three levels of *truth*: *true*, *false* or *indeterminate* when the status of the content is borderline between *true* and *false*. I also distinguish three levels of *honesty*: *honest*, *dishonest* and *imprecise* when sources are less cooperative than they should according to Gricean principles. By combining levels of truth and honesty, I then identify nine categories of messages in the descriptive space of the intelligence scale. That way, both standard and non-standard cases of deceptive attitudes can be recovered during information evaluation.

In Chapter 4 entitled *A Dynamic Procedure for Information Evaluation*, I combine conceptual and formal resources to devise a new procedure for intelligence evaluation. This procedure aims to comply with empirical findings researchers have made concerning the dimensions of credibility and reliability. In fact, experiments have shown that credibility and reliability are seen as *highly correlated*, and even *redundant*, by officers on the field [e.g. Baker *et al.* 1968, Samet 1975]. Credibility is perceived as the dominant dimension when evaluating messages, whereas reliability, whose role is secondary, helps mark a score on this dominant dimension. In that respect, the existing scale is ill-founded from a procedural perspective and should be revised to comply with empirical findings.

I propose a procedure in numerical belief revision [e.g. Aucher 2004, van Ditmarsch 2005, van Ditmarsch & Labuschagne 2007] that abides by the overwhelming importance of the credibility dimension and the ancillary role of the reliability dimension. The distinct credibility ratings are captured through *various degrees* expressing the conditional plausibility of the message under evaluation based on the evidence the officer has for, or against, its content being true. Reliability ratings are represented through *various update rules* that modify these degrees depending on the reliability of the source. This proposal is comprehensive in the sense that classical and non-classical cases of deceptive messages can be spotted in the course of evaluation.

In the Conclusion, I draw some perspectives for future work. Beside the main chapters in this thesis, the Appendix contains a paper I have written with my Advisor for the *Oxford Handbook of Lying* published in 2018 (J. Meibauer ed.). Entitled "Lying and Vagueness", our contribution is devoted to clarifying the link between standard cases of lies and non-standard cases of vague assertions. We start out by presenting the conceptual issues around linguistic vagueness through the distinction between main categories of vague language, namely *pragmatic imprecision* vs. *semantic indeterminacy*, as well as sub-categories of imprecision (*generality* vs. *approximation*) and of indeterminacy (*degree-vagueness* vs. *open-texture*). We then analyze deceptive cases of pragmatic imprecision based on *hiding information* as well as deceptive cases of semantic indeterminacy based on asserting *half-truths* (i.e. on making utterances whose truth status is borderline between truth and falsity). If hiding information is not lying (since the speaker does not breach the Gricean Maxim of Quality in that case), we argue that half-truths can be classified as lies if the speaker's utterance is true only under some very peculiar precisifications. Some of the content of this Appendix infuses the distinctions made in Chapter 3, regarding the evaluation of information.

Chapter 1

Two Definitions of Lying

1.1 Introduction

Since Plato fighting against sophistry (*Gorgias*) and Augustine' two essays on lying (*De Mendacio*) and against lying (*Contra Mendacio*), the topics of lying and deception have aroused continuous debates. They have spurred a lot of interest amongst (developmental) psychologists [e.g. Piaget 1932, Peterson *et al.* 1983, Bussey 1999], linguists and semanticists [e.g. Dynel 2011 2016, Meibauer 2014], experimental pragmaticists [e.g. Danziger 2010] and ethnologists. But philosophers, of course, have been on the front stage, especially *analytical epistemologists* who have tried to offer proper accounts of the (verbal) act of lying — that is adequate definitions of what it means *for someone to utter a lie to someone else*. In this definitional entreprise, most of them have offered *iff*-accounts of lying, sometimes called "checklist definitions" [e.g. Carson 2006, Fallis 2009b 2010, Mahon 2015]. The goal of such definitions is to provide a *set of necessary and sufficient conditions* for lying, the aim of which being to capture people's *theoretical* intuitions on the notion.

By doing so, however, analytic epistemologists have usually left aside people's *pre*-theoretical intuitions on the notion, — either because they consider that the concept of lying can be fully determined by pure inquiry,[1] or that *pre*-theoretical intuitions are strictly reducible to *theoretical* ones and don't need to be investigated

[1] As it is thought to be the case for abstract concepts such as "*set*", "*relation*", "*knowledge*", etc.

on their own. As a result, these epistemologists have generally assumed that lying is not sensitive to context and can be entirely manipulated by thought experiments. Whether intentional or not, this aprioristic bias necessarily sets apart extra dimensions (such as *blame* and *conditions of success* for the speaker, *costs* for the hearer, etc.) that strongly matter for defining lies since lying is a social action. Indeed, lying typically involves *two social agents*, one of them — *the speaker* — having an intention to deceive the other — *the hearer* — by uttering a proposition he or she believes to be false. As a consequence, lying is a multidimensional concept which is sensitive to external constraints and cannot be grasped by thought experiments only.

For this reason, some epistemologists have joined the definitional debate through experimental means. Their ambition is to try to assess, complete or even challenge the aprioristic definitions (analytic) epistemologists have put forward for centuries. By asking questions about scenarios that involve a dishonest speaker, the aim is to evaluate people's commonsensical intuitions about lying to determine whether these intuitions actually match with the aprioristic definition. We can distinguish two different challenging strategies.

The first strategy rejects the *iff* definitional project as being structurally inadequate and illusory. Amongst those who consider that *iff* accounts are too rigid to capture all the features of lying are *prototypical* semanticists [since Fillmore 1975]. For instance, Coleman & Kay [1981] hold that a proper definition of lying should only *highlight* the definitional *traits* that are *constitutive* of lying, namely the features called "*typical*" which materialize when using the verb "*to lie*". These prototypical definitions are more flexible than the previous *iff* ones. They are less rigid (since definitional constituents are no longer necessary and/or sufficient), and they introduce a semantic dimension that makes them more sensitive to contextual variations.

The second strategy proposes to adapt the *iff* project by adding (or removing) necessary conditions in order to reach adequate sufficiency. In this endeavour, some experiments have been conducted by John and Angelo Turri in empirical epistemology. In 2015, they collected data to show that contrary to the aprioristic credo which they dub the "*subjective account of lying*", uttering lies requires to make

objectively false statements and not only *believed-to-be-false* statements. Accordingly, the subjective account should be reinforced by a so-called "*falsity condition*". This is the "*objective account of lying*". Immediately, though, this claim was challenged by Wiegmann *et al.* [2016] who argued that the falsity condition was superfluous, and that the subjective definition was satisfactory as it stands.

A compromise strategy could be to integrate the *iff* strategy to the *prototypical* strategy. The concept of lying would be *many-sided* and more precisely, *two-sided* between the subjective and objective definitions. On that view, lying is pragmatically ambiguous and may alternate between *two acceptable definitions* of the same notion. As a matter of fact, the Turris themselves recognized that their results are consistent with lying being *many-sided*: "*An alternative interpretation of our findings is that there are multiple senses of "lie," some of which require objective falsity and some of which do not*" [166]. But they dismiss this *many-sense interpretation* as being an easy way out from offering a straightforward account of lying based on the *one-sense interpretation*. So, preferably to any compromise strategy, they first attempted to give evidence for the objective view of lying.

In this chapter, I argue that the *subjective account* offers a better explanation of people's understanding of "lie" than the *objective account*. I present the theoretical features to both definitions, and provide empirical data supporting the subjective view. More specifically, I test people's intuitions on various predicates derived from the root verb "lie": "*lied successfully*", "*lied*", "*liar*", etc. My strategy is to elicit people's intuitions in order to show that they favor the subjective definition in critical condition. Then, I also discuss the mechanisms from which the objective definition can be retrieved from the subjective one in particular circumstances.

My chapter is structured as follows. In section 1.2, I briefly review the epistemological literature behind the aprioristic accounts. I first present the *subjective definition* which is known as the *traditional* (or *standard*) *definition* on lying (subsection 1.2.1). I then point out how epistemologists have been challenging its limits concerning the necessity and sufficiency of definitional conditions (subsection 1.2.2). I particularly insist on the *objective account* that requires a *falsity condition* as mandatory.

Section 1.3 presents the main results of the experiments I have mentioned above [viz. Turri & Turri 2015, Wiegmann *et al.* 2016]. I start out by reviewing the Turri & Turri' plea for the objective view (subsection 1.3.1), and then present Wiegmann & al.' response to support the subjective view (subsection 1.3.2). Reviewing those results will help me sketch the main issues involved in the empirical debate about lying. It will also give some preliminary insight to distinguish Wiegmann & al.' strategy from my own.

My purpose in section 1.4 is to defend the *subjective hypothesis* by testing different predicates derived from the root verb "to lie" (subsection 1.4.1). I first present a pilot experiment supporting this subjective hypothesis. Basically, I test two distinct predicates (viz. "is a liar" and "has lied successfully") which respectively encode the subjective vs. objective definitions due to their dispositional vs. episodic forms (subsection 1.4.2). As it turns out, these data are not replicated for more basic predicates, namely "is a liar" and "has lied" (subsection 1.4.3). However, a post hoc analysis of the results shows that "has lied" behaves exactly like "is a liar" in the critical condition (subsection 1.4.4). Both are tracking identical properties (viz. the speaker's intention to deceive). This confirms that people endorse the subjective definition as a core definition of lying.

In section 1.5, I discuss the subjective definition of lying. Though this definition accounts for people's handling of "lie", some points need clarification to make it fully operative. In particular, one should explain why the falsity of the speaker's utterance is not mandatory for lying but tends to increase the ascription of the subjective definition. My interpretation is that falsity generates a Knobe effect [viz. Knobe 2003ab 2006] which naturally increases people's ascription of the predicate "has lied". By clarifying this issue, I hope to secure a stronger basis for the subjective account when falsity enters the scene.

1.2 Defining Lying: standard account and theoretical challenges

1.2.1 The Traditional Definition: the Subjective View

Definitions of lying have a long history. Augustine [395] provided the first account of the notion by writting that one "*may say a true thing and yet lie, if he or she thinks it to be false and utters it for true, although in reality it be so as he or she utters it*". According to Augustine, lying only requires for speakers to utter contents that they *believe to be false*. Whether contents happen to be *objectively true* or *objectively false* does not matter in that respect. In the same way, having some *intention to deceive* and directing this intention to some *defined addressee* are not required for lying *per se*. But since Augustine, many other definitions of lying have been put forward by philosophers. For instance, Isenberg [1965] writes: "*A lie is a statement made by one who does not believe it with the intention that someone else shall be led to believe it*", or by Primoratz [1984]: "*Lying is making a statement believed to be false, with the intention of getting another to accept it as true*". Again, these definitions do not specify that the utterance should be directed towards a *specific addressee* to qualify as a lie. In addition to that, all these philosophers agree that a lie does not need to be *false* but only *believed to be false* and *intended to be believed as true*. [see also Fried 1978, Williams 1985, Chisholm & Feehan 1977]

The standard definition of lying is more precise, however. Mahon [2015] frames it in the following way: "*To make a believed-false statement to another person with the intention that the other person believe that statement to be true*". From this account, an *iff*-definition can be given that lists all the necessary and (mutually) sufficient conditions for lying in the traditional sense. This definition is generally assumed to be both descriptive and normative of what "lying" means. Here is the definition:

Subjective Def. A speaker X lies to a hearer Y on a proposition p *iff*

(1) X believes that p is false.

(2) X intends Y to believe that p is true.

(3) X tells Y that p.

In this definition, condition **(1)** is usually called the "*untruthfulness clause*" and explains why the standard definition is called "*the Subjective View on Lying*". The speaker's belief is crucial here: this is the perspective from which X starts deploying a deceptive strategy towards Y. In fact, X determines what he or she aims Y to believe about p (here, that p is true) by first considering what he or she actually believes about p (here, that p is false). Finally, X utters that "p" to fulfill his or her intention towards Y.

Note that the *untruthfulness clause* is stronger than the clause whereby the speaker simply disbelieves proposition p. In this latter case, X has a weaker propositional attitude towards p, whether he or she is skeptic or doubtful that p, or absolutely unaware of the fact that p. Then, it would be counterintuitive to say that X lies to Y on p in that case since there is *no belief* that *not-p* from which X could deceive Y by making him or her believe that p. One last remark about the untruthfulness condition is the following: this subjective clause differs from the "*falsity clause*" — the latter being objective: p is *objectively false*. According to the subjective view, p can be objectively true so long as X *believes it to be false*.

Condition **(2)** is called the "*intention-to-deceive clause*" and states that X *sees to it that* Y believe a proposition p that X believes to be false. This intention is a purely mental state and must be distinguished from the effective *attempt* to make Y believe that p. In fact, X's intention to deceive can fail to materialize in two different ways. First, X can decide to leave plans unexecuted. For instance, X makes no attempt to realize his or her intention by not trying to lie after all. Second, even if X *does* try to lie, X can fail to achieve this attempt in case he or she either does not believe that *not-p* and/or does not tell that p to Y (but to some distinct addressee for instance).

Finally, condition **(3)** is named the "*addressed statement clause*": X has to utter that "p" to a specific addressee Y. Making a statement, more precisely uttering a declarative sentence to Y, is then a necessary condition for lying. As a consequence, so-called "*lies of omission*" (which are "lies" with *no assertions* involved) are not *lies* in the colloquial sense. Lies are always *lies of commission* in that sense. Furthermore, one cannot lie by cursing, by asking a question, by making an excla-

mation or by giving an order [see Carson 2010, for many other cases] since lying necessarily implies to assert a *declarative sentence*. Also, a lie should be addressed to a *human hearer* or to a group of *human hearers* (a crowd, a targeted audience, etc.), namely to cognitive human beings who are able to understand (and not only process) the utterance at stake. In this regard, one cannot lie to other sentient beings like animals or plants, and to non-living entities like robots or minerals [contra Chisholm & Feehan 1977].

With respect to condition **(3)**, one may wonder which norm of assertion is violated when we utter a lie.[2] But answering to this question first requires to agree on which account of assertion is the correct one, that is on whether *knowledge* [e.g. Williamson 2000, DeRose 2002, Hawthorne 2003, Engel 2008, Schaffer 2008, Turri 2010], *belief* [Bach 2008], *rational credibility* [Douven 2006 2009], *reasonable belief* [Kvanvig 2009 2011, Lackey 2007] or *truth* [Weiner 2007, MacFarlane 2014] is the correct norm of assertion. According to the subjective definition of lying, however, condition **(1)** implies that the speaker is insincere. In other words, the speaker believes that the content he or she is lying about is *false*, no matter whether this content turns out to be *true* or *false*. Believing that this content is false implies that the speaker does not believe it and thus, that he or she does not know it. Then, norms based on knowledge and belief are breached when one utters a lie in the subjective sense. If the utterance also turns out to be false, norms based on truth are also violated in that case.

Compared to other verbs such as "*to deceive*" and "*to mislead*", "*to lie*" is not usually considered as an *achievement* or *success verb* [e.g. Ryle 2009, Carson 2010]. Since *to deceive* (and *to mislead*) can be defined (in the broad sense) as *causing someone to hold a false belief* — whether it is intentional or not, one *cannot attempt to deceive* (or to *mislead*) without *succeeding at deceiving* (or *misleading*) — that is at making someone else believe a falsehood. On the contrary, it is generally admitted that one *can attempt to lie* without *succeeding at lying*. Two causes can be invoked for such lie failures.

One cause of lie failure is if one fails regarding conditions **(1)**-**(2)**-**(3)** above.

[2] See in particular Lackey [2007], Engel [2008 2016] and McKinnon [2016] on the interplay between lies and norms of assertion.

First, reaching untruthfulness is not always straightforward. Even though X wants to believe that p is false, X can fail to make it because he or she involuntarily believes that p is true, or both believes that p is true *and* that p is false with inconsistency. Second, X can fail to intend to deceive because he or she does not really entertain such a malevolent intention, or because they have mixed intentions towards their intended addressee. Finally, X can fail uttering that p due to mispronunciation (for instance), and to address it to someone else due to a defective communication channel.

But the main cause invoked for lie failure is the following: X fails to lie because X fails to "deceive". That is, X does not succeed in making Y *trust* him into believing a *falsehood*. Either Y does not trust X and then does not come to believe that p is true; or Y does trust X but, finally, p happens to be true and not misleading after all. In this second sense of lie failure, X does not manage to lie because X does not reach trust and/or objective falsity.

1.2.2 Main Conceptual Challenges to the Traditional Definition

The standard subjective view, though being standard, has also been strongly debated by epistemologists over the years. The traditional account is considered to be either *too narrow* or *too broad*. In the first case, epistemologists have claimed that conditions taken as necessary by predecessors, were not so in reality. In the second case, they have claimed that the three conditions listed above are not jointly sufficient for capturing the essence of lying.

Concerning necessity, some hold that the "*untruthfulness clause*" is not always required for lying. For instance, Carson [2006 2010] and Shiffrin [2014] are *permissive* about untruthfulness. Carson argues that the uttered content does not need to be *believed-as-false*: it is sufficient that the content is *not believed to be true* or simply *believed to be probably false*. For Shiffrin, even a statement that is *not believed to be either true or false* (as it is the case in "*agnoticism*") can be a lie if the speaker also has some intention to deceive the addressee. In the same way, Davidson [1985] and Barnes [1994] have made even *more radical* proposals: they advocate that a *truthful statement* that is made with an intention to deceive also counts as a lie. Moreover, Adler [1997], Dynel [2011] and Meibauer [2011] share the Gricean view that a

truthful statement which *conversationally implicates a believed-to-be-false statement* is a lie.

Concerning necessity again, others have proposed to slightly modify the "*intention-to-deceive clause*" [e.g. Chisholm & Feehan 1977, Williams 2002]: the speaker does not want the hearer to believe a statement he himself believes to be false but he or she aims for the hearer to believe that he or she believes the statement to be true. Consequently, these authors propose to substitute clause **(2')** to clause **(2)** in **Subjective Def**:

(2)' X intends Y to believe that X believes that p is true.

Mahon [2015] goes a step further: he proposes not to substitute **(2')** to **(2)** but to supplement **(2)** with **(2')**. But the more radical challengers have been [Sorensen 2007], Fallis [2009b] and Stokke [2013]. They argue that the intention-to-deceive clause is *not necessary for lying* since there are borderline cases such as "*bald-faced lies*" in which the speaker makes an untruthful assertion but lacks any intention to deceive. A bald-faced liar is effectively lying according to them.[3]

The objections made against the necessity of the "*addressed statement condition*" have been twofold. First, some have defended that *making a statement is not required for lying stricto sensu* [e.g. Vrij 2000, Smith 2007]. Withholding information with an intent to deceive *is* lying in that sense. Then, lies of omission are lies in the same way lies of commission are [e.g. Ekman 1985, Scott 2006]. Second, although a deceptive statement is made, there is no need to direct this statement to some specific addressee to count it as a lie [e.g. Shibles 1987, Griffiths 2010]. The lie may even be directed to *no addressee* as a matter of fact. From this perspective, making a believed-false statement while intending to deceive an unspecified addressee is sufficient for lying *per se*.

Another major source of disagreement over the subjective definition is that condi-

[3] Think for instance of muddy-faced children who are making the untruthful assertion to their mother that they have not been playing in the mud. Are they lying or not? In case they *do* lie to their mother and given that the mother can easily read on their faces that they do, are the children having any intention to deceive her?

tions **(1)-(2)-(3)** are not mutually sufficient for lying. For instance, Simpson [1992], Frankfurt [1999] and Faulkner [2007] have successively claimed that condition **(2)** must be reinforced by the clause **(2')** mentioned above. The main point about sufficiency is that a "*falsity condition*" should be added to the standard account: *p has to be objectively false* [e.g. Krishna 1961, Grotius 2005, Grimaltos & Rosell 2013]. In other words, the adequate view on lying is *more surely* objective (*p* is *objectively* false) than purely subjective (*p* is *subjectively* believed-to-be-false). Others, like Carson [2006 2010] and Saul [2012], have argued that this extended definition is *not the only right* definition of lying. This definition is an other account of lying that should be considered *on top of* the traditional one.

Let us give an *iff*-definition listing all necessary and sufficient conditions for lying in the objective sense:

Objective Def. A speaker *X* lies to a hearer *Y* on a proposition *p iff*

(1) *X* believes that *p* is false.

(2) *X* intends *Y* to believe that *p* is true.

(3) *X* tells *Y* that *p*.

(4) *p* is false.

As I said, this objective definition is usually considered too strong: *most analytical epistemologists reject the falsity condition as unnecessary*. But empirically, this definition is supported since one of the first experimental studies on lying. In *Prototypical Semantics: The English Word LIE* published in 1981, Coleman & Kay asked participants to read eight different stories and then, asked them to rate whether a statement made by the main character was a lie or not (On scale: 1 = 'Very sure non-lie' - 7 = 'Very sure lie'). One of the vignettes involved a statement that (i) was believed to be false, (ii) was intended to deceive and (iii) *was objectively false* (Story I), while an other vignette involved a statement that (i) was believed to be false, (ii) was intended to deceive but (iii) *was objectively true* (Story IV). Although both statements were believed to be false and were intended to deceive, the first statement — which was objectively false, was considered *more of a lie* than the second statement which was objectively true (6,96 contra 5,16).

More recently, the objective view has found new ardent advocates in Turri & Turri' empirical work. I will now present the experimental protocol they have promoted to support the "*falsity condition*". Then I will review Wiegmann *et al.* [2016]' response in favor of the subjective account.

1.3 The Subjective vs. Objective View on Lying: previous experimental results

1.3.1 Turri & Turri' Plea for Falsity: the Objective View

In *The Truth about Lying* published in 2015, Angelo and John Turri attempted to provide empirical data to support the objective definition of lying. They aimed to show that folk intuitions required that the speaker uttered a statement that was *objectively false* in order to lie in the colloquial sense. To this end, they ran three different experiments involving a slight variation on the same story.

In the *first experiment* (and the subsequent *replication*), they considered the following scenario:

> *"Jacob's friend Mary recently posted information on the internet that will alert the public to serious corruption in the government. Soon some federal agents visit Jacob and ask where Mary is. Jacob realizes that they want to detain her for what she did. Jacob thinks that Mary is at her brother's house. He is very concerned about Mary and does not want her to be detained. Jacob tells the agents, "She is at* [*the grocery store/her brother's house*] *on the corner." However, unbeknownst to Jacob, Mary is at the grocery store, where the agents find and arrest her."*

From this story, Turri & Turri built a 2×2 between-subject design, based on 2 factors (Intention/Predicate), with 2 levels of the factor Intention ("Honest"/"Dishonest") and 2 levels of the factor Predicate ("Lied"/"Told the truth"). In dishonest conditions, Jacob told the authorities something *true* (viz. that "Mary is at the grocery store") despite his *malevolent intention* to deceive the authorities. In honest conditions, he told the agents something *false* (viz. that "Mary is at her brother's

house") despite his *benevolent intention* to inform them. Both in honest and dishonest conditions, some participants were asked the test question "Did Jacob lie about Mary's location?" (Yes/No) while others had to answer the following one: "Did Jacob tell the truth about Mary's location?" (Yes/No). In the critical case [Dishonest/True] in which Jacob said something he believed to be false but that was *objectively true*, 78,3 % of subjects answered that he *did lie*. This result was then replicated with graded measures (On scale: 1 = "Strongly disagree" - 7 = "Strongly agree") instead of dichotomous ones. Participants were asked to rate the statements "Jacob lied about Mary's location" and "Jacob told the truth about Mary's location". The preceding results were confirmed: subjects agreed that Jacob lied at 5,59/7 (see Figure 1.1).

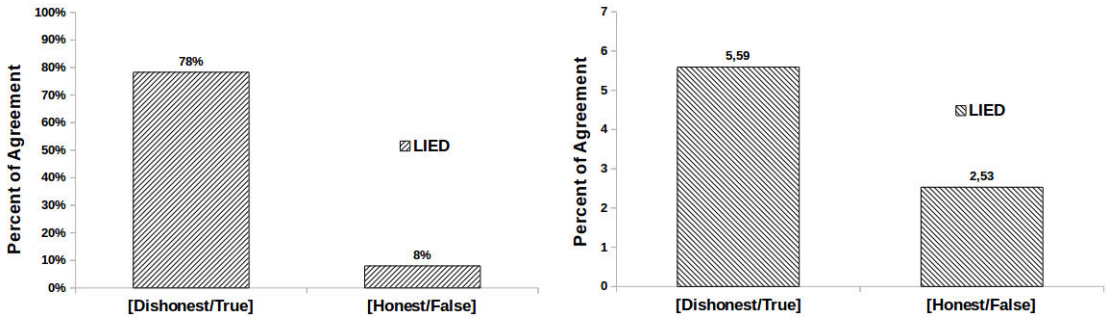


Figure 1.1: The Turris' *first experiment* (left) and *replication* (right).

All these data appear to support the subjective view on lying: "*A dishonest assertion is a lie even if it turns out to be objectively true*". But the Turris claimed that interpreting the data this way is biased for two distinct reasons (even though they did not test those hypotheses afterwards). First, the Turris consider that participants may have taken the perspective of the speaker for rating his utterance as a lie. In other words, people would assess how the protagonist *would rate his own attitude* if he or she was Jacob: "*If the agent [Jacob] makes a dishonest assertion, then he thinks that he is lying*" [163] and his assertion must be rated as a straightforward lie. Second, the Turris think that the subjects did not really answer to the test question "Did Jacob lie about Mary's location?" but used it "*as an opportunity to express their approval or disapproval of the protagonist's conduct*" [163]. This means that they reinterpreted the test question in the following way "Is Jacob's conduct morally good or wrong?" and the collected data were thus purely "*artifactual*".

For those two reasons, Turri & Turri decided to make a *second experiment* offering higher flexibility to respondents. They esteemed that the biases they identified (*perspective-taking*, *blame-opportunity*) were created by the formulations of the test question ("Did Jacob lie about Mary's location?") and of the response option ("Jacob lied about Mary's location"). The Turris make the following point: intending to lie is not the same thing as lying *per se*, viz. as succeeding at lying. Even though both cases require to assert a believed-false statement while intending to deceive, one *only succeeds at lying* if one actually asserts a *false statement*. But the formulations at stake did not make this kind of distinction and relied on ambiguity. To decide whether Jacob lied in the dishonest case, participants could only evaluate Jacob's intention and due to perspective-taking and blame-considerations, were led to answer positively. To offset these unwanted effects, Turri & Turri slightly modified the initial scenario in the following way:

> "*Jacob's friend Mary recently posted information on the internet that will alert the public to serious government corruption. Soon some federal agents visit Jacob and ask where Mary is, in order to detain her. Jacob thinks that Mary is at* [*the grocery store/her brother's house*]*, so he tells the agents, "She is at* [*the grocery store/her brother's house*]*." In fact, Mary is at the grocery store.*"

From this modified version, they built a 2×2 between-subject design, based on 2 factors (Intention/Content), with 2 levels of the factor Intention ("Honest"/-"Dishonest") and 2 levels of the factor Content ("True"/"False"). Participants were randomly assigned to four kinds of vignettes: [Dishonest/True], [Dishonest/False], [Honest/True] and [Honest/False]. Note that Jacob could now utter a statement that was either true or false, whether or not he had a benevolent or malevolent intention towards the authorities. Both in honest and dishonest conditions, participants were then asked to choose between four different options describing Jacob's attitude when he was talking to the authorities:

(1) He tried to tell the truth and succeeded in telling the truth;

(2) He tried to tell the truth but failed to tell the truth;

(3) He tried to tell a lie but failed to tell a lie;

(4) He tried to tell a lie and succeeded in telling a lie.

By using these formulations that separated Jacob's utterance between a *trying-part* ("tried to tell a lie") and a *result-part* ("succeeded in telling a lie/failed to tell a lie"), the Turris wanted subjects to rate Jacob's assertion (as a *lie* or *not*) rather than his moral conduct (as *blameworthy* or *not*). More specifically, they wanted participants to leave aside Jacob's perspective by asking if he actually succeeded at lying or failed. In the dishonest conditions in which Jacob intended to deceive the agents, 88 % of the participants agreed that Jacob tried to lie but failed in the [Dishonest/True] case, while 95 % of the participants agreed that Jacob tried to lie and succeeded at lying in the [Dishonest/False] case (see Figure 1.2). The Turris concluded that intending to deceive and saying something you believed to be false but that turned out to be objectively true was not lying *per se*.

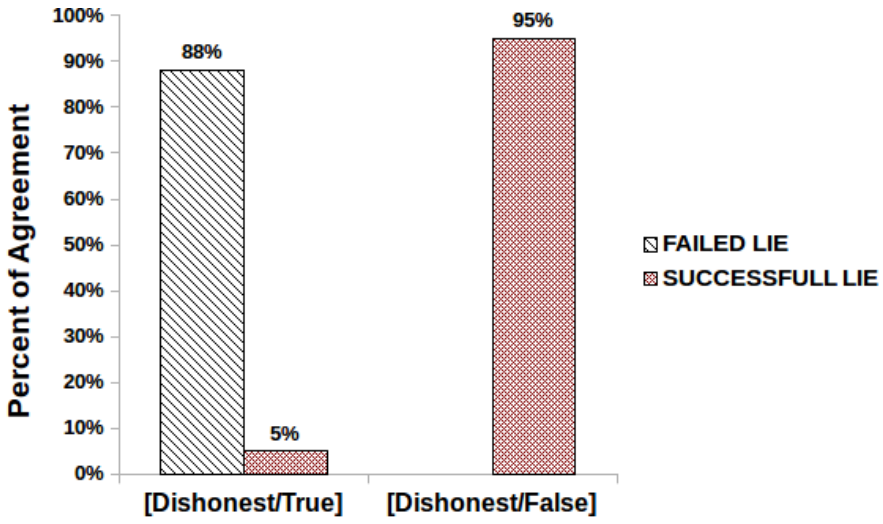


Figure 1.2: The Turris' *second experiment*.

According to the Turris, these results defeat the *subjective view* since, by checking option (4), "*only 5 % of the participants (2 of 41) classified a dishonest but true assertion as a lie*" [165]. On the contrary, the Turris claimed that the data strongly supported the *objective account*: only statements that were both dishonest and false were judged as being "*lies*" stricto sensu. Besides that, they argued that it could be held that "*a failed lie is still a lie, just as a failed attempt is still an attempt*" [165]. In other words, when 88 % of the participants agreed that a dishonest but true assertion

was a failed lie, they may have misunderstood what a "failed lie" is. They may have thought that Jacob *still lied* (since he committed the act of lying) but that *he failed to say something objectively false*.

As a consequence, the Turris conducted a *third experiment* based on the previous one. They kept the second scenario and the 2×2 Intention vs. Content between-subject design from the *second experiment*. But instead of asking people to choose between four options ("Which better describes Jacob?"), they asked them to choose between only two:

(1) He tried to lie and actually did lie;

(2) He tried to lie but only thinks he lied.

They also put a prime before presenting the test options: "What Jacob said is objectively ——." (True/False). By doing so, they expected people to pay more attention to the difference between *what Jacob subjectively believed* and *what was objectively the case*, — the aim of such a strategy being to offset any potential perpective-taking. The results confirmed those of the *second experiment* in the (critical) dishonest conditions (see Figure 1.3): only 10 % of the participants (4 out of 40) qualified a [Dishonest/True] assertion as a lie while 90 % of the participants (36 out of 40) qualified a [Dishonest/False] assertion as a lie.

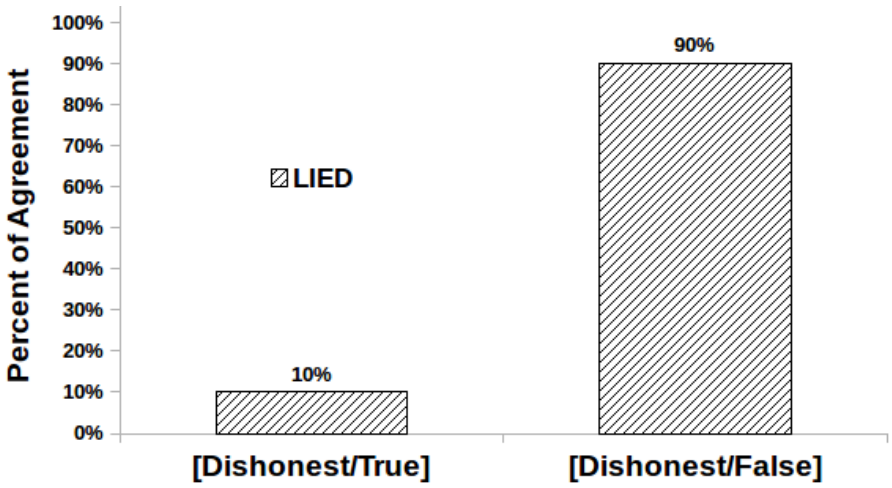


Figure 1.3: The Turris' *third experiment*.

From all those experiments, the Turris concluded that *only dishonest false statements count as lies stricto sensu*. Contrary to the subjective view, *dishonest but true assertions are not lies* according to them: they are only attempts to lie but not real lies. As a result, the adequate view on lying is the objective definition which requires a falsity condition as a supplement. Nevertheless, this radical conclusion has been immediately challenged by three psychologists a year later. Wiegmann, Samland and Waldmann explained that the Turris' plea for the objective view was not conclusive at all. The Turris' main arguments to dismiss the subjective view (*perspective-taking*, *blame-opportunity*, etc.) were not compelling and could be easily blocked. Moreover, Wiegmann & al. claimed that the Turris' protocols were defective since they generated order and framing effects favoring the objective view over the subjective one. I now present Wiegmann & al.' defense of the subjective definition against the Turris' attacks.

1.3.2 Wiegmann & al.' Defense of the Subjective View

In *Lying Despite Telling the Truth* published in 2016, Wiegmann, Samland and Waldmann answer to the Turris by taking into account conversational and experimental pragmatics [e.g. Grice 1989, Noveck & Reboul 2008]. Their main hypothesis is that participants in the Turris' protocols *did support the subjective view* but "*were led by the two-part response options to interpret the test question being not merely about whether Jacob lied but about whether what Jacob said was objectively false*" [37-38]. Splitting the options between a trying-part and a result-part being pragmatically misleading, Wiegmann & al. proposed to come back to one-part response options. They aimed to prove that dishonest but true assertions were lies in the same sense dishonest and false assertions were. More generally, they wanted to show that according to folk intuitions, the standard account is a more adequate definition of lying than the revised one. To this end, they used Turri & Turri's following *original scenario* as a basis to build four different experiments — the last protocol being a comparison between the verbs "to lie" and "to promise" to address the Turris' framing effect(s). Note that this *original scenario* corresponds to the Turris' [Dishonest/True] condition in which Jacob intended to deceive the authorities but unexpectedly turned out to tell them the truth (i.e. that Mary was at the grocery store):

> *"Jacob's friend Mary recently posted information on the internet that will alert the public to serious government corruption. Soon some federal agents visit Jacob and ask where Mary is in order to detain her. Jacob thinks that Mary is at her brother's house, so he tells the agents, "She is at the grocery store." In fact, Mary is at the grocery store."*

Let us first recall that the Turris dismissed theirs findings as *artifactual* in their first experiment and replication. In particular, they held that people may have taken the test questions and options as *blame-opportunities* to "*express their approval or disapproval of the protagonist's conduct*". "*Of the available options, the best way to register disapproval is to say that the protagonist lied*" when he or she utters a believed-false statement while intending to deceive, even though the statement happens to be objectively true [Turri & Turri 2015, 163]. Nevertheless, the Turris did not provide any empirical data to account for this explanation. So Wiegmann & al.' planned their *first experiment* to fill this gap.

They designed a 2×2 between-subject design, based on 2 factors (Scenario/Predicate), with 2 levels of the factor Scenario ("Original"/"Dictator") and 2 levels of the factor Predicate ("Lie"/"Blame"). Subjects were presented with either the *original scenario* or the *dictator scenario*, the latter being a modified version of the story taking place in a dictatorship:

> *"Jacob's friend Mary recently posted information on the internet that will alert the public to serious government corruption. Soon some federal agents visit Jacob and ask where Mary is in order to detain her. Jacob and Mary live in a country ruled by a merciless dictator who commits crimes against humanity. Mary has reported these crimes and therefore faces torture if she gets caught. Jacob thinks that Mary is at her brother's house, so he tells the agents, "She is at the grocery store." In fact, Mary is at the grocery store."*

Then, participants were asked *either* "Did Jacob lie about Mary's location?" (Yes/No) *or* "Is Jacob (morally) to be blamed for trying to deceive the agents?" (Yes/No). The data are summed up in Figure 1.4. Concerning *lying*, results were comparatively the same in the original and dictator versions: 85 % vs. 84 % of

"Yes" responses. Concerning *blame*, significantly more participants judged that Jacob should be blamed in the original scenario than in the dictator one: 33% vs. 16 % of "Yes". Thus Wiegmann & al. concluded that the Turris' blame argument for dismissing the subjective view was irrelevant.

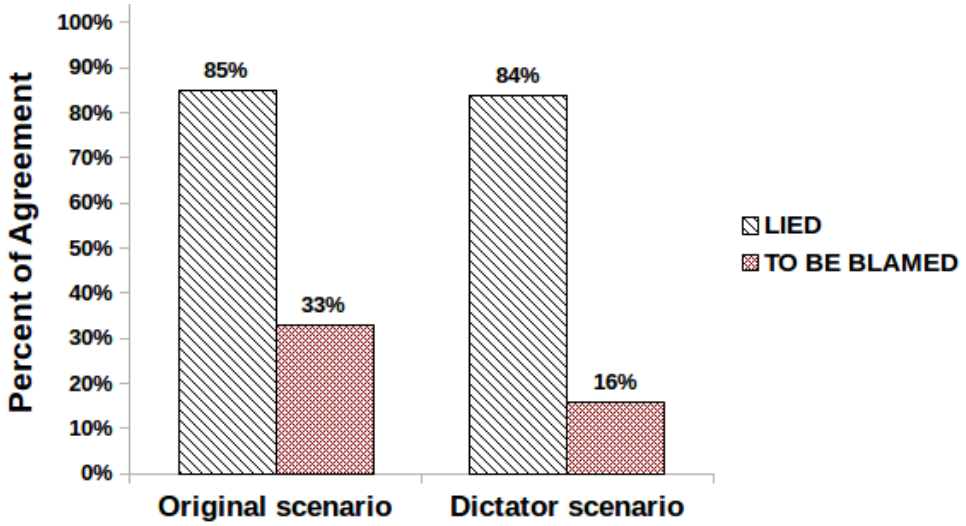


Figure 1.4: Wiegmann & al.' *first experiment*.

Wiegmann & al. then ran a *second experiment* to address the Turri & Turri's claim that *perpective-taking* benefited to the subjective account over the objective account. They devised a 2×2 between-subject design that was similar to the design of the *first experiment*, except that they conceived an alternative version of the original scenario by adding an extra component to the test options — the aim of which being to emphasize that Jacob's statement was *objectively* true when he tried to lie. If participants still agreed that Jacob lied when his utterance turned out to be true, Wiegmann & al. could conclude that the Turris' argument on *perspective-taking* was also disproved. The material used in this experiment consisted of the preceding *original scenario* but participants were now assigned to test options that were put either *without ending* or *with an additional ending* (see plus sign and square brackets not displayed on the screen):

(1) He tried to tell a lie but failed to tell a lie

+[because what he said turned out to be true];

(2) He tried to tell a lie and succeeded in telling a lie

+[although what he said turned out to be true].

When the *original scenario* was followed by test options *without ending*, the Turris' second experiment was replicated (see Figure 1.5): only 22 % of subjects said that Jacob succeeded at lying while 78 % said that he did not. But when the *original scenario* was followed by test options *with an additional ending*, the Turris' claim was proven false (see Figure 1.5 again): 72 % of the participants judged that Jacob succeeded at lying when his deceptive utterance turns out to be true while 28 % judged that he did not. Accordingly, *perspective-taking* could not be taken as a confound to invalidate the subjective view of lying.

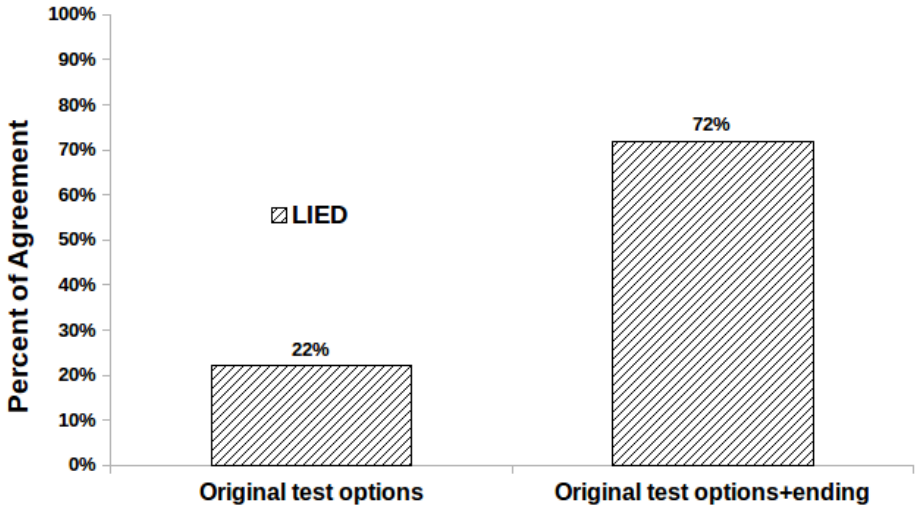


Figure 1.5: Wiegmann & al.' *second experiment*.

Wiegmann & al. criticized the Turris' prejudices against the subjective view but they also tried to explain why their experiments favored the objective view over the subjective one. They ran *two further experiments* to point out that the objective account benefited from *order effects* and *framing effects* in the Turris' protocols.

As a reminder, the Turris conducted a *third experiment* because they thought that the data obtained from rephrasing the response options (*second experiment*) could be seen as "artifactual" (due to *perspective-taking*). They modified the two-response options to put more emphasis on the difference between subjective

truth (viz. Jacob's belief about Mary's location) and objective truth (viz. Mary's actual location): (1) He tried to lie and actually did lie vs. (2) He tried to lie but only thinks he lied. Besides that, the Turris put a prime before the test options ("What Jacob said is objectively ——." (True/False)) to force participants to pay more attention to the subjective/objective contrast. Nevertheless, Wiegmann & al. argued that this prime actually disadvantaged the subjective view of lying by appearing as a *demand characteristic "alerting subjects that (...) objective truth or falsity [was] particularly relevant for answering the test question"* [40]. More precisely, they highly suspected that this prime led participants to reinterpret the result-part ("actually did lie", "only thinks he lied") as a request to assess objective falsity. But since falsity was missing in the Turris' [Dishonest/True] condition, subjects were necessarily led to answer that Jacob *did not lie* in this case.

Thus, Wiegmann & al.' *third experiment* aimed at testing this potential *order effect*. They devised a 2 (Prime: "Present" vs. "Absent") × 2 (Ending: "Added" vs. "Not added") between-subject design based on their *original scenario* (that matched with the Turris' [Dishonest/True] condition). In the "present" conditions, subjects were asked the same question as before: "What Jacob said is objectively ——." (True/False). In the "absent" conditions, they were not asked any question. Then, in all the conditions, participants were assigned to the test question "Which better describes Jacob?" and provided with the following response options put either *without ending* or *with an additional ending*:

(1) He tried to lie and actually did lie

+[although what he said turned out to be true];

(2) He tried to lie but only thinks he lied

+[because what he said turned out to be true].

Wiegmann & al.' prediction was confirmed (see Figure 1.6). When a prime was added before the *original scenario* given without ending, the Turris' results were replicated: 74 % of subjects agreed that Jacob did not lie while only 26 %

agreed that he did. But the results reversed when a *prime* was added before the *original scenario* now followed by an additional ending on objective truth: 58 % of the participants agreed that Jacob did lie while 42 % agreed that he did not. The results were even more significant when *no prime* was added to the ending case: 81 % of people said that Jacob lied while 19 % said that he did not. As a consequence, adding a prime on objective truth created an *order effect* that strongly benefited to the objective view.

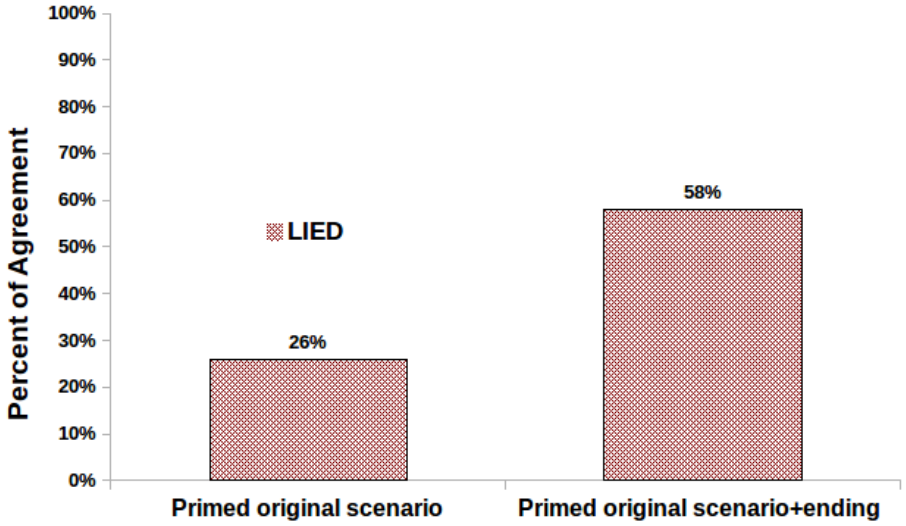


Figure 1.6: Wiegmann & al.' *third experiment*.

Finally, Wiegmann & al. introduced conversational and experimental pragmatics to argue that the Turris' two-parts response options generated a *framing effect*. We know that their main point was that splitting the response options led participants to reinterpret the experimentator's intentions. Given that one-part response options were interpreted as requests to assess whether someone lied or not, moving to two-parts options was understood as a request to assess objective falsity. To better support their claim, Wiegmann & al.' *fourth experiment* drew a comparison between verbs "to lie" and "to promise" whose understandings seem to be affected by framing effects. I won't go into the details of Wiegmann & al.' *fourth experiment* but they did observe the following phenomena: in case a speaker made a promise and participants were presented with one-part dichotomous options to pick ("He made a promise" vs. "He did not make a promise"), participants strongly agreed that the speaker *did make* a promise. But in the same situation, they were immediately less prone to agree that the speaker did make a promise when

presented with two-parts dichotomous options ("He tried to make a promise but failed to make a promise" vs. "He tried to make a promise and succeeded in making a promise").

From this observation, Wiegmann & al. infered that "*as in the case of 'lying', participants could be pushed into answering that the agent failed to make a promise although in the other conditions they clearly expressed that he did make a promise*" [41]. Thus, manipulating the response options strongly influenced the subjects' understandings of the experimentator's intended meaning(s). But, according to Wiegmann & al., taking into account these pragmatical reinterpretations should not lead to modify our common grasp of the verbs "to promise" and "to lie". *From the lying perspective*, the good reason to keep the standard subjective view is that moving to the objective one would only result from a misunderstanding caused by framing effects.

In *Lying Despite Telling the Truth*, Wiegmann & al. addressed the Turris' arguments on *perspective-taking* and *blame-opportunity*, and then pointed out two distinct flaws in their experimental protocols, namely *order* and *framing effects*. We know that their main goal was to rehabilitate the subjective definition against the Turris' attacks. In the remaining part of this chapter, I also advocate that the standard account is more adequate than the revised one. I provide new empirical data showing that people naturally endorse the subjective definition in critical condition.

1.4 The Subjective Core Definition of Lying

1.4.1 Main Point

Let us first remind the Turris' results for the predicate "Lied" in their *first experiment* and *replication*. Their initial story involved a possibly sly speaker, Jacob, who was telling the authorities a content he believed to be false but that turned out to be objectively true. Note that Jacob was alternatively honest or dishonest depending on whether (or not) he intended to deceive the authorities. In the dishonest case where Jacob actually intended to deceive the agents, 78,3 % of sub-

jects answered that he "lied" even though he told the truth. This result was then replicated with graded measures; participants again strongly agreed that Jacob lied at 5,59 (out of 7) in that case. Nevertheless, these results were immediately dismissed: the Turris condemned their protocols as being defective and decided to run further experiments to fully invalidate the subjective view. Immediately, though, Wiegmann & al. argued that the Turris' initial data were perfectly acceptable ("*the best test for assessing how people understand the concept of lying*" [38]) and that the subjective account was the proper definition of lying.

I do believe that Wiegmann & al. brought enough factual evidence that the subjective definition is appropriate as it stands. But I ran further experiments to shore up the subjective account through new empirical paths. I hypothesized that testing dispositional and episodic predicates derived from the root verb "to lie" could help frame the difference between the objective and subjective definitions. My goal was to correct a major confound in the Turris' experiments. Their designs gave an unfair advantage to the objective view by just providing episodic predicates[4] to assess a dishonest utterance as a lie or not.

So my own protocols aimed to reserve a better place to the subjective view by testing the respondents' intuitions on a dispositional predicate in addition to the episodic ones. I ran a pilot experiment to test people's intuitions on the dispositional/episodic pair "is a liar"/"has lied successfully"; and then I attempted to replicate these results for the more basic pair "is a liar"/"has lied". By doing so, participants were allowed to answer that a dishonest speaker *lied in the subjective sense* (by ascribing the dispositional predicate "is a liar") even though *he did not in the objective sense* (by *not* ascribing the episodic predicates "has lied successfully" or "has lied"), — *and the other way around*.

But I also wanted to know which role the speaker's false utterance and intention-to-deceive play when ascribing both predicative forms. In my experiments, I observed that the subjective sense of "lie" was driven *equally* by the speaker's false utterance and intention-to-deceive, while its objective sense was driven *mainly* by the speaker's intention-to-deceive. I now present the data I collected in a *pilot experiment* (for pair "liar"/"lied successfully") and in a tentative

[4] Such as "lied", "tried to tell a lie", "succeeded in telling a lie", etc.

replication (for pair "liar"/"lied").

1.4.2 Pilot Experiment

1.4.2.1 Pilot Hypothesis

The Turris' test questions ("*X* lied?") and options ("*X* tried to lie" vs. "*X* succeeded at"/ "*X* failed to"/"*X* only thinks") put a strong emphasis on the act of *having lied successfully*, that is on the episodic dimension of the verb "to lie". Hence, according to my hypothesis, their design was inevitably biased in favor of the objective view of lying. In my *pilot experiment*, the main goal was to correct this potential confound by offering a more-balanced testing protocol. To do so, my *pilot hypothesis* was that asking participants to rate "*X* is a liar?" (in addition to "*X* lied successfully" and to the Turris' predicate "*X* tried to lie"), could help emphasize the dispositional dimension of the verb "to lie" which is crucial but not treated in an equal way by our predecessors. By doing so, I aimed to show that the subjective definition more adequately captured people's understanding of "lie".

1.4.2.2 Design and Materials

The experiment rests on a 3×2×2 between-subject design, based on 3 factors (Predicate/Intention/Content), with 3 levels of the factor Predicate ("*X* is a liar"/"*X* tried to lie"/"*X* lied successfully"), 2 levels of the factor Intention ("Honest"/ "Dishonest") and 2 levels of the factor Content ("True"/"False"). The study was conducted online on Qualtrics and Mechanical Turk with a number of 122 participants (Raw Number $N = 145$). Subjects were randomly assigned to (alternative versions of) this general story:

> "*John believes that* [*Toronto/Ottawa*] *is the capital of Canada when in fact it is Ottawa. He is playing a trivia game in a bar. The question for this round is « What is the capital of Canada ? » An acquaintance from work, playing on one of the opposite teams, leans over and asks John for help. John* [*though*] *really wants to* [*win the match/help his friend*]. *So he responds* [*Ottawa/Toronto*]. *His colleague trusts John and answers accordingly.*"

Four conditions were built from this scenario depending on whether or not John intended to deceive his acquaintance and/or told him the truth about the Capital of Canada. These different conditions are summed up in the following 2×2 matrix:

		CONTENT	
		True	***False***
INTENTION-TO-DECEIVE	***Yes***	Dishonest/True	Dishonest/False
	No	Honest/True	Honest/False

Table 1.1: The 2×2 Matrix for the Pilot Conditions.

Let us clarify our four distinct conditions. In the critical condition [Dishonest/True], John intends to deceive his acquaintance but tells him the true answer (viz. "Ottowa"). In [Dishonest/False], John also intends to deceive his acquaintance and actually tells him a false answer (viz. "Toronto"). In [Honest/True], John is absolutely trustworthy (since he does not intend to deceive his acquaintance) and tells the truth (viz. "Ottowa"). Finally, in [Honest/False], John does not intend to deceive his acquaintance but mistakenly tells a falsehood (viz. "Toronto").

In a randomized order, I asked participants to "agree" or "disagree" concerning three statements about John. In each case, I put a confidence rating: "*How confident are you about your answer?*" (Scale: 0 - 10). Then, all participants were assigned to two controls in order to evaluate their general understanding of the story. Here are the three statements and two controls:

Statements:

(1) "*John is a liar.*" [Agree/Disagree] (+ confidence rating)

(2) "*John tried to lie.*" [Agree/Disagree] (+ confidence rating)

(3) "*John lied successfully.*" [Agree/Disagree] (+ confidence rating)

Controls:

1. "*Which game is John playing*?" [Chess/Trivia/Cards]

2. *"Which city does John believe the capital of Canada to be?"*

[London/Toronto/Ottawa]

Here I was mostly interested in the predicates "is a liar" and "lied successfully". "Tried to lie" was less essential to my prediction but was included to match the Turris' overall designs.

1.4.2.3 Results

Before we elaborate on the data, let us agree on two conventions I will use from now on. First, if at least 50% of the respondents agree that some factor α is crucial for ascribing some predicate Γ, then this factor α is assumed as a "*necessary condition*" for defining the meaning of Γ. Second, if at least 50% of the respondents all agree that a finite set of factors α, β, γ, etc. is enough for ascribing Γ, then this set of factors is considered to provide jointly "*sufficient conditions*" for capturing the essence of Γ.

The data of the pilot are summed up in Figure 1.7 above. Concerning the predicate "lied successfully", the data confirm the Turris' ones: a speaker succeeds at lying *if and only if* he is dishonest and his dishonest utterance is objectively false. Having an intention to deceive appears to be a necessary condition for success at lying: 100% of subjects agree that John *succeeded at lying* in condition [Dishonest/False] and 41% in [Dishonest/True] while only 6% and 32% of subjects do agree that John *succeeded at lying* in conditions [Honest/True] and [Honest/False]. In addition to that, uttering a false statement is also a necessary condition for a successful lie (compare [Dishonest/False]: 100% and [Dishonest/True]: 41%). Since the difference for "Lied successfully" between the [Dishonest/True] and [Dishonest/False] conditions is significant (Fisher Test, $p \leq 10^{-6}$, two-tailed), we can conclude that *"lied successfully" is driven by both the agent's intention to deceive* as well as the *falsity of his or her utterance* (in an equal way).

Note that in the [Dishonest/True] condition, 41% of agreement is close to average. It suggests that if John did not *succeed* at lying in this case, he may still *have lied in some sense*. Let us see if the data confirm predictions for predicate "is a liar".

As a matter of fact, the prediction is fulfilled: in the critical case [Dishonest/True], John is called a "liar" by 69 % of the participants. Even though he has not lied successfully, he is still a liar when he intends to deceive but happens to tell the truth.

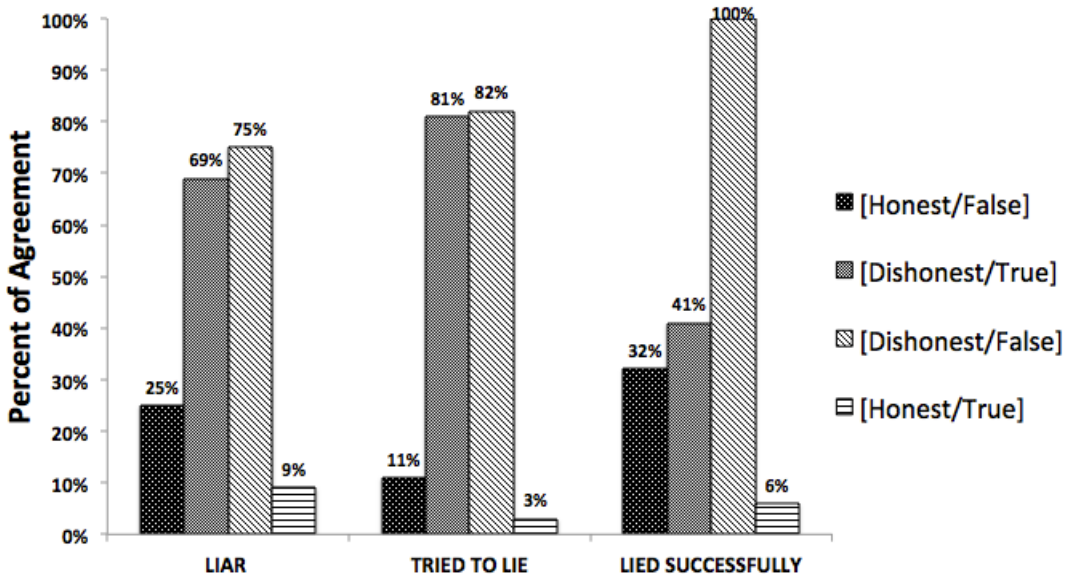


Figure 1.7: Pilot Results.

The results also show that intending to deceive is a necessary condition for being a liar: 69 % and 75 % of the subjects agree that John is a liar in conditions [Dishonest/True] and [Dishonest/False] (respectively) contra 9 % and 25 % in conditions [Honest/True] and [Honest/False] (respectively). As a matter of fact, the difference between "is a liar" is not significant in the two dishonest conditions (Yates' Chi-Square Test, χ^2 = 0,062). In addition to that, intending to deceive is also sufficient for being a liar: compare results from [Dishonest/True]: 69 % and [Dishonest/False]: 75 % which are above average with results for [Honest/True]: 9 % and [Honest/False]: 25 % which are below average. But the falsity of the asserted content also plays a role in saying that one is a liar. First, when John intends to deceive his acquaintance, he is *more of a liar* when he says something false rather than something true (compare [Dishonest/False]: 75 % vs. [Dishonest/True]: 69 %). Second, when John does not intend to deceive his acquaintance, more participants judge that he is a liar when he says an objective falsehood (compare [Honest/False]: 25 % vs. [Honest/True]: 9 %). Hence, the falsity of the utterance

increases the tendency to say that one "is a liar" but is not a necessary condition for being a liar. In that sense, predicate *"is a liar" is driven mainly by the agent's intention to deceive*.

Data for "is a liar" seem to be strongly correlated with the agent's attempt to lie. Each time John is said to be a liar (see dishonest conditions), people also strongly agree that he "has tried to lie" (see [Dishonest/True]: 81% and [Dishonest/False]: 82%). One plausible hypothesis is that the predicate "has tried to lie" also partially encodes the agent's intention to deceive. But, I have not deepened this hypothesis any further.

1.4.2.4 Analyses

We know that the literature distinguishes the subjective vs. objective understanding of "lie" by considering two essential factors: the speaker's intention to deceive and/or his ability to say something false. In that respect, a disputed case named "[Dishonest/True]" is when an agent intends to say something false, but ends up saying something true. Did the agent lie in this case? The Turris responded "No", but Wiegmann & al. answered "Yes".

My pilot experiment supported Wiegmann & al.' claim: people's understanding of "lie" is tied to the subjective definition. In [Dishonest/True], even though the agent did not "lie successfully" in the objective sense, he is "a liar" in the subjective one. This relies on the fact that "is a liar" is driven mainly by the agent's intention to deceive whereas "lied successfully" is driven equally by the agent's intention to deceive and the presence of a false utterance. Since the agent's utterance is true in [Dishonest/True], "lie" is understood through its *dispositional predicative form* ("is a liar") but not through its *episodic predicative form* ("lied successfully"). I proposed to replicate the pilot results to see more clearly into the behaviour of these dispositional vs. episodic forms.

1.4.3 Replication

1.4.3.1 Replication Hypothesis

I chose to investigate the effect of using distinct predicates based on the same root verb "to lie". I compared ascriptions of "*X* lied" vs. "*X* is a liar" vs. "*X* lied to *Y*". In the critical condition [Dishonest/True], I expected the result to be: "*X* is a liar" > "*X* lied" > "*X* lied to *Y*", namely that *X* is *more of a liar* than someone who *has lied* or *has lied to someone else* (*Y* in this case). The reason was that for "liar", I expected the intentional factor to prevail over the contentual one, for "lied to *Y*" I expected the contentual factor to prevail over the intentional one, and for "lie", I expected results to be mixed.[5]

1.4.3.2 Design and Materials

I came up with a 3×2×2 between-subject protocol based on 3 main factors: Predicate, Intention and Content. I had 3 levels of the factor Predicate ("*X* is a liar"/"*X* lied"/"*X* lied to *Y*"), 2 levels of the factor Intention ("Honest"/ "Dishonest") and 2 levels of the factor Content ("True"/"False"). The study was run online on the same platforms as before but with a larger number of the participants: 294 (Raw Number $N = 310$). People were randomly assigned to slightly modified versions of the initial scenario:

> "*John believes that* [*Toronto/Ottawa*] *is the capital of Canada* [*when/and*] *in fact it is Ottawa. He is playing a trivia game in a bar. The question for this round is « What is the capital of Canada? ». Sam, an acquaintance from work playing on one of the opposite teams, leans over and asks John for help. John* [*though*] *really wants to* [*fool/help*] *his colleague. So John tells Sam: "It is* [*Ottawa/Toronto*]*". The latter trusts John and answers accordingly.*"

The general scenario was modified in order to clarify *to whom* John lied to (viz. Sam). I clarified that Sam always trusts John such that if John performs a real "lie", his lie is necessarily successful. As in the *pilot experiment*, four different conditions were obtained depending on John's honesty and factual utterance:

[5] Central to my prediction was that "*X* is a liar" > "*X* lied". Whether "*X* lied" > "*X* lied to *Y*" or the opposite, was less essential. I decided to treat "*X* lied to *Y*" more as a control than as a variable of interest.

[Dishonest/True], [Dishonest/False], [Honest/True] and [Honest/False]. After having read *one* of these possible conditions, participants were asked to rate their "agreement"/"disagreement" about *one statement* out of the three proposed. In each case, I also added a confidence rating: *"How confident are you about your answer?"* (Scale: 0 - 10). Finally, all participants were assigned to six controls. Here are the three statements and six controls:

Statements:

(1) *"John is a liar."* [Agree/Disagree] (+ confidence rating)

(2) *"John lied."* [Agree/Disagree] (+ confidence rating)

(3) *"John lied to Sam."* [Agree/Disagree] (+ confidence rating)

Controls:

1. *"Which game is John playing?"* [Chess/Trivia/Cards]
2. *"Which response John gave to Sam?"*[Ottawa/London/Toronto]
3. *"Which city does John believe the Capital of Canada to be?"*

   [London/Toronto/Ottawa]
4. *"What is the name of John's colleague?"*[Billy/Jacob/Sam]
5. *"Did John say something true or false?"*[True/False]
6. *"Did John intend to deceive Sam?"*[Yes/No]

1.4.3.3 Results

My prediction was that we should find the pattern "*X* is a liar" > "*X* lied" > "*X* lied to *Y*" concerning the proportion of "Agree" in the critical condition [Dishonest/True]. I was expecting that when someone has an intention to deceive, but says something objectively true, the person is more readily called a "liar" than someone who "have lied". I was also entertaining the milder expectation that "John lied to Sam" is accepted in the cases in which "John lied" is accepted. The results are presented in Figure 1.8.

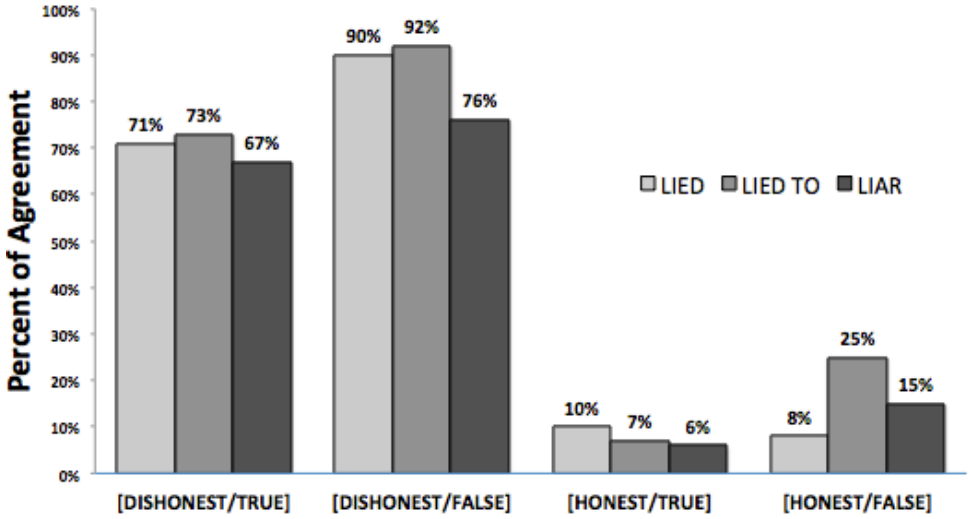


Figure 1.8: Replication Results.

The prediction is defeated. The new set of results indicates that there is basically *no difference* between the three kinds of predicates in the critical condition [Dishonest/True]. A majority of the participants (around 70%) considers that "John lied", "lied to Sam" and "is a liar" — despite saying something objectively true. In details, 73% of subjects agreed that John *lied to Sam*, 71% that he *lied* and 67% that he is a *liar*. This difference is in fact not significant (Yates' Chi-square, Test, $\chi^2 = 0{,}027$).

Consistent with the previous results (see Pilot results), the use of predicate "is a liar" appears to be driven mainly by the speaker's intention to deceive which is still a necessary and sufficient condition for being called a "liar". Concerning necessity, John is said to be a liar only in the dishonest conditions (see [Dishonest/True]: 67% and [Dishonest/False]: 76%), not in the honest ones (see [Honest/True]: 6% and [Honest/False]: 15%). Concerning sufficiency, predicate "is a liar" is ascribed above average only in the dishonest conditions, not in the honest ones (compare honest and dishonest results for "is a liar"). Nevertheless, scores confirm that falsity increases the tendency to say that one "is a liar" (first compare [Honest/False]: 15% vs. [Honest/True]: 6%, then compare [Dishonest/True]: 67% and [Dishonest/False]: 76%). But *"is a liar" is still driven mainly by the agent's intention to deceive*, not by the falsity of his dishonest assertion.

Unlike for "has lied successfully" (see Pilot results again), the predicate *"has lied" seems to be driven mainly by the agent's intention to deceive*, not by the falsity of his or her utterance (whose role is now secondary). Contrary to the Turris' claims, intending to deceive is a necessary and sufficient condition for *having lied*. First, John "has lied" only in the dishonest conditions: results for [Dishonest/True]: 71% and [Dishonest/False]: 90% are above average, not results for [Honest/True]: 10% and [Honest/False]: 8%. Second, even if there is a significative difference between conditions [Dishonest/True] (71%) and [Dishonest/False] (90%), the agent's intention to deceive appears to be sufficient for having lied. The falsity of the speaker's utterance plays a secondary role.

Finally, predicate *"has lied to Y" is driven mainly by the falsity of the agent's assertion* which is a necessary condition, though not sufficient, for *having lied to Sam*. When John is dishonest, more participants are prone to say that he has lied *to Sam* when he also tells him a falsehood (compare [Dishonest/True]: 73% and [Dishonest/False]: 92%). Similarly, when John is honest, more participants are willing to judge that he has lied *to Sam* when his assertion is false rather than true (compare [Honest/True]: 7% and [Honest/False]: 25%).

1.4.3.4 Post Hoc Analysis

Because of these surprising results compared to the first *pilot experiment*, I led a post hoc analysis of the *pilot experiment*, in which each participant answered questions concerning three different predicates, namely "*X* is a liar"/"*X* tried to lie"/"*X* lied successfully". I did this analysis because there may have been *order effects* influencing the results. So, for condition [Dishonest/True] in the *pilot experiment*, I calculated the percentage of participants' who responded to "*John lied successfully*" and to "*John is a liar*" when each of these sentences appears *first*, or *second after a control question*. The results are given in Figure 1.9. In fact, post hoc calculations are consistent with my pilot results for the [Dishonest/True] condition. So I can conclude that there are *no order effects* in people's responses for predicates "lied successfully" and "liar" in this case.

As I said, I performed this analysis to make a comparison with the attempted replication in which participants did not have any potential contrast effect due to

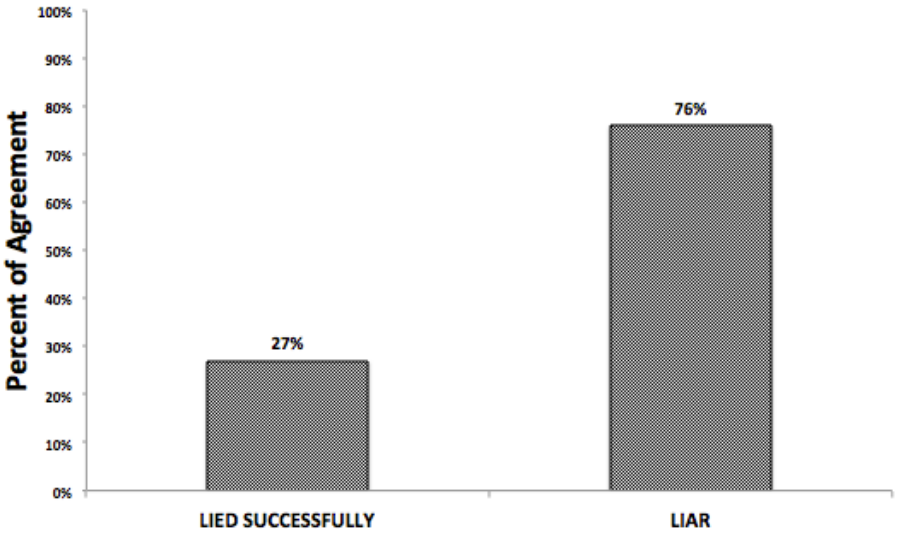


Figure 1.9: Post Hoc Analysis of the Pilot Results.

the successive presentation of the different predicates ("*John is a liar.*", "*John lied.*", "*John lied to Sam.*"). Let us now compare the data for predicates "*lied successfully*", "*lied*" and "*liar*" in the *pilot experiment* and *replication*.

1.4.4 Comparing the Results

Now *pilot* and *replication* results can be treated as between-subject data of the same type. Data for "*lied successfully*" (Post Hoc Analysis) and "*lied*" (Replication) in the [Dishonest/True] conditions are summed up in Table 1.2. Concerning predicates "*Lied successfully*" and "*Lied*", subjects are 27% to judge that John "lied successfully" in the *pilot experiment* and 71% that he "lied" in the tentative *replication*. A Fisher test confirms that this difference is *significant* at the 0.05 level (p = 0.0271, two-tailed).

Experiment	*Post Hoc Analysis*	*Replication*
Predicate	*Lied successfully*	*Lied*
Number "Agree" Responses	3	17
Number "Disagree" Responses	8	7
Total	11	24
Proportion of "Agree"	27%	71%

Table 1.2: Difference between "*lied successfully*" and "*lied*".

The results for the predicate "*Liar*" in the [Dishonest/True] conditions are summed up in Table 1.3. Subjects are 76% and 67% to judge that "John is a liar" in the *pilot* and *tentative replication* (respectively). In this case, a Fisher test indicates that this difference is *not significant* (p = 0.7106, two-tailed). The two tables are gathered in the comparative Figure 1.10.

Experiment	*Post Hoc Analysis*	*Replication*
Predicate	*Liar*	*Liar*
Number "Agree" Responses	10	16
Number "Disagree" Responses	3	8
Total	13	24
Proportion of "Agree"	76%	67%

Table 1.3: Differences for *"liar"*.

From the tentative *replication* and comparison with the *post hoc analysis*, we can conclude that "lied" and "liar" are in fact tracking *identical properties* to lay participants, who handle them *alike* (Fisher Test, p = 0,7672, two-tailed). On the contrary, "liar" and "lied successfully" really are treated *differently* (Fisher Test, p = 0,0377, two-tailed). In the [Dishonest/True condition], a consistent majority of people agrees that "John lied" (71%) and "is a liar" (at 76% and 67%), but only a small minority is willing to say that "John lied successfully" (27%) in this case. So the adverb *successfully* clearly directs participants to the truth or falsity of what is said. But *the default understandings of "lied" and "liar" are driven the agent's intention to deceive*.

1.5 Discussion

Remind that the main purpose of this chapter was *definitional*: how can we define "lying" properly? What is the adequate (viz. "good") definition of lying? I started investigating on this issue through conceptual means but rapidly took an empirical path to provide a more refined analysis of the notion. Conceptually,

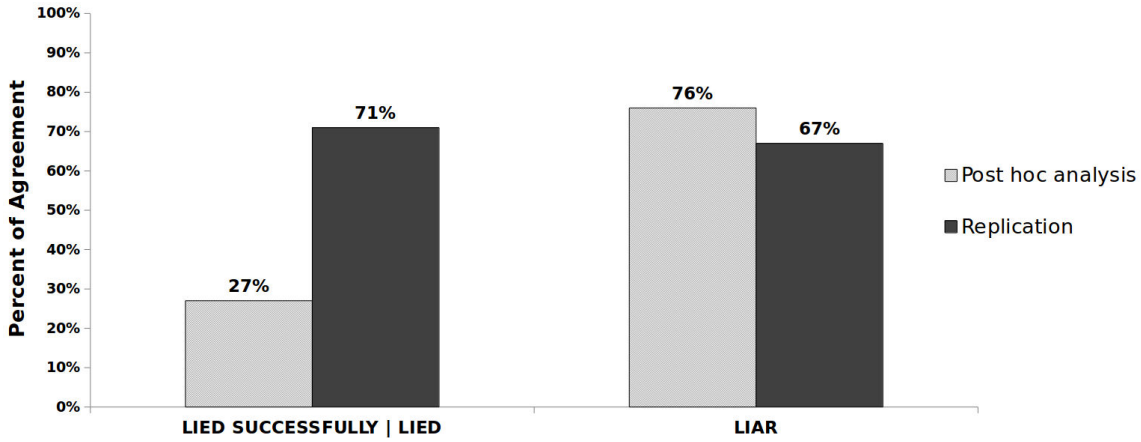


Figure 1.10: Comparison for [Dishonest/True].

I presented the standard subjective account which has been held by (analytical) epistemologists for centuries:

Subjective Def. A speaker X lies to a hearer Y on a proposition p *iff*

(1) X believes that p is false.

(2) X intends Y to believe that p is true.

(3) X tells Y that p.

Immediately, though, I explained that this aprioristic account has been challenged in two empirical ways leading to two distinct theories about lying. One of these challenging strategies was part of the *iff* definitional project and was promoted by Turri & Turri [2015]. Based on empirical results concerning the predicate "succeeded in telling a lie", the Turris called for an *objective theory* of lying. They argued that a *falsity condition* should be added to the traditional subjective definition to promote a new definition called "objective":

Objective Def. A speaker X lies to a hearer Y on a proposition p *iff*

(1) X believes that p is false.

(2) X intends Y to believe that p is true.

(3) X tells Y that p.

(4) p is false.

Their argument was that conditions **(1)-(2)-(3)** from the subjective definition were sufficient for the speaker to "have lied" but not to "have succeeded in telling a lie". In fact, success is acribed *only* when condition **(4)** is also met on top of conditions **(1)-(2)-(3)**. Even though dishonesty is sufficient for an utterance to be called a "lie", falsity is also necessary for the utterance to be called a "successful lie". From this, the Turris concluded that the objective definition was more adequate as a standard definition of lying than the traditional one. In other words, they used the data obtained for "succeeded in telling a lie" to make a more general claim concerning the very definition of lying.

In this request for objectivity, another challenging strategy was promoted by Coleman & Kay [1981] and stemmed from *prototype semantics*. Basically, Coleman & Kay condemned the *iff* project as illusory: no set of necessary and sufficient conditions could adequately capture the very essence of lying. Their argument was that "lie" behaves exactly the same as all other semantic categories. The category of "lie" has "*blurry edges and allows degrees of membership*" in such a way that "*applicability of a word [such as "lie"] to a thing is in general not a matter of 'yes or no', but rather of 'more or less'*" [Coleman & Kay 1981, 27]. As a consequence, Coleman & Kay concluded that there was *no core definition of lying* but only a *prototypical one* — the parameters of which being untruthfulness, intention-to-deceive *and* falsity. Here is a prototype for "lie" we can propose from their own:

Prototypical Lie. A prototypical lie is an utterance in which a speaker X asserts some proposition p to a hearer Y *such that...*

(1) X believes that p is false;

(2) In uttering p, X intends to deceive Y;

(3) p is false.

The Turris' *objective definition* and Coleman & Kay' *prototypical claim* have been the main challenging theories opposed to the subjective account (a summary is proposed in Table 1.4). But in this discussion part, I argue that the standard subjective account is sufficient for our definitional matters, viz. capturing the notion of lying in the colloquial sense. The necessary conditions tied to the

Turris' Objective Def.	Coleman & Kay' Prototype.
Speaker X lies to hearer Y on p *iff*	Speaker X lies to hearer Y on p *when*
(1) X believes that p is false.	**(1)** X believes that p is false.
(2) X intends Y to believe that p is true.	**(2)** In uttering p, X intends to deceive Y.
(3) X tells Y that p.	**(3)** p is false.
(4) p is false.	

Table 1.4: The Two Challenging Theories to the Subjective Definition.

subjective view prove to be sufficient for capturing the predicate "lied". On the contrary, the challenging theories (whose claims were based on results for predicate "lied successfully") are too demanding and fail to capture the root predicate "lied". That being said, the data leave two issues concerning predicates "lied successfully" and "lied". First, scores for "lied successfully" are quite high in the [Dishonest/True] case both in the pilot experiment and the post hoc analysis.[6] How can we account for this surprising result? Second, scores for "lied" increase when falsity is met by the deceptive speaker.[7] Again, how can we explain that falsity increases the tendency to say that one "lied"?

The replication results shows that people ascribe the predicate "lied" both in conditions [Dishonest/True] (71%) and [Dishonest/False] (90%), not in conditions [Honest/True] (10%) and [Honest/False] (8%). So the speaker's intention-to-deceive appears to be sufficient to qualify an utterance as a "lie". The falsity of the utterance is not required for ascribing the predicate "lied". Falsity pulls upwards the tendency to say that one "lied" but only plays a secondary role. A deceptive speaker is said to have "lied" no matter what she says turned out to be true or false. As a consequence, the subjective account (which does not adjudicate on the truth or falsity of the speaker's utterance) is peoples' default definition of lying.

[6] Respectively, 41% and 27 % of agreement.
[7] Compare the replication results for "lied" in [Dishonest/True] (71%) and [Dishonest/False] (90%).

Let us remind that the Turris' objective claim is based on results they got for predicate "lied successfully". Even though a dishonest speaker is said to have "lied" in [Dishonest/True], they observe that this speaker is considered to have "lied successfully" only in [Dishonest/False]. For this reason, they hold falsity to be a *necessary condition* for "lying successfully" and, basically, for *lying* properly.[8] Following a similar intuition, Coleman & Kay take falsity to be a *typical property* of a lie. My own pilot results converge in this *objective* direction for "lied successfully": a dishonest utterance is considered a successful lie only in [Dishonest/False] (100%), not in [Dishonest/True] (41%). The impact of falsity is also patent when we compare [Honest/True] (6%) and [Honest/False] (32%) for the same predicate.

Nevertheless, we can argue that an objective claim for "lied" based on results for "lied successfully" is possibly ill-founded. The adverb "successfully" seems to direct people to the falsity of what is said. Adding this complement is likely to be taken as a "demand characteristic", namely as a request to assess whether the speaker said something objectively false instead of whether the speaker has some intention-to-deceive the addressee. As a consequence, results for "lied successfully" cannot be used to adjucate on people's handling of "lied". The complement "successfully" leads them to set aside the speaker's intention-to-deceive to concentrate on his or her success (or failure) to utter a blatant falsehood. Being biased as such, the Turris' plea for objectivity stands as a poorer account of lying that the standard subjective definition.

But in order to be fully operative, the subjective account has to address two remaining issues concerning predicates "lied successfully" and "lied". Regarding the first, even though "lied successfully" does not obtain in the [Dishonest/True] case, scores are quite high for this specific condition in the pilot experiment (41%) as well as in the post hoc analysis (27%). How can we explain this outcome? My interpretation is that these results are caused by the way I phrased my experimental scenarios. All the vignettes I have proposed specified that the listener trusts the dishonest speaker and thus, believes his utterances as true in all cases. Concretely, I obtained this outcome by ending the *pilot scenario* with: "His colleague

[8] Since one cannot be said to have "lied successfully" without also being said to have "lied".

[the listener] trusts John and answers accordingly", and the *replication scenario* with: "The latter [the listener] trusts John and answers accordingly".

In doing so, however, I created a potential confound encouraging people to agree that *in some sense*, the speaker "lied successfully" in the [Dishonest/True] condition although she said something objectively true. In this case, the adverb "successfully" is not elicited by falsity (such as before) but by the speaker' success to deceive the addressee, namely to make him or her believe the opposite of what he or she believes. In fact, the word "successfully" is ambiguous in some way or another. "Having lied successfully" can be understood as "having succeeded in telling a falsehood" (first sense) but "having lied successfully" can also be understood as "having succeeded in deceiving the addressee" (second sense). This latter sense is elicited by the way I completed the scenarios and explains the high scores I get for "lied successfully" in the [Dishonest/True] condition.

Concerning the second issue, falsity appears to have an enhancing effect on the ascription of "lied". This is patent in the replication when we look at conditions [Dishonest/True] (71%) and [Dishonest/False] (90%). How can we account for this effect? How can we explain that people are more willing to agree that a dishonest speaker lied when he or she says something false instead of something true? This could be interpreted as a positive reason to favor one of the challenging definitions above over the subjective account. Some could argue that people being more likely to ascribe predicate "lied" when falsity obtains, falsity *is a necessary condition* or *a typical property* of lies. But a rationale can be given for explaining this increasing effect from the subjective perspective.

In the replication experiment, the potential inconvenience caused to the addressee differs between conditions [Dishonest/True] and [Dishonest/False]. In the [Dishonest/True] case, the hearer is deceived by the speaker (which makes him believe the opposite of his own belief) *but* he ends up believing something true (Ottawa *is* the capital of Canada). So the hearer is deceived about the speaker's belief *but* he is not deceived about the world. In the [Dishonest/False] case, however, the hearer is deceived about the speaker's belief *and* about the world (Toronto *is not* the capital of Canada). As a result, falsity in *false lies* allows *bad side-effects* that truth in *true lies* does not. So false lies are *more harmful* than true lies for

the addressee and participants are more willing to blame them for their greater inconvenience.

Following Joshua Knobe's observations in experimental psychology [e.g. Knobe 2003ab 2006], this moral asymmetry between false lies and true lies might explain the higher score we obtain for "lied" in the [Dishonest/False] condition. Falsity tends to create a "Side-Effect Effect", now known as a "Knobe Effect", in which the greater *inconvenience* (or *harmfullness*) caused by falsity leads people to judge a false assertion to be *more intentional* than a true assertion. As Knobe [2003a] puts it: people "*seem considerably more willing to say that a side-effect was brought about intentionally when they regard that side-effect as bad than when they regard it as good*". So a *dishonest false assertion* (which is seen as "bad" by participants) has greater chances to be called a "lie" than a *dishonest true assertion* (which is seen as "good" by them or as "better" than the first). To sum up, the impact of falsity on people's ascriptions of "lied" most likely results from a Knobe effect caused by falsity itself.

1.6 Conclusion

Contra Coleman & Kay [1981], we can conclude that there is more to the definition of lying than just prototypical information. There is a *core definition* of lying. The definitional clauses are not just typical properties, they are necessary conditions. The speaker's intention-to-deceive is one of these necessary conditions but the falsity of the utterance is not. Falsity enhances peoples' ascriptions of "lied" but its role is ancillary. So, contra Turri & Turri [2015], we can conclude that the subjective account is a better candidate as a core definition of lying. The objective account is too demanding in that respect.

But endorsing the subjective perspective raised several issues I tried to address successively. First, my pilot experiment confirmed one of the Turris' main observation: a dishonest speaker is said to have "lied successfully" only when she says something false. My interpretation is that adding "successfully" to the root predicate "lied" moves people's interest from assessing the speaker's utterance as a "lie" to assess it as "true" or "false". Based on this confound, I argued that

the Turris' defense of the objective definition is possibly ill-founded.

Second, I tried to explain the high scores I got for "lied successfully" in the [Dishonest/True] condition. My explanation is that this outcome results from the way I phrased the protocols. In the scenarios, I forced the lie to be *successful* by specifying that the hearer *always* comes to believe the dishonest utterance one it has been made. Quite naturally then, I got substantive scores for "lied successfully" even though the speaker's dishonest utterance turned out to be true.

Third, I tried to account for the enhancing effect falsity has on people's ascriptions of "lied". I argue that falsity generates a Knobe effect: being more harmful for the addressee, *false assertions* are interpreted as *more intentional* than *true assertions* to lay participants. As a result, ascriptions of "lied" tend to increase when the dishonest speaker tells a falsehood instead of the truth.

Lies are *standard strategies* of deception. The present chapter has reviewed the traditional debates on the definition of lying, and defended the subjective account based on new empirical data. In contrast, Chapter 2 investigates *non-standard strategies* of deception that are *misleading default inferences* and *omissions of information*. For a dishonest speaker, strategic omission consists in hiding relevant pieces of information to some intended addressee. A misleading default inference is an inference whereby some addressee reaches a false conclusion through default reasoning. In case of deception, a sly speaker can perfectly trigger such a misleading inference to fool the addressee by omission. This is what happens in the *Surprise Deception Paradox* I analyze in Chapter 2.

Chapter 2

The Surprise Deception Paradox

2.1 Introduction

In one of his riveting books on paradoxes entitled *What is the Name of this Book?*, the logician Raymond Smullyan tells an anecdote concerning his first introduction to logic [see Smullyan 1978]. As he was suffering from flu on April 1st 1925, his older brother — *Emile* — came to his bedroom and said to him: "Well, *Raymond*, today is April Fool's Day, and I will deceive you as you have never been deceived before!". After this announcement, *Raymond* waited all day long to be deceived but (apparently) nothing happened... So, late at night, he was no longer expecting to be deceived (or so he thought) but still felt highly concerned by his brother's announcement. As a consequence, *Raymond*'s mother intervened to ask *Emile* why he had (apparently) not deceived him. At this moment, *Emile* turned to *Raymond* and said to him:

Emile: "*So, you expected me to deceive you, didn't you?*

Raymond: *Yes.*

Emile: *But I didn't, did I?*

Raymond: *No.*

Emile: *But you expected me to, didn't you?*

Raymond: *Yes.*

Emile: *So I deceived you, didn't I?*"

After this explanation, *Raymond* lay in his bed and started reasoning about it. On the one hand, supposing that he wasn't deceived, then he did not get what he expected (because he expected to be deceived after *Emile*'s announcement!), and hence he was actually deceived. But on the other hand, supposing that he was deceived, then he exactly did get what he expected, and hence he was not "deceived".

In this chapter, I follow two investigative paths. First, I define the *theoretical notions* at stake in the story. At first glance, the notions of *deception*, *deception by omission* and *surprise* are involved in this deceptive tale. In fact, *Emile* performs a *deception by omission* on April 1^{st} 1925 that also leads his brother to be *surprised*. But, as a matter of fact, I will show that *Emile*'s "action" of deception is more complex than it seems at first sight. It revolves around *distinct deceptive moves* which lead to *consecutive states of deception* and *surprise*. Second, I deal with some *theoretical issues* raised by Smullyan's story. I want to better understand what is puzzling in the reasoning *Raymond* performs later in the evening. But contrary to *Raymond*'s own argument, I show that this reasoning is not of a paradoxical nature. Moreover, I argue that this reasoning is of a lesser significance than the announcement *Emile* makes earlier in the morning. This early announcement should not be successful *in principle* due to its self-defeating feature. But the announcement is successful (*Raymond* is surprised at the end of the day), only because *Raymond* is not a perfect reasoner.

This chapter is structured as follows. In section 2.2, I introduce a logical apparatus for analyzing Smullyan's story in details. Based on work by van Benthem [2007] and Baltag & Smets [2006 2008b], the framework consists in an epistemic language extended with a dynamic operator of *Belief Radical Upgrade* (subsection 2.2.1). The syntax is then interpreted in epistemic plausibility models that can be fully axiomatized by reduction axioms (subsection 2.2.2). This language is finally matched with Smullyan's story for clarity purposes (subsection 2.2.3).

This formal setting helps me define in section 2.3 the deception that takes place in the story. First, I point out that the announcement *Emile* makes in the morning is pragmatically misleading (subsection 2.3.1). Then, I focus on the distinct states

of deception *Raymond* goes through during that day (subsection 2.3.2). That being done, I describe the actions that lead to those states of deceptions (subsection 2.3.3). I end by summarizing *Emile*'s whole deceptive plot (subsection 2.3.4).

My purpose in section 2.4 is to model the dynamics of *Emile*'s later explanation step-by-step. To do so, I use the plausibility machinery introduced in section 2.2. Through this modelization, *Emile*'s explanation appears as an opportunity to unveil his whole deceptive trick (subsection 2.4.1). But the story shows that this disclosure leads *Raymond* to fall in puzzling thoughts. However, I argue that those puzzling thoughts are not of a paradoxical nature (subsection 2.4.2). The steps that might lead to contradictions can be easily bypassed.

Finally, section 2.5 is devoted to analyze the surprise which is induced by *Emile*'s explanation. By comparing *Emile*'s announcement with the announcement of surprise a teacher makes in the famous Surprise Exam Paradox [e.g. O'Connor 1948, Scriven 1951, Shaw 1958], I argue that *Emile*'s announcement is also an announcement of surprise since *Emile* claims that he will deceive his brother as *he has never been deceived in the past* (subsection 2.5.1). However, I argue that *though being true* such an announcement cannot be successful *in principle*. *Raymond*'s announcement is successful only because he is not a perfect reasoner. If he were a perfect reasoner, *Emile*'s announcement could not be successful and his deception would fail to be a surprise (subsection 2.5.2). That being said, I finally describe *Raymond*'s *states of surprise* in detail (subsection 2.5.3).

More broadly, I aim to show that Smullyan's story is informative of the way one can deceive while saying the truth. We may call *veridical deception* that strategy by which one causes someone else to hold a false belief through a true piece of information [e.g. Adler 1997, Meibauer 2014, Fallis 2015]. In Smullyan's tale, veridical deception is used when *Emile* makes a *true announcement* ("*Today, I will deceive you as you have never been deceived before!*"), which suggests the false conclusion that he will deceive *Raymond* by doing some particular action (*deception only by commission*) when in fact he does not (*deception only by omission*). But veridical deception is a broader category including other types of strategies I will evoke such as *false implicatures*, *pretending to deceive* and *presupposition faking*.

2.2 A Language for Analysis

2.2.1 A Dynamic Epistemic Syntax $\mathcal{L}_{(B,K,[\Uparrow])}$

I start out by defining a dynamic epistemic logic $\mathcal{L}_{(B,K,[\Uparrow])}$ which is a propositional syntax containing static operators for (conditional) belief and knowledge as well as a dynamic operator for Lexicographic Upgrade [e.g. van Benthem 2007 2014] or Belief Radical Upgrade [e.g. Baltag & Smets 2006 2008b]. The set $\mathcal{F}_{(B,K,[\Uparrow])}$ of all $\mathcal{L}_{(B,K,[\Uparrow])}$-*formulas* is given by the following Backus-Naur Forms:

$$\langle \textbf{\textit{Formulas}} \rangle \qquad \varphi, \psi \ := \ \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid \boldsymbol{B}_R^{\psi}\varphi \mid \boldsymbol{K}_R\varphi \mid [\Uparrow \varphi]\psi$$

$$\langle \textbf{\textit{Atoms}} \rangle \qquad p \ := \ d^{+} \mid d^{-}$$

Here $\top$ is the common abbreviation for tautologies, the subscript R designates the particular agent *Raymond* and p is some atomic proposition d^{+} or d^{-}. The additional propositional connectives ($\bot$,$\vee$,$\rightarrow$,$\leftrightarrow$) are defined as usual. According to standard usage in the epistemic logic literature, the intended interpretations of the operators $\boldsymbol{B}_R^{\psi}\varphi$, $\boldsymbol{K}_R\varphi$ and $[\Uparrow \varphi]\psi$ are respectively " *Raymond believes that* φ *conditional on formula* ψ"[1], "*Raymond knows that* φ" and "*after a belief radical upgrade with formula* φ, ψ *is the case*". The operator $\boldsymbol{B}_R\varphi$ which stands for "*Raymond plainly believes that* φ" can be defined from $\boldsymbol{B}_R^{\top}\varphi$ standing for "*Raymond conditionally believes that* φ *on* $\top$": $\boldsymbol{B}_R\varphi \ := \boldsymbol{B}_R^{\top}\varphi$.

2.2.2 Epistemic Plausibility Models for Language $\mathcal{L}_{(B,K,[\Uparrow])}$

I give a semantic interpretation to $\mathcal{L}_{(B,K,[\Uparrow])}$ in terms of "Epistemic Plausibility Models" [see van Benthem 2007, Baltag & Smets 2006 2008b, van Benthem & Smets 2015] instead of more classical "Kripke Models" used in Public Announcement Logics for instance [e.g. Gerbrandy & Groeneveld 1997, Baltag *et al.* 1998, Plaza 2007].

I will show hereafter that this semantic choice is motivated by Smullyan's own

[1] Following Baltag & Smets [2011], we can think of conditional beliefs $\boldsymbol{B}_R^{\psi}\varphi$ as 'contingency' plans for belief change: in case *Raymond* will find out that ψ was the case, he will believe that φ was the case.

recollection. In fact, *Raymond* uses a *default rule* to interpret his brother's announcement of a future deception as being a deception of a particular kind, namely *deception only by commission*. This default rule can be conceived as a *plausibility rule* corresponding to *the most plausible way* people would interpret *Emile*'s early announcement: "Well, *Raymond*, today is April Fool's Day, and I will deceive you as you have never been deceived before!". Let us first introduce the analytic framework.

A pointed plausibility model for $\mathcal{L}_{(B,K,[\Uparrow])}$ is a relational structure $\mathbb{S}$ such that:

$$\mathbb{S} = \langle\, \mathcal{S}, \leqslant_R, \| \cdot \|, s_0^\star \,\rangle$$

The structure $\mathbb{S}$ consists of:

- $\mathcal{S}$ which is a finite non-empty set of "*possible states*" s (or "*worlds*").
- $\leqslant_R$ which is a "*plausibility relation*" for *Raymond* such that $\leqslant_R \subseteq \mathcal{S} \times \mathcal{S}$. This relation $\leqslant_R$ is a total *preorder*, that is to say a transitive, strongly connected and thus reflexive relation over the set $\mathcal{S}$.
- $\| \cdot \| : At \to \wp(\mathcal{S})$ which is a standard "*valuation map*" where At is the set of all atomic formulas p, and $\wp(\mathcal{S})$ is the set of subsets of $\mathcal{S}$.
- A pointed state $s_0^\star \in \mathcal{S}$ is called the "*actual state*" in model $\mathbb{S}$.

The conventional reading of the plausibility order is the following: when $s \leqslant_R t$ (for all $s, t \in \mathcal{S}$), *Raymond* considers the state t to be "*at least as plausible as*" the state s. In other words, he considers t to be equally or more plausible than s.

For simplicity's sake, we define *comparability relations* $\sim_R$ from *plausibility orders* $\leqslant_R$ as follows: $s \sim_R t$ *iff* either $s \leqslant_R t$ or $s \geqslant_R t$ (where $\geqslant$ is the converse of $\leqslant$). This relation is a *symmetric* preorder since the conditions for $s \sim_R t$ and $t \sim_R s$ to obtain are strictly identical. That is: when one obtains, the other is given for free. So $\sim_R$ is an *equivalence relation* over $\mathcal{S}$.

Truth of a formula φ at a possible state s in an epistemic plausibility model $\mathbb{S}$, denoted $\mathbb{S}, s \models \varphi$, is defined inductively by simply extending the valuation map to all formulas belonging to $\mathcal{F}_{(B,K,[\Uparrow])}$:

- $\mathbb{S}, s \models \top$ *Always*.
- $\mathbb{S}, s \models p$ *iff* $s \in \| p \|$.
- $\mathbb{S}, s \models \neg\varphi$ *iff* $\mathbb{S}, s \not\models \varphi$.
- $\mathbb{S}, s \models \varphi \wedge \psi$ *iff* $\mathbb{S}, s \models \varphi$ and $\mathbb{S}, s \models \psi$.
- $\mathbb{S}, s \models \boldsymbol{K}_R\varphi$ *iff* for all $t \in \mathcal{S}$, if $s \sim_R t$ then $\mathbb{S}, t \models \varphi$.
- $\mathbb{S}, s \models \boldsymbol{B}_R^{\psi}\varphi$ *iff* for all $t \in Max_{\leqslant_R}^{\|\psi\|}$: $\mathbb{S}, t \models \varphi$.

Here $Max_{\leqslant_R}^{\|\psi\|}$ is the set of states that satisfy ψ and are maximal for the ordering $\leqslant_R$ in model $\mathbb{S}$. That is: $Max_{\leqslant_R}^{\|\psi\|} := \{u \in \mathcal{S} \mid \forall v \in \mathcal{S}\ v \leqslant_R u \text{ and } \mathbb{S}, u \models \psi\}$. Those states on which Raymond conditions his beliefs in φ are the states he considers to be the most plausible ψ-states that also satisfy φ. As I said, plain beliefs are recovered from conditional beliefs by conditioning on $\top$, so the last clause can be reduced to:

- $\mathbb{S}, s \models \boldsymbol{B}_R\varphi$ *iff* for all $t \in Max_{\leqslant_R}^{\|\top\|}$: $\mathbb{S}, t \models \varphi$.

Following van Benthem [2007] and Baltag & Smets [2008b], I define a Belief Radical Upgrade with a formula φ, written $[\Uparrow \varphi]$, as a mapping:

$$[\Uparrow \varphi] : \mathbb{S} \rightarrow \mathbb{S}^{[\Uparrow\varphi]}$$

Note that $\mathbb{S}$ is the initial model and $\mathbb{S}^{[\Uparrow\varphi]}$ is the transformed model obtained after having performed the intended operation $[\Uparrow \varphi]$. Here is the semantic clause for belief radical upgrade:

- $\mathbb{S}, s \models [\Uparrow \varphi]\psi$ *iff* $\mathbb{S}^{[\Uparrow\varphi]}, s \models \psi$.

The epistemic plausibility model $\mathbb{S}^{[\Uparrow\varphi]}$ is defined from $\mathbb{S}$ in the following way:

$$\mathbb{S}^{[\Uparrow\varphi]} = \langle\, \mathcal{S}^{[\Uparrow\varphi]}, \leqslant_R^{[\Uparrow\varphi]}, \|\cdot\|^{[\Uparrow\varphi]}, s_0^{\star} \,\rangle$$

where:

- $\mathcal{S}^{[\Uparrow\varphi]} := \mathcal{S}$
- $\leqslant_R^{[\Uparrow\varphi]} := \big(\leqslant_R \cap (\mathcal{S} \times \| \varphi \|)\big) \cup \big(\leqslant_R \cap (\| \neg\varphi \| \times \mathcal{S})\big) \cup \big(\sim_R \cap (\| \neg\varphi \| \times \| \varphi \|)\big)$
- $\| \, . \, \|^{[\Uparrow\varphi]} := \| \, . \, \|$

Here, $\mathcal{S}^{[\Uparrow\varphi]}$ and $\| \, . \, \|^{[\Uparrow\varphi]}$ are identical to $\mathcal{S}$ and $\| \, . \, \|$ from the initial model $\mathbb{S}$. The special feature of $\mathbb{S}^{[\Uparrow\varphi]}$ is the plausibility order $\leqslant_R^{[\Uparrow\varphi]}$. Following van Benthem & Liu [2007], this relation is defined as being equal to the initial plausibility order $\leqslant_R$ in which the ordered pairs (s,t) satisfying respectively formulas φ and $\neg\varphi$ have been put at the bottom of the ordering. This guarantees that all $\neg\varphi$-states are eliminated from the top of the ordering and put at the bottom in model $\mathbb{S}^{[\Uparrow\varphi]}$. In practice then, when *Raymond* makes a belief radical upgrade with a formula φ, he puts all φ-states from his initial model $\mathbb{S}$ *at the top of the plausibility ordering* in his upgraded model $\mathbb{S}^{[\Uparrow\varphi]}$, leaving everything else the same. So the states satisfying φ become the *most plausible* states in model $\mathbb{S}^{[\Uparrow\varphi]}$.

Rather than the technicalities, what matters is the particular doxastic attitude corresponding to the dynamic operation $[\Uparrow \varphi]$. When some agent performs a belief radical upgrade with a piece of information φ, he or she does not know the source of the incoming φ to be *absolutely reliable*, thus *completely honest* and *truthful*. But he or she believes the source to be *highly reliable*, thus *strongly honest* and *truthful*. According to van Benthem's terminology, the agent does not have a *hard* doxastic attitude towards the source of information, but a *softer* one. The agent is not sure that the source always tells the truth but he or she is *highly confident* that the source does.

From a semantic point of view, a belief radical upgrade is a *plausibility rule of interpretation*. More precisely, this is a *strong* plausibility rule which takes the incoming information φ to be *highly plausible* and state $(\varphi \wedge \psi)$ to be *more plausible* than state $(\varphi \wedge \neg\psi)$. Many other plausibility rules can be found in the literature which are *weaker* and take the incoming information and states to be *less plausible* than before. They are either *conservative upgrades* [see van Benthem 2007, Baltag

& Smets 2008b] or more subtle dynamic attitudes such as *positive* and *negative upgrades*, *semi-positive* and *semi-negative ones*, *minimal attitudes*, *trivial ones*, etc. [see Baltag *et al.* 2012, Rodenhäuser 2014].

Basically, a belief radical upgrade can be conceived as a *default rule of interpretation* for a given semantic content. It takes an incoming formula φ and computes that, based on φ, formula ψ is *more plausible* than formula $\neg\psi$ in all the cases in which φ obtains. This notion of *higher plausibility* is also the core intuition governing non-monotonic reasoning since the 1990's. Veltman [1996] has tried to offer a *formal rule* to capture default reasoning. Like belief upgrades, this default rule is based on the notion of *higher plausibility*. Formally, the rule is written $(\varphi \Rightarrow \psi)$ and means that formulas "φ *are normally* ψ" or, to put it another way, that worlds satisfying the conjunction $(\varphi \wedge \psi)$ are *more plausible* that worlds satisfying the conjunction $(\varphi \wedge \neg\psi)$. For this reason, I propose in this chapter to approximate belief radical upgrades with default rules.

Following van Benthem [2007] and Baltag & Smets [2008b], the language $\mathcal{L}_{(\boldsymbol{B},\boldsymbol{K},[\Uparrow])}$ can be axiomatized completely by the following axioms:

- *All instances of propositional tautologies.*
- *The S5 axioms for Knowledge.*
- *The S4 axioms for plain Belief.*
- *This recursion axiom for Conditional Belief:*

$$\vdash \boldsymbol{B}_R^{\psi}\varphi \quad \leftrightarrow \quad \neg\boldsymbol{K}_R\neg\psi \rightarrow \neg\boldsymbol{K}_R\neg(\psi \wedge \boldsymbol{B}_R(\psi \rightarrow \varphi))$$

- *The usual recursion axioms for Belief Radical Upgrade:*

$$\begin{aligned}
&\vdash [\Uparrow \varphi]p && \leftrightarrow && p \text{ for all atomic proposition letters } p.\\
&\vdash [\Uparrow \varphi]\neg\psi && \leftrightarrow && \neg[\Uparrow \varphi]\psi\\
&\vdash [\Uparrow \varphi](\psi \wedge \phi) && \leftrightarrow && [\Uparrow \varphi]\psi \wedge [\Uparrow \varphi]\phi\\
&\vdash [\Uparrow \varphi]\boldsymbol{K}_R\varphi && \leftrightarrow && \boldsymbol{K}_R[\Uparrow \varphi]\varphi\\
&\vdash [\Uparrow \varphi]\boldsymbol{B}_R^{\psi}\phi && \leftrightarrow && (\neg\boldsymbol{K}_R\neg(\varphi \wedge [\Uparrow \varphi]\psi) \wedge \boldsymbol{B}_R^{\varphi\wedge[\Uparrow\varphi]\psi}[\Uparrow \varphi]\phi)\\
& && && \vee(\boldsymbol{K}_R\neg(\varphi \wedge [\Uparrow \varphi]\psi) \wedge \boldsymbol{B}_R^{[\Uparrow\varphi]\psi}[\Uparrow \varphi]\phi)
\end{aligned}$$

➤ *The necessitation rule for Belief Radical Upgrade:*

If $\vdash \psi$, then $\vdash [\Uparrow \varphi]\psi$

This axiomatization is not purely ornamental. It will ensure the soundness of the logical derivations I will make concerning *Raymond*'s conclusion that he is deceived *if and only if* he is not deceived (subsection 2.4.2).

2.2.3 Matching $\mathcal{L}_{(B,K,[\Uparrow])}$ with Smullyan's Story

In the context of Smullyan's story, the atomic propositions d^+ and d^- express the two ways by which one can understand *Emile*'s early announcement of deception. The first way is given by proposition d^+ meaning that *Emile* will deceive *Raymond by doing some action* on April 1^{st} 1925. In other words, *Emile* will make *some move* such that *Raymond* will be *deceived on some fact* after this very move. Such a strategy is known as *"deception by commission"* in the literature [e.g. Chisholm & Feehan 1977]. But another way of understanding *Emile*'s announcement is given by proposition d^- meaning that *Emile* will deceive *Raymond by no action* on April 1^{st} 1925. In that case, *Emile* won't perform any action but by doing so, he will actually deceive *Raymond* by depriving him of some piece of information. This is *"deception by omission"* in the literature [e.g. Chisholm & Feehan 1977]. Here are listed the exact meanings of propositions d^+ and d^-:

d^+ : "*Raymond* will be deceived by commission on April 1^{st} 1925".
d^- : "*Raymond* will be deceived by omission on April 1^{st} 1925".

Now, we let a complex formula D stand for the content of *Emile*'s announcement in the morning:

D : "Well, (...) I will deceive you as you have never been deceived before".

Propositions d^+, d^- and formula D can be conceived as *types of (predictive) actions of deception*. They encode the different ways by which *Emile* can deceive his brother throughout the day, namely the *types* of deceptive "actions", and their combinations, he can deploy to trick him. Actually *three distinct combinations* can

be considered in the logical space: $(d^+ \wedge \neg d^-)$, $(\neg d^+ \wedge d^-)$ and $(d^+ \wedge d^-)$. The first combination means that *Raymond* will be deceived *only* by commission on April 1^{st} 1925, the second means that he will be deceived *only* by omission on that date and the last combination means that he will be deceived *both* by commission *and* by omission. In case *Raymond* will *neither* be deceived by commission *nor* by omission on April 1^{st} 1925, then propositions d^+ and d^- are false $(\neg d^+ \wedge \neg d^-)$ and the announcement itself would turn out to be false: $\neg D$.

If we do not take *Emile*'s announcement as a *bona fide* action, *Raymond won't be deceived only by commission* but only *by omission* on April 1^{st} 1925. First, *Raymond won't be deceived by commission*, written $\neg(d^+ \wedge \neg d^-)$, since *Emile* does not undertake *any action* to deceive him after his early announcement. Second, *Raymond* will be *deceived only by omission*, written $(\neg d^+ \wedge d^-)$, since his brother will keep *doing nothing* for the rest of the day to trick him. That being said, we can remark that the combination $(d^+ \wedge d^-)$ does not play any role in Smullyan's story. Though possible, this combination is not taken into consideration by *Emile* and *Raymond*. None of them invoke this combination as plausible in their respective parts. As a consequence, I will set aside this type of deception in my upcoming analyses.

In the pragmatic sense, formula D seems to imply the conjunction $(d^+ \wedge \neg d^-)$ stating that *Emile* will deceive *Raymond* only by commission. This relies on the default inference one would make to interpret some announcement of *deception* in common circumstances. After such an announcement, written $[\Uparrow D]$, one would naturally infer that deception only by commission $(d^+ \wedge \neg d^-)$ is a *more plausible* option than the negation of deception only by commission $\neg(d^+ \wedge \neg d^-)$ and even more plausible than deception only by omission $(\neg d^+ \wedge d^-)$. In other words, one would infer that the conjunction $D \wedge (d^+ \wedge \neg d^-)$ is a *more plausible* option than the conjunction $D \wedge \neg(d^+ \wedge \neg d^-)$ and even more plausible than the conjunction $D \wedge (\neg d^+ \wedge d^-)$. The underlying default rule at stake here is then $D \Rightarrow (d^+ \wedge \neg d^-)$ and can be approximated by the formula $[\Uparrow D](d^+ \wedge \neg d^-)$ in language $\mathcal{L}_{(B,K,[\Uparrow])}$.

The problem is that this latter formula is false in the context of Smullyan's tale since it is not the case that *Raymond* will be deceived only by commission after his

brother's annnouncement. Actually, *Emile*'s explanation reveals that *Raymond* is deceived *only* by omission at the end of the day. Though being perfectly natural as a default, the rule $D \Rightarrow (d^+ \wedge \neg d^-)$ should not be used by *Raymond* on April 1^{st}. In fact, using such a rule leads him to draw the false conclusions that he will be deceived only by commission and thus, that he won't be deceived only by omission. I will show later that having those false beliefs makes him being surprised afterwards. Before doing this, I take the opportunity to describe *Emile*'s misleading announcement in more details.

2.3 The Deceptive Plot

2.3.1 *Emile*'s Misleading Announcement

I first concentrate on the *pragmatic meaning* of *Emile*'s announcement to show that it is completely misleading. Though semantically true, this announcement is *pragmatically* misleading since it naturally conveys the wrong idea that *Emile* will deceive *Raymond* by doing some specific action (*deception only by commission*) when, in fact, he only uses deception by omission to trick his brother (*deception only by omission*).

So in formal terms, the announcement D is misleading as a content since it pragmatically conveys the wrong idea that formulas $(d^+ \wedge \neg d^-)$ and $(\neg d^+ \wedge d^-)$ are respectively *true* and *false*. But after *Emile*'s early announcement and then on, *Raymond* will not be deceived by commission, and for this reason, he will actually be deceived only by omission. So formula $(d^+ \wedge \neg d^-)$ is *false* while formula $(\neg d^+ \wedge d^-)$ is *true*.

In Smullyan's story, *Raymond*'s states of deception result from the fact that he has wrong beliefs about formulas $(d^+ \wedge \neg d^-)$ and $(\neg d^+ \wedge d^-)$ throughout the day. I will now describe these two states of deception *Raymond* successively goes through during the day.

2.3.2 *Raymond*'s Successive States of Deception

I start out by defining what it means for *Raymond* to be deceived on a formula ψ. Let us agree that *Raymond* is in a state of deception about ψ, written *Deceived* $_{On\ \psi}$, if and only if *Raymond* believes that ψ is *false* (*true*) when in fact, ψ happens to be *true* (*false*). So, for some formula $\psi \in \mathcal{L}_{(\boldsymbol{B},\boldsymbol{K},[\Uparrow])}$, the state *Deceived* $_{On\ \psi}$ is defined by the conjunction:

$$\textit{Deceived}_{\ On\ \psi} := \ \psi \wedge \boldsymbol{B}_R \neg \psi$$

On this basis, I make the stipulation that *Raymond* is deceived simpliciter on April 1^{st} 1925, written *Deceived*, if there is (at least) one formula ψ on which *Raymond* is deceived. Let us write [Postulate] this stipulation that being deceived on at least one formula ψ, written *Deceived* $_{On\ \psi}$, implies to be deceived simpliciter, written *Deceived*:

$$\text{[Postulate]} \qquad \textit{Deceived}_{\ On\ \psi} \rightarrow \textit{Deceived}$$

Emile's morning announcement that D brings about *Raymond*'s *first state of deception*. By applying the default rule $D \Rightarrow (d^+ \wedge \neg d^-)$ to this announcement, *Raymond* computes the wrong conclusion $(d^+ \wedge \neg d^-)$ that he will be deceived only by commission on April 1^{st} 1925. From then on, *Raymond* believes that the formula $(d^+ \wedge \neg d^-)$ is true when in fact, this is not the case: $\neg(d^+ \wedge \neg d^-)$. So after his brother's announcement and until his brother's explanation, *Raymond* holds the false belief that $(d^+ \wedge \neg d^-)$ and his state of deception can be described by the conjunction:

$$\textit{Deceived}_{\ On\ (d^+ \wedge \neg d^-)} := \ \neg(d^+ \wedge \neg d^-) \wedge \boldsymbol{B}_R(d^+ \wedge \neg d^-)$$

This first state of deception vanishes right before *Emile*'s explanation. *Raymond* realizes that he has not been deceived by commission. Then, a *second state of deception* substitutes the first. After having waited all day long to be deceived by commission, his mother asks *Emile* why he has not deceived *Raymond* by commission. This way, *Raymond* confesses that he does not believe that he could be deceived (only) by omission. Since he will be deceived as such (that's the sense of *Emile*'s explanation), formula $(\neg d^+ \wedge d^-)$ is true and *Raymond* falls in a *second state of deception* expressed by the conjunction:

$$Deceived_{\ On\ (\neg d^+ \wedge d^-)} := (\neg d^+ \wedge d^-) \wedge \boldsymbol{B}_R \neg(\neg d^+ \wedge d^-)$$

To sum up, *Raymond* goes through *two states of deception* in Smullyan's tale. First, he is deceived on the fact that he will be deceived only by commission: $Deceived_{\ On\ (d^+ \wedge \neg d^-)}$. Then, he is deceived on the fact that he will be deceived only by omission: $Deceived_{\ On\ (\neg d^+ \wedge d^-)}$. While the first state of deception concerns a *content*, namely the content that can be infered from *Emile*'s announcement that D, the second state of deception concerns a *type of action*, namely the kind of strategy that can be used to trick one on April 1^{st} 1925. It will prove helpful to keep this distinction in mind from now on.

The first state of deception is made possible by the default rule $D \Rightarrow (d^+ \wedge \neg d^-)$ *Raymond* mistakenly applies after hearing *Emile*'s announcement that D. The second state of deception is triggered by the fact that *Raymond* does not realize that he will be deceived only by omission after realizing that he has not been deceived (only) by commission. The next subsection is devoted to studying the events leading to those states of deception.

2.3.3 The Events Leading to Raymond's States of Deception

Properly speaking, the events that lead to those states are not "*actions*" made by *Emile* for deceiving *Raymond* since *Emile* does not perform any action to deceive his brother on April 1^{st} 1925. Those events are better conceived as epistemic "*reactions*" from *Raymond* corresponding to the distinct ways his mental states evolve throughout the day. These mental reactions are captured by belief radical upgrades with formulas φ, written $[\Uparrow \varphi]$. More precisely, *Emile* is said to deceive *Raymond* on a formula ψ by a formula φ, written $Deceive^{\ By\ \varphi}_{On\ \psi}$, if and only if ψ is *true* (*false*) and after upgrading his beliefs with formula φ, *Raymond* comes to believe that ψ is *false* (*true*). So for some formulas formulas $\varphi, \psi \in \mathcal{L}_{(\boldsymbol{B},\boldsymbol{K},[\Uparrow])}$, the event $Deceive^{\ By\ \varphi}_{On\ \psi}$ is defined by the conjunction:

$$Deceive^{\ By\ \varphi}_{On\ \psi} := \psi \wedge [\Uparrow \varphi]\boldsymbol{B}_R \neg\psi$$

Mental reactions with formulas φ lead *Raymond* to be deceived on a formula ψ ($Deceived_{\ On\ \psi}$). So, the actions $Deceive^{\ By\ \varphi}_{On\ \psi}$ can be defined from formula $Deceived_{\ On\ \psi}$ as follows:

$$Deceive^{\ By\ \varphi}_{On\ \psi} := [\Uparrow \varphi] Deceived_{\ On\ \psi}$$

In order to simplify the expression of the first deception that takes place in Smullyan's tale, I will use the abbreviation "*Default*" to name the default rule $D \Rightarrow (d^+ \wedge \neg d^-)$ *Raymond* applies:

$$Default \ := \ D \Rightarrow (d^+ \wedge \neg d^-)$$

Raymond's *first state of deception* is that he wrongly believes that he will be deceived only by commission on April 1st 1925: $\neg(d^+ \wedge \neg d^-) \wedge \boldsymbol{B}_R(d^+ \wedge \neg d^-)$. As I said, this state results from the *default rule* he applies to *Emile*'s announcement that D. So, the event that leads *Raymond* to be deceived a first time can be captured by an upgrade with formula $(D \wedge Default)$:

$$Deceive^{\ By\ (D \wedge Default)}_{On\ (d^+ \wedge \neg d^-)} \ = \ \neg(d^+ \wedge \neg d^-) \wedge [\Uparrow (D \wedge Default)] \boldsymbol{B}_R(d^+ \wedge \neg d^-)$$

Or to put it more simply:

$$Deceive^{\ By\ (D \wedge Default)}_{On\ (d^+ \wedge \neg d^-)} \ = \ [\Uparrow (D \wedge Default)] Deceived_{\ On\ (d^+ \wedge \neg d^-)}$$

Late in the afternoon, *Raymond* realizes that he has been deceived on the fact that he would be deceived only by commission. But though realizing that, he does not come to conclude that he will be deceived only by omission. In other words, he does not draw the right conclusion that $(\neg d^+ \wedge d^-)$ from learning that $\neg(d^+ \wedge \neg d^-)$. In terms of belief upgrade, this learning can be captured by a belief radical upgrade with formula $\neg(d^+ \wedge \neg d^-)$. We can now sum up the event that leads to *Raymond*'s *second state of deception*:

$$Deceive^{\ By\ \neg(d^+ \wedge \neg d^-)}_{On\ (\neg d^+ \wedge d^-)} \ = \ (\neg d^+ \wedge d^-) \wedge [\Uparrow \neg(d^+ \wedge \neg d^-)] \boldsymbol{B}_R \neg(\neg d^+ \wedge d^-)$$

Or to put it more simply:

$$Deceive^{\ By\ \neg(d^+ \wedge \neg d^-)}_{On\ (\neg d^+ \wedge d^-)} \ = \ [\Uparrow \neg(d^+ \wedge \neg d^-)] Deceived_{\ On\ (\neg d^+ \wedge d^-)}$$

2.3.4 *Emile*'s Whole Deceptive Plot on April Fool's Day

Emile's whole deceptive trick can be expressed by embedding the content of $Deceive^{\ By\ \neg(d^+\wedge\neg d^-)}_{\ On\ (\neg d^+\wedge d^-)}$ into the one of $Deceive^{\ By\ (D\wedge Default)}_{\ On\ (d^+\wedge\neg d^-)}$. The result is given by formula *Deception* which describes what *Emile*'s deception comes down to:

$$\begin{aligned} Deception \ = \ & [\Uparrow (D \wedge Default)]\Big(Deceived_{\ On\ (d^+\wedge\neg d^-)} \\ & \wedge\, [\Uparrow \neg(d^+ \wedge \neg d^-)]Deceived_{\ On\ (\neg d^+\wedge d^-)}\Big) \end{aligned}$$

So far, I have shown that *Emile*'s announcement naturally triggered a pragmatic interpretation (*deception only by commission*) that was false. But I have explained that learning the falsity of such a content did not help *Raymond* infer the right strategy *Emile* would use to deceive him on April 1^{st} (*deception only by omission*). In a broader sense, *Raymond* has been deceived because he did not reach the exact sense that *Emile*'s announcement implied. He did not reach the true semantic content that D encompassed. After hearing *Emile*'s announcement, he reached a pragmatic conclusion that he mistakenly equated with D's literal meaning and that turned out to be false. This conclusion being false, he had been tricked.

Later on that day comes *Emile*'s explanation. *Raymond*'s mother asks *Emile* why he has not deceived *Raymond* (at least, as he expected). As a result, *Emile* unveils the trap that led his brother to be deceived. In some sense, *Emile*'s explanation acts as a way of teaching *Raymond* the exact meaning of D. To make things clearer, I will use dynamic plausibility models in the next section to describe *Emile*'s explanation step-by-step.

2.4 Unveiling the Deception

2.4.1 *Emile*'s Explanation on Deception

As I said, *Emile*'s explanation can be conceived as a way of teaching *Raymond* the deceptive strategy he has been preyed to. By using epistemic plausibility models to represent the dynamics of *Emile*'s explanation, I aim to put more emphasis on the reasons why *Raymond* is surprised on April 1^{st} 1925. Before modelling these

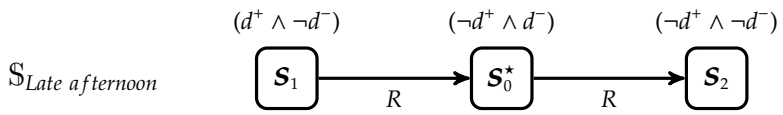
dynamics, let us describe the epistemic landscape of the day. The set of all possible states $\mathcal{S}$ is:

$$\mathcal{S} = \{s_0^{\star\,(\neg d^+ \wedge d^-)}, s_1^{(d^+ \wedge \neg d^-)}, s_2^{(\neg d^+ \wedge \neg d^-)}, s_3^{(d^+ \wedge d^-)}\}$$

The upperscripted formula means that the formula holds at the state. For instance, formula $(\neg d^+ \wedge d^-)$ holds at the pointed state $s_0^\star$ because *Raymond* will be deceived only by omission in the actual state. Note that *Raymond* will be deceived in states $s_0^\star$, s_1 and s_3 but not in state s_2 which corresponds to the failure of *Emile*'s announcement $(\neg D)$. As I said (see subsection 2.2.3), the combination $(d^+ \wedge d^-)$ encodes a type of deception which is irrelevant in the context of Smullyan's story. Neither *Emile* nor *Raymond* takes this combination as a plausible deceptive option. For this reason, I set aside state $s_3^{(d^+ \wedge d^-)}$ in my modelization.

Graphically, to represent the fact that the state t is *at least as plausible as* the state s for *Raymond* ($s \leqslant_R t$), I draw a subscripted right arrow "$\underset{R}{\rightarrow}$" from state s to state t. When states s and t are *equally plausible* for *Raymond* ($s \leqslant_R t$ and $t \leqslant_R s$), I draw a subscripted left-right arrow "$\underset{R}{\leftrightarrow}$" between states s and t. Reflexive arrows are omitted at each state to simplify the pictures. Let us now model the different steps of the story.

Late in the afternoon, *Emile*'s deceptive entreprise is achieved but *Raymond* has still not gotten the trick. On the one hand, he now believes that the default conclusion he has derived in the morning is false: $B_R\neg(d^+ \wedge \neg d^-)$. But on the other hand, he does not conclude that he will be deceived only by omission: $B_R\neg(\neg d^+ \wedge d^-)$. Actually, *Raymond* believes that the most plausible option is that he won't be deceived at all: $B_R(\neg d^+ \wedge \neg d^-)$. Here is the epistemic model representing *Raymond*'s beliefs at this stage:



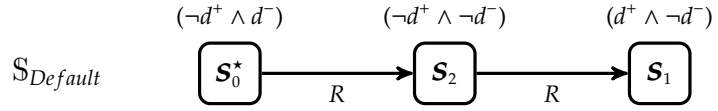
Model $\mathbb{S}_{Late\ afternoon}$ shows that *Raymond* is no longer deceived on the false formula $(d^+ \wedge \neg d^-)$ since he now believes this formula to be false: $\mathbb{S}_{late...}, s_0^\star \models$

$\neg$*Deceived* $_{On\ (d^+ \wedge \neg d^-)}$. But model $\mathbb{S}_{Late\ afternoon}$ also shows that *Raymond* is still deceived on the true formula $(\neg d^+ \wedge d^-)$ since he believes this formula to be false: $\mathbb{S}_{Late...}, s_0^\star \models$ *Deceived* $_{On\ (\neg d^+ \wedge d^-)}$. Immediately after *Raymond* complains to his mother that *Emile* has not deceived him (at least as he expected). So next comes *Emile*'s explanation that teaches *Raymond why* and *how* he has been deceived by his brother. The first part of *Emile*'s explanation is the following one:

Emile: "*So, you expected me to deceive you, didn't you?*

Raymond: *Yes.*"

Here *Emile*'s question can be understood as "*So, you expected me to deceive you by some action, didn't you?*". Since this question is in the affirmative form, it can be formally expressed by *Raymond* upgrading his beliefs with formula $\boldsymbol{B}_R(d^+ \wedge \neg d^-)$ corresponding to the belief he reached by default interpretation. The matching operation $[\Uparrow \boldsymbol{B}_R(d^+ \wedge \neg d^-)] : \mathbb{S}_{Late...} \to \mathbb{S}_{Default}$ returns the following model:

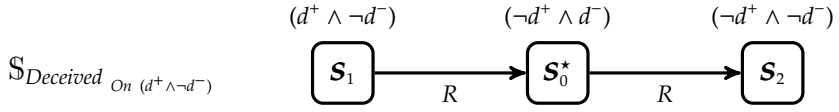


The arrows being reflexive, we can easily verify that $\mathbb{S}_{Default}, s_0^\star \models \boldsymbol{B}_R\boldsymbol{B}_R(d^+ \wedge \neg d^-)$. Here *Raymond* acknowledges that he came to believe that $(d^+ \wedge \neg d^-)$ after his brother's announcement that D. Immediately, though, *Emile* informs him that this default interpretation was wrong:

Emile: "*But I didn't, did I?*

Raymond: *No.*"

By those lines, *Emile* invites his brother to retract his belief in formula $(d^+ \wedge \neg d^-)$. The dynamic operation corresponding to such a removal can be expressed by $[\Uparrow \neg(d^+ \wedge \neg d^-)] : \mathbb{S}_{Default} \to \mathbb{S}_{Deceived\ On\ (d^+ \wedge \neg d^-)}$. Graphically, the model we obtain is:



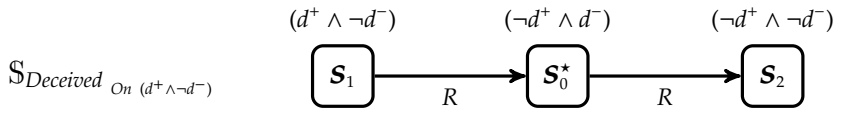
In model $\mathbb{S}_{Deceived\ On\ (d^+\wedge\neg d^-)}$, *Raymond* now believes that formula $(d^+ \wedge \neg d^-)$ is false: $\mathbb{S}_{Deceived\ On\ (d^+\wedge\neg d^-)}, s_0^\star \models \boldsymbol{B}_R\neg(d^+ \wedge \neg d^-)$. By doing so, he acknowledges that he was wrong to believe that $(d^+ \wedge \neg d^-)$ after his brother's announcement and thus, that he has been deceived on $(d^+ \wedge \neg d^-)$. But *Emile*'s demonstration goes even further. In the remaining part of his explanation, he shows that not only has *Raymond* been deceived on formula $(d^+ \wedge \neg d^-)$ but he has also been deceived on formula $(\neg d^+ \wedge d^-)$. The reason is that *Raymond* kept dismissing the true fact that he would be deceived only by omission $(\neg d^+ \wedge d^-)$ after realizing, as he just did, that he would not be deceived only by commission $\neg(d^+ \wedge \neg d^-)$. *Emile* starts explaining why in the following lines:

Emile: "*But you expected me to, didn't you?*

Raymond: *Yes.*"

Here *Emile* repeats the same question he has asked before ("*So, you expected me to deceive you, didn't you?*") to produce a second-order thought in his brother's mind. This repetition, and the way the question is introduced by the contrastive word "*But*", tend to generate a meta-level interpretation which can be translated by reformulating *Emile*'s question: "*But you kept dismissing that you could be deceived only by omission, didn't you?*". That way, *Raymond* is invited to realize that he kept considering deception only by omission as an implausible option even though he knew that he could not be deceived only by commission.

This recognition can be captured by *Raymond* upgrading his beliefs with the doxastic formula $\boldsymbol{B}_R\neg(\neg d^+ \wedge d^-)$ expressing his initial disbelief to be deceived only by omission. From a modeling perspective, however, this recognition leaves the model $\mathbb{S}_{Deceived\ On\ (d^+\wedge\neg d^-)}$ as it is since the formula $\boldsymbol{B}_R\neg(\neg d^+ \wedge d^-)$ is a validity of the model. Then, the corresponding operation $[\Uparrow \boldsymbol{B}_R\neg(\neg d^+ \wedge d^-)]$ is simply vacuous. The model is still:

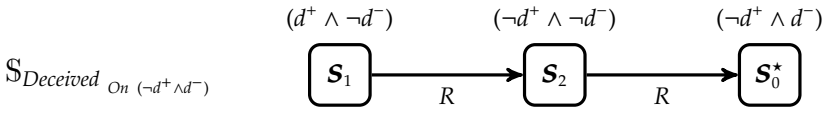


Immediately, though, *Emile* informs *Raymond* that his disbelief was unfortunate and led him to be deceived a second time:

Emile: "*So I deceived you, didn't I?*

Raymond: *Yes.*"

By those lines, *Emile* invites *Raymond* to acknowledge the truth of formula $(\neg d^{+} \wedge d^{-})$, namely to recognize that he has been deceived only by omission. The corresponding operation is $[\Uparrow (\neg d^{+} \wedge d^{-})] : \mathbb{S}_{Deceived\ On\ (d^{+} \wedge \neg d^{-})} \rightarrow \mathbb{S}_{Deceived\ On\ (\neg d^{+} \wedge d^{-})}$ and returns the final model:



Through the latter upgrade, *Raymond* comes to believe that formula $(\neg d^{+} \wedge d^{-})$ is true: $\mathbb{S}_{Deceived\ On\ (\neg d^{+} \wedge d^{-})}, s_0^{\star} \models \boldsymbol{B}_R(\neg d^{+} \wedge d^{-})$. Now he has adequate beliefs on the situation (since $s_0^{\star\ (\neg d^{+} \wedge d^{-})}$), and he also acknowledges that he has been deceived on this very formula all along.

I have described in detail the different steps by which *Raymond* learns why and how he has been fooled by his brother on April 1st 1925. By going backward from model $\mathbb{S}_{Deceived\ On\ (\neg d^{+} \wedge d^{-})}$ to the initial model $\mathbb{S}_{Late...}$, we can reconstitute the whole formula capturing the dynamics of *Emile*'s explanation. Let us start out by listing the four distinct steps of the dialogue:

(1) $[\Uparrow \boldsymbol{B}_R(d^{+} \wedge \neg d^{-})]$

(2) $[\Uparrow \neg(d^{+} \wedge \neg d^{-})]$

(3) $[\Uparrow \boldsymbol{B}_R \neg(\neg d^{+} \wedge d^{-})]$

(4) $[\Uparrow (\neg d^{+} \wedge d^{-})]$

Formulas **(1)-(2)-(3)-(4)** correspond to the distinct steps by which *Raymond* comes to learn the deception *Emile*'s announcement implied. At step **(1)**, he learns that after hearing that D, he made the default interpretation that this announcement

meant that he would be deceived only by commission: $[\Uparrow \boldsymbol{B}_R(d^+ \wedge \neg d^-)]$. But at step **(2)**, *Emile* reveals that this default interpretation was wrong since, as *Raymond* realized by himself late in the afternoon, he had not been deceived (only) by commission: $[\Uparrow \neg(d^+ \wedge \neg d^-)]$. By concatenating steps **(1)** and **(2)**, *Raymond* can conclude that he has been deceived on the pragmatic meaning of the announcement: $Deceived_{\ On\ (d^+ \wedge \neg d^-)}$. At step **(3)**, *Emile* tells *Raymond* that though realizing that he had been deceived on this pragmatic meaning, he kept dismissing that he could be deceived only by omission: $[\Uparrow \boldsymbol{B}_R \neg(\neg d^+ \wedge d^-)]$. But at step **(4)**, *Emile* finally informs *Raymond* that since he has done nothing to deceive him throughout the day, the right conclusion he should have reached is that he would be deceived only by omission: $[\Uparrow (\neg d^+ \wedge d^-)]$. By concatenating steps **(3)** and **(4)**, *Raymond* can conclude that he has been deceived on the type of action *Emile* would use to trick him: $Deceived_{\ On\ (\neg d^+ \wedge d^-)}$.

By learning the steps above (see Table 2.1 for a summary), *Raymond* can reconstitute *Emile*'s whole deceptive plot on April 1^{st} 1925:

$$\begin{aligned} Deception \ = \ & [\Uparrow (D \wedge Default)]\Big(Deceived_{\ On\ (d^+ \wedge \neg d^-)} \\ & \wedge\, [\Uparrow \neg(d^+ \wedge \neg d^-)] Deceived_{\ On\ (\neg d^+ \wedge d^-)}\Big) \end{aligned}$$

To conclude, *Emile*'s explanation can be conceived as a way of teaching *Raymond* the deceptive plot he has been preyed to. Through the learning process, *Raymond* understands *why* and *how* he has been deceived by his malicious brother. But I will show that this learning also throws *Raymond* into paradoxical thoughts.

2.4.2 *Raymond*'s Self-Referential Reasoning

After his brother's explanation, *Raymond* lies in bed and starts reasoning about the deception he has been preyed to. He eventually falls in a paradox which can be summed up as follows:

> *On the one hand, supposing that he wasn't deceived, then he didn't get what he expected (because he expected to be deceived after his brother's announce-*

Explanation	Dynamic Step	Deceptive State
E: "*So you expected...?* **R:** *Yes.*"	$[\Uparrow \boldsymbol{B}_R(d^+ \wedge \neg d^-)]$	...
E: "*But I didn't...?* **R:** *No.*"	$[\Uparrow \neg(d^+ \wedge \neg d^-)]$	*Deceived* $_{On\ (d^+ \wedge \neg d^-)}$
E: "*But you expected...?* **R:** *Yes.*"	$[\Uparrow \boldsymbol{B}_R \neg(\neg d^+ \wedge d^-)]$	...
E: "*So I deceived you...?* **R:** *Yes.*"	$[\Uparrow (\neg d^+ \wedge d^-)]$	*Deceived* $_{On\ (\neg d^+ \wedge d^-)}$

Table 2.1: *Emile*'s Explanation Flow.

ment!), and hence he was actually deceived. But on the other hand, supposing that he was deceived, then he exactly did get what he expected, and hence he was not "deceived."

I will now express *Raymond*'s reasoning in more formal terms. Actually, only the first line of his reasoning ("On the one hand...") leads to a contradiction. The second line ("On the other hand...") is not contradictory as it stands. Let us start with the second line before studying the first in details.

Raymond's second line of reasoning can be reformulated as such: if *Raymond* supposes that he is deceived by *Emile*, he has to suppose that the formula *Deceived* is true after having processed *Emile*'s announcement: $[\Uparrow (D \wedge Default)]Deceived$.

But then, *he exactly did get what is expected* (because he expected to be deceived after his brother's announcement): $[\Uparrow (D \wedge Default)]\boldsymbol{B}_R Deceived$. From those starting points, the following derivation shows that *Raymond* can finally conclude that he is not deceived *on the fact that he is deceived*: $[\Uparrow (D \wedge Default)]\neg Deceived_{\ On\ Deceived}$. But from this conclusion, *Raymond* cannot conclude that he is not deceived *at all*. Doing so would be logically incorrect: it would consist in denying the antecedent in the (stipulative) statement [Postulate]: $Deceived_{\ On\ Deceived} \rightarrow Deceived$. However, *Raymond* does so when he concludes that he is not deceived from the assumption that he is. Let us examine his reasoning to point out the guilty step:

[1]$[\Uparrow (D \wedge Default)]Deceived$	[Hypothesis].
[2]$[\Uparrow (D \wedge Default)]\boldsymbol{B}_R Deceived$	[Hypothesis].
[3]$[\Uparrow (D \wedge Default)]Deceived \wedge [\Uparrow (D \wedge Default)]\boldsymbol{B}_R Deceived$	[1]-[2], Conjunction.
[4]$[\Uparrow (D \wedge Default)](Deceived \wedge \boldsymbol{B}_R Deceived)$	[3], Reduction Axiom.
[5]$[\Uparrow (D \wedge Default)]\neg Deceived_{\ On\ Deceived}$	[4], Failure of $Deceived_{\ On\ \ldots}$.
[6]$\neg Deceived_{\ On\ Deceived} \rightarrow \neg Deceived$	[5], By [Postulate].
[7]$[\Uparrow (D \wedge Default)](\neg Deceived_{\ On\ Deceived} \rightarrow \neg Deceived)$	[6], Necessitation rule.
[8]$[\Uparrow (D \wedge Default)]\neg Deceived_{\ On\ Deceived} \rightarrow [\Uparrow (D \wedge Default)]\neg Deceived$	[7], Axiom K.
[9]$[\Uparrow (D \wedge Default)]\neg Deceived$	[5]-[8], MP. **Contra [1].**

In the derivation above, the guilty step is **[6]**: it is logically incorrect to derive that $\neg Deceived_{\ On\ Deceived} \rightarrow \neg Deceived$ by denying the antecedent in formula $Deceived_{\ On\ Deceived} \rightarrow Deceived$. For this reason, the conclusion *Raymond* derives is wrong: he cannot logically conclude that after having processed his brother's announcement that D, he is deceived *only if* he is not deceived. So this line in *Raymond*'s reasoning does not lead to a genuine contradiction. Concerning the first, I will now point out that it does lead to a genuine contradiction since one of the assumptions (viz. that *Raymond* is not deceived after his brother's announcement) is false. Let us examine this reasoning.

Raymond's first line of reasoning can be reformulated as follows: on the one

hand, if *Raymond* supposes that he is not deceived by *Emile*, he has to suppose that formula $\neg$*Deceived* is true after processing his brother's announcement: $[\Uparrow (D \wedge Default)]\neg Deceived$. But then, *he did not get what he expected* because he clearly expected to be deceived after his brother's announcement: $[\Uparrow (D \wedge Default)]\boldsymbol{B}_R Deceived$. We can now derive a contradiction by applying the following principles:

[1]$[\Uparrow (D \wedge Default)]\neg Deceived$	[Hypothesis].
[2]$[\Uparrow (D \wedge Default)]\boldsymbol{B}_R Deceived$	[Hypothesis].
[3]$\neg[\Uparrow (D \wedge Default)]Deceived$	[1], Reduction Axiom.
[4]$[\Uparrow (D \wedge Default)]\neg Deceived \wedge [\Uparrow (D \wedge Default)]\boldsymbol{B}_R Deceived$	[1]-[2], Conjunction.
[5]$[\Uparrow (D \wedge Default)](\neg Deceived \wedge \boldsymbol{B}_R Deceived)$	[4], Reduction Axiom.
[6]$[\Uparrow (D \wedge Default)]Deceived_{\ On\ \neg Deceived}$	[5], Def. of *Deceived* $_{On\ \ldots}$.
[7]$Deceived_{\ On\ \neg Deceived} \rightarrow Deceived$	By [Postulate].
[8]$[\Uparrow (D \wedge Default)](Deceived_{\ On\ \neg Deceived} \rightarrow Deceived)$	[7], Necessitation rule.
[9]$[\Uparrow (D \wedge Default)]Deceived_{\ On\ \neg Deceived} \rightarrow [\Uparrow (D \wedge Default)]Deceived$	[8], Axiom K.
[10]$[\Uparrow (D \wedge Default)]Deceived$	[6]-[9], MP. **Contra [3].**

But though perfectly sound, *Raymond*'s first line of reasoning is based on a false premise, namely **[1]**: $[\Uparrow (D \wedge Default)]\neg Deceived$. It is false that *Raymond* is not deceived after computing the default interpretation of his brother's morning announcement. We saw in the previous subsection that *Raymond* is deceived on formula $(d^+ \wedge \neg d^-)$ after this computation since this default interpretation is wrong: $[\Uparrow (D \wedge Default)]Deceived_{\ On\ (d^+ \wedge \neg d^-)}$. Due to the stipulation we assumed to hold, one who is deceived on a specific formula after *Emile*'s announcement is deceived simpliciter. So the formula $[\Uparrow (D \wedge Default)]Deceived$ is also true. This way, the contradiction with conclusion **[10]** simply vanishes.

To sum up, the puzzling thoughts *Raymond* has late on the evening are not of a paradoxical nature. In his first line of reasoning, the assumption that he is

not deceived after his brother's announcement can be suspended. In the second line of reasoning, the step leading to the contradiction can be blocked. Of more interest is the state of surprise *Raymond* is led to by his brother's explanation. Such a surprise is noticeable in *Raymond*'s late reasoning. He is surprised both by the *deceptive content* and the *type of deception Emile*'s announcement involved. In the next section, I explain why *Raymond* is surprised as such at the end of his brother's explanation.

2.5 A Source of Surprise

2.5.1 *Emile*'s Surprise Announcement of Deception

From now on, let us go back to *Emile*'s early announcement: "Well, *Raymond*, today is April Fool's Day, and I will deceive you as you have never been deceived before!". We have shown that this announcement pragmatically suggests the false conclusion that *Raymond* would be deceived only by commission. But what does this announcement really mean? What is the exact meaning of *Emile*'s announcement? How can it be expressed in natural and more formal terms? To give an answer, I introduce a famous epistemic puzzle called the 'The Surprise Examination Paradox' in the literature [e.g. O'Connor 1948, Scriven 1951, Shaw 1958]. In the latter, a teacher announces to his or her students that he or she will give a (single) surprise exam on the next week. The way it is formulated, the teacher's announcement of surprise can be compared with *Emile*'s announcement along similar lines. Both announce a surprising event to come and both raise similar issues concerning the possibility for the announcement to be fulfilled. In this section, we focus on the *two-day* version of the Surprise Exam Paradox and provide a formal expression of *Emile*'s announcement based on the formalizations Gerbrandy [1999 2007] and van Ditmarsch & Kooi [2006] have given to the teacher's announcement in the Surprise Exam case.[2]

In the *two-day* version of the Surprise Exam Paradox, the teacher makes the fol-

[2] See also van Benthem [2004] and Bonnay & Égré [2011] on discussions of related epistemic paradoxes in dynamic logic settings (Fitch's paradox, Moore's paradox and Williamson's Margin for Error paradox).

lowing suprise announcement: "I will give you a (single) surprise exam either on Monday or on Tuesday next week". Between the teacher and his or her students, it is common knowledge that an exam comes as a surprise if the students do not know the evening before the exam that it will happen the next day. But after this announcement of a surprise exam, the students (who are assumed to be perfect reasoners) start reasoning and finally conclude that such an exam is impossible. Logically speaking, the exam would fail to be a surprise on any day of the week. Their argument is based on a type of reasoning called *backward induction*. If the test takes place on Tuesday, the students would know on Monday evening that the test is on Tuesday because Tuesday is the last available working day of the week. Then, it would not be a surprise. But nor will it be a surprise if the test is on Monday because the students would know on Sunday evening that the test is on Monday (for they know that the test is not on Tuesday due to the previous reasoning). At the end of this backward reasoning, the students would conclude that the test cannot be a surprise on any day of the week. Unfortunately, the teacher unexpectedly gives the exam on Tuesday and this is a complete surprise for the students.

Let us reformulate *Emile*'s early announcement. If we consider the *deceptive options* he has, a more explicit formulation of his announcement might be: "I will surprise you by deceiving you *either* only by commission *or* only by omission". We know that *Raymond* misinterprets this explicit announcement in the following sense: "I will surprise you by deceiving you only by commission". But from a semantic point of view, both deceptive options (namely *deception only by commission* and *deception only by omission*) should be taken into account to capture the exact meaning of *Emile*'s announcement.

This informal formulation of *Emile*'s announcement can inspire many formal expressions depending on the way we capture the concept of *surprise*. Following a basic intuition we will elaborate on later (see subsection 2.5.3 for details), a state of surprise is generally triggered by a *mismatch of beliefs*: facts or events come out to be true but remain unacknowledged by agents until they come out to be true. As a result, the latter agents undergo states of *surprise*, or even of *astonishement*,

when they learn the existence of those facts and events. To see more clearly into this, let us first give the formal expression of the announcement we favor:

$$\begin{aligned} D \;=\; & ((d^+ \wedge \neg d^-) \wedge \boldsymbol{B}_R \neg (d^+ \wedge \neg d^-)) \\ & \vee ((\neg d^+ \wedge d^-) \wedge [\Uparrow \neg (d^+ \wedge \neg d^-)] \boldsymbol{B}_R \neg (\neg d^+ \wedge d^-)) \end{aligned}$$

This formal expression of *Emile*'s announcement is loosely based on Gerbrandy's formalization of the teacher's announcement in the Surprise Exam case.[3] Formula D encodes the fact that *Emile*'s deception will be a surprise for the reason that if *Raymond* is deceived only by omission, formula $(\neg d^+ \wedge d^-)$ is true but *Raymond* denies it (viz. $\boldsymbol{B}_R \neg (\neg d^+ \wedge d^-)$) even after having recognized that he won't be deceived only by commission: $[\Uparrow \neg (d^+ \wedge \neg d^-)]$. But if *Raymond* is deceived only by commission $(d^+ \wedge \neg d^-)$, he denies it too: $\boldsymbol{B}_R \neg (d^+ \wedge \neg d^-)$. So no matter whether he is deceived only by commission or only by omission, in both cases it will come as a surprise according to D.

That being said, we can rewrite formula D by using the abbreviations we have used for *Emile*'s deceptive states. So formula D becomes:

$$D \quad = \quad \textit{Deceived}_{\,\textit{On}\ (d^+ \wedge \neg d^-)} \vee [\Uparrow \neg (d^+ \wedge \neg d^-)] \textit{Deceived}_{\,\textit{On}\ (\neg d^+ \wedge d^-)}$$

Let us now remind the full deceptive action *Emile* put forward to trick his brother on April 1^{st} 1925:

[3] A (two-day) Gerbrandy's formalization of the teacher's announcement of surprise would be: $(\textit{mon} \wedge \neg \boldsymbol{K}\, \textit{mon}) \vee (\textit{tue} \wedge [!\neg \textit{mon}] \neg \boldsymbol{K}\, \textit{tue}) \vee \boldsymbol{K} \bot$, such that "*mon*" and "*tue*" stand for *monday* and *tuesday*, and "!" is some learning operation. Gerbrandy's formalization means that the exam will come as a surprise for the reason that if the exam is on tuesday (*tue*), the agent does not know this ($\neg \boldsymbol{K}\, \textit{tue}$) after having learned that it is not on monday ($[!\neg \textit{mon}]$); if the exam is on monday (*mon*), the agent does not know it either on monday ($\neg \boldsymbol{K}\, \textit{mon}$); if the agent has contradictory information about the date of the exam ($\bot$), the latter will also come as a surprise (according to Gerbrandy). This proposal has been restated by van Ditmarsch & Kooi [2006] in the following way: $(\textit{mon} \rightarrow \neg \boldsymbol{K}\, \textit{mon}) \wedge (\textit{tue} \rightarrow [!\neg \textit{mon}] \neg \boldsymbol{K}\, \textit{tue})$. Though van Ditmarsch & Kooi' formalization does not take inconsistency as a source of surprise, both forms opt for a *non-self-referential reading* of the teacher's announcement. Failure of the Success Axiom in Public Announcement Logic makes *self-referential sentences* of the Moorean type impossible to express in those settings. In my case, the formalization of *Emile*'s announcement also relies on a *non-self-referential reading* of the announcement. Self-referentiality is excluded since the announcement itself is not considered as a *bona fide* action (see subsection 2.2.3 for details).

$$Deception \;=\; [\Uparrow (D \wedge Default)]\Big(Deceived_{\;On\;(d^{+}\wedge\neg d^{-})} \wedge [\Uparrow \neg(d^{+} \wedge \neg d^{-})]Deceived_{\;On\;(\neg d^{+}\wedge d^{-})}\Big)$$

Setting aside the default interpretation *Raymond* makes of his brother's announcement, namely $[\Uparrow (D \wedge Default)]$, it is clear that being *deceived* the way he is, *Raymond* is also necessarily *surprised*. In language $\mathcal{L}_{(B,K,[\Uparrow])}$, the *conjunctive* subform in *Deception* implies the *disjunctive* formula D, namely $\vdash Deception \rightarrow D$. This is perfectly consistent with Smullyan's story: we do observe that at the end of the day, *Raymond* is surprised by the deception he has been preyed to. Crucially, the deception *Emile* has performed (viz. *Deception*) is the deception that has made *Raymond* being surprised.

Nonetheless, a problem immediately arises with this type of announcement. In the Surprise Exam case, it is usually accepted that the teacher's announcement is an "epistemic blindspot" for the students and that it cannot succeed *in principle* for this exact reason [see Sorensen 1988]. In the same way, I argue that *Emile*'s announcement cannot be fulfilled *in principle* and that assuming the opposite leads to a paradox. My argument is the following: if *Raymond* is a perfect reasoner, he has the (temporal) resources to conclude that *Emile*'s announcement is false, and then he cannot be surprised after this very announcement. But then, he necessarily ends up in a paradox since the announcement turns out to be true at the end of the day: he *is* surprised by his brother's explanation. So how can this be possible? How can *Raymond* be surprised without inconsistency ?

2.5.2 *Emile*'s Successful Action of Surprise

To put it briefly: *Emile*'s announcement of surprise is fulfilled only because *Raymond* is not a perfect reasoner. If *Raymond* was a perfect reasoner, his brother's announcement could not succeed, *Raymond* could not be surprised afterwards. For the sake of argument, let us suppose that *Raymond* is a perfect reasoner. If he now computes his brother's announcement by using the *default rule*, then, no matter whether he will be deceived only by commission ($d^{+} \wedge \neg d^{-}$) or will be deceived only by omission ($\neg d^{+} \wedge d^{-}$), in both cases it cannot be a surprise for him.

This relies on the fact that *Raymond* has the temporal resources to *adjust* his beliefs in time before looking for his brother to get more information. After *Emile*'s announcement, he can come to believe that he will be deceived only by commission. In case he is, it won't be a surprise since he would have expected to be deceived as such. But in case he is not deceived as such, it won't be a surprise either. The reason is that if he is not deceived only by commission, he will necessarily be deceived only by omission. Late in the afternoon and before going to see his brother, he would have had enough time to realize it and to come to rightly believe that he would be deceived only by omission. As a consequence, it cannot be a surprise either. In other words, no matter whether he will be deceived only by commission or only by omission, his brother's explanation will fail to be a surprise. *Emile*'s announcement cannot be fulfilled in principle. That being said, it is time to study the *distinct states of surprise Raymond* goes through late on that day.

2.5.3 *Raymond*'s Distinct States of Surprise

In Smullyan's tale, *Raymond*'s surprise does not follow from the completion of *Emile*'s deceptive plot but from his brother unveiling this very plot. The reason is that a cognitive agent reaches a state of surprise when he or she is led to recognize an *inconsistency*, that is a *discrepancy* or a *mismatch*, between what he or she believes about the world and what the actual world is [e.g. Meyer *et al.* 1997, Lorini & Castelfranchi 2006].

Following a basic intuition, an event which is unexpected is considered "surprising" and the more unexpected it is, the more surprising it turns out to be [e.g. Ortony & Partridge 1987, Meyer *et al.* 1997]. In other words, an event which was expected but does not happen is surprising. But an event which actually happens while being totally unexpected is even more surprising.

In that sense, Lorini & Castelfranchi distinguish two main kinds of surprise. The first kind is called "*mismatch-based surprise*" and results from a "*conflict between a perceived fact and a scrutinized representation*". The agent is surprised because he or she has some anticipatory representation of a fact or event, but he or she cannot

make the incoming data fit with this anticipatory representation they have. In that case, the intensity of the induced surprise depends of the probability, more crucially on the *implausibility*, which is assigned by the agent to the conflicting data he or she receives.

But Lorini & Castelfranchi contrast this first type of surprise with a stronger form named "*astonishment*" (or "*surprise in recognition*"). Contrary to the first, this form of surprise is of a *second-order nature* because it is rooted on the *recognition* of the implausibility of a perceived fact compared to expectations: "*I perceive a certain fact and recognize the implausibility of this*". Lorini & Castelfranchi also precise that this astonishment can depend upon two distinct mental processes. First, I can be astonished by a fact/event φ because I assigned a high probability to $\neg\varphi$ and after perceiving φ I realize that "*I would not have expected that event* φ" [see Lorini & Castelfranchi 2006, 3]. Second, I can be astonished by φ because perceiving φ leads me to infer the falsity (and incongruity) of my initial disbelief that φ or belief that $\neg\varphi$.[4]

In Smullyan's tale, both kinds of deeper surprise are involved because both cases of unexpectation are involved. On the one hand, *Raymond* expects to be deceived only by commission after his brother's announcement but later in the evening, he is led to notice that nothing has happened as such. In fact, *Emile* has not perfomed any particular action to deceive him. Due to this obvious discrepancy between his scrutinized representation (to be deceived only by commission) and the lack of any deceptive action, *Raymond* is preyed to a surprise of the mismatch-based kind.

But on the other hand, *Emile*'s explanation also leads *Raymond* to be *astonished* (or *surprised in recognition*). Indeed, *Emile* tells him that he has actually been deceived only by omission although he kept discarding such an option as being plausible (even after noticing that it was no longer plausible that he would be deceived only by commission). Since he judged deception only by omission as being very implausible, *Raymond* is astonished when he recognized that *Emile*

[4] Lorini & Castelfranchi contrast those "*deeper and slower forms of surprise which are due to symbolic representations of expected events*" [see Lorini & Castelfranchi 2006, 1] to a *first-hand surprise* which corresponds to a perceptual mismatch between a stimulus (viz. what the agents can *see* or *hear* in their immediate environment) and their sensory-motor expectations.

used this type of deception to trick him. Consistent with Lorini & Castelfranchi's distinction is the fact that *Raymond*'s *astonishment* is stronger in term of surprise than his first *mismatch-based surprise*. His astonishement is tied to his strongly entrenched disbelief that he could be deceived only by omission. In contrast, his belief that he would be deceived only by commission is less entrenched and only leads him to be moderately surprised.

2.6 Conclusion

Emile's trick relies on a *misleading default* and pertains to the broader category of *veridical deception*. Veridical deception concerns all the deceptive strategies whereby a non-cooperative speaker uses a *true* piece of information to induce a *false belief* in some intended addressee. Apart from misleading defaults, non-cooperative speakers can use other forms of veridical deception. For instance, they can use *false implicatures* to make someone falsely believe a piece of information φ by telling them a true information ψ while intending for them to infer that φ from ψ [e.g. Adler 1997, Meibauer 2005 2014]. But veridical deception also includes strategies like *pretending to deceive* [e.g. Vincent & Castelfranchi 1981, Fallis 2014]: to make someone disbelieve a true piece of information φ by arousing his or her suspicion towards the source of φ; or *presupposition faking* [e.g. Harder & Kock 1976, Vincent & Castelfranchi 1981]: to make someone inadequately believe a true formula φ he or she fails to semantically or pragmatically account for.

A I said, veridical deception is performed through a *misleading default* in Smullyan's story. *Emile*'s strategy relies on a *true* but *misleading* announcement. As a semantic content, *D* happens to be *true*: *Emile will deceive Raymond as he has never been deceived before*. But pragmatically, *D* leads *Raymond* to draw a *false* conclusion: contrary to what *Emile* suggests, *he won't deceive Raymond (only) by commission*. Though being false, however, this conclusion is based on a natural process. This is usually the case that one who is deceived is deceived by some action (*commission*) instead of by none (*omission*). So the interpretation *Raymond* makes is the natural interpretation one would make in common circumstances.

From a theoretical perspective, I have shown that the surprise deception paradox can be seen as a variant, and interesting illustration, of the *two-day* version of the Surprise Exam Paradox. But an important difference between the two puzzles is that the (projected) failure of the teacher's announcement only relies on logical reasoning, whereas the (projected) failure of *Emile*'s announcement relies both on logical and pragmatic reasoning. *Emile*'s trick succeeds because *Raymond* first makes the wrong pragmatic inference that *deception* implies *deception by commission*. This inference makes him logically, and falsely, conclude that being not deceived by commission, he is not deceived simpliciter. If *Emile* was cynical enough, his last word and final recommendation to his brother could be: "*If you want to deceive someone else, whether or not you do so actively, start with pragmatics*!"

This chapter was devoted to investigating *deception by omission* provoked by a *misleading default inference*. Contrary to the *standard case* of lies (Chapter 1), misleading defaults and strategic omissions are *non-standard cases of deception*, and have received less attention from epistemologists working on deception. Chapter 3 aims to integrate these standard and non-standard cases into a definitional account of informational messages.

Chapter 3

The Definition of Intelligence Messages

3.1 Introduction

One of the activities provided by intelligence agencies is called "Information Evaluation" and consists in assessing "*an item of information in respect of the reliability of the source, and the credibility of the information*" [see STANAG-2511 2003, DIA-2 2010]. A competent authority also known as *intelligence officer* performs the evaluation along a 6×6 alphanumeric matrix based on 2 ratings ("Content Credibility" vs. "Source Reliability") with 6 levels of the rating Credibility and 6 levels of the rating Reliability. Both ratings are then crossed to give an overall insight into the quality of sources and contents.

But the alphanumeric scale has been criticized for being based on an assumption we may call "*fact vs. interpretation*" and such that "*intelligence reports transmit facts and/or assessments. The distinction between fact and interpretation must always be clearly indicated*" [see STANAG-2511 2003, 2]. As a matter of fact, experimental results show that intelligence officers fail to comply with this distinction between objective facts and subjective interpretations when they evaluate intelligence data [e.g. Baker *et al.* 1968, Kelly & Peterson 1971, Johnson 1973]. A straightforward explanation is that the scale is built on confused notions that should be clarified and better specified. Researchers have observed that officers misunderstand the

dimensions of credibility and reliability that underlie the scale, their distinct levels of evaluation, as well as resultant scores.

But to help officers better separate facts from interpretations, it seems imperative to first point out the objective-subjective divide that remains implicit in the existing scale. My hypothesis is that this divide is nothing over and beyond the distinction between *descriptive* and *evaluative* terms. Basically, credibility and reliability are *evaluatives* that are used for assessing *descriptive* terms of truth for message contents, and of honesty for message sources. More precisely, credibility and reliability are used to elicit *degrees* of truth for contents and *degrees* of honesty for sources. In the first case, *contextual evidence* is used to assess contents while in the second case, *past experience* with the source is used to determine the source reliability.

Once elicited, degrees of truth and degrees of honesty are *descriptive features* of intelligence messages. They are objective dimensions that *define* intelligence messages and can be used for isolating distinct *categories*, or *types*, for them. But these dimensions remain implicit in the existing scale and spotting types is tedious, especially concerning deceptive messages. No clear link can be drawn from the existing scale to *necessary* and *sufficient* conditions commonly used to frame informational types: *information*, *misinformation*, *lie*, *omission*, etc. Lies are notoriously difficult to rate along the scale [Capet & Revault d'Allonnes 2013], but even more difficult is spotting mischievous messages that involve omission of information.

In this chapter, I present a descriptive account to see more clearly into informational types. I propose a 3×3 matrix that distinguishes 3 levels of Truth for contents ("*True*", "*False*", "*Indeterminate*") and 3 levels of Honesty for sources ("*Honest*", "*Dishonest*", "*Imprecise*"). This will help me isolate 9 categories of messages depending on the extent to which messages are *true* or *false* and sources *honest* or *dishonest*. Regarding deceptive messages, I mainly focus on categories that appeal to omission from the standpoint of *linguistic vagueness* [based on Égré & Icard 2018]. Even more than lies, omission poses an increased difficulty to intelligence

officers for omitting information always involves keeping intact the Maxims of Quality that would imply easy detection. As a result, spotting vague messages remains a critical issue for intelligence raters but I believe that a precise taxonomy of the phenomenon would be a crucial advance in that direction. Accordingly, I distinguish two main categories of vague messages: *semantic indeterminacy* and *pragmatic imprecision*, each of which with more specific varieties (semantic degree-vagueness and open-texture vs. pragmatic generality and approximation).

This chapter is structured as follows. In section 3.2, I start out by describing the wide range of activities intelligence agencies usually engage in (subsection 3.2.1). I then focus on the activity of *information evaluation* by describing the scale officers commonly use in this endeavour (subsection 3.2.2). I then insist on some virtues of the alphanumeric scale before pointing out its shortcomings for information evaluation (3.2.3).

My purpose in section 3.3 is to analyze the *fact vs. interpretation assumption* to show that this assumption is not respected by intelligence officers on the field (subsection 3.3.1). As I said, a straightforward explanation for this failure is that the evaluative dimensions of the scale, as well as their descriptions, are *confused*. Based on the empirical literature, I argue that the definitions of the credibility and reliability ratings, as well as their distinct levels of evaluation, lead to misunderstandings and inconsistencies amongst officers and between them (subsection 3.3.2). This hypothesis is valid. But making the *fact vs. interpretation assumption* efficient first requires to shed light on the objective dimensions of truth and honesty that remain in the background of the scale (subsection 3.3.3).

Section 3.4 is devoted to presenting my *descriptive account* in detail. I start out by explaining the 3×3 matrix I propose to *define* messages and to *isolate* their informational types (subsection 3.4.1). I first concentrate on *classical types* obtained when the content of the message is either *true* or *false* and the source, *honest* or *dishonest* (subsection 3.4.2). Following Égré & Icard [2018], I then put more emphasis on some *borderline types* that rely either on *semantic vagueness* — in case the status of the content is semantically indeterminate (subsection 3.4.3), or

on *pragmatic vagueness* — in case sources are less informative than they should according to the Gricean Maxims (subsection 3.4.4).

3.2 Information Evaluation in Intelligence

3.2.1 The Intelligence Cycle in a Nutschell

Intelligence agencies usually engage in a cyclic sequence of activities [TTA 2001, DIA-2 2010] consisting (roughly)[1] in four distinct stages (see Figure 3.1).

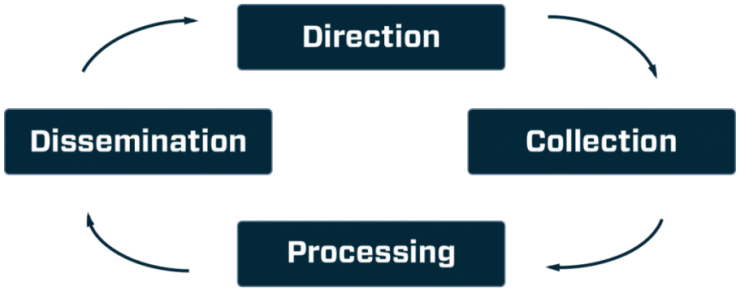


Figure 3.1: The Intelligence Cycle.

In the first stage called "*Direction*" (or "*Planning*"), a competent authority or decision maker determines the "*intelligence requirements*" of the agency, namely the kind of data the service needs on a specific matter of interest (that can be of a strategical, tactical or operational nature). These requirements are expressed by a "Request for Information" (RFI).

The stage of "*Collection*" (or "*Gathering*") immediately follows: in response to the requirements defined above, the intelligence organization recruits *sources* that are relevant on the topic of interest. These sources can have different *sensor* types: they can be *human sources* (HUMINT) if intelligence is collected from human agents on the field. There also exists many different types of technical

[1] To be fully accurate, we should distinguish *four-stages cyclic models* (on which the French model is based [DIA-2 2010]: "*Orientation*", "*Collection & Processing*", "*Analysis*" and "*Dissemination*") from *five-stages (or more) cyclic models* (on which the American model is based [ODN 2011]: "*Planning & Direction*", "*Collection*", "*Processing & Exploitation*", "*Analysis & Production*" and "*Dissemination & Integration*"). To keep things simple, the model I use in this chapter is based on *four stages*. This model is usually accepted as a minimal ground by the intelligence community.

sources like geospatial imagery (IMINT), measurement and signature intelligence (MASINT), interception of signals (SIGINT), cyber intelligence (CYBINT), analysis of monetary transactions (FININT), etc.

Then comes a third stage called "*Processing*" (or "*Exploitation*") in which the collected data are classified into categories of the same nature (*Grouping*), evaluated depending on the quality of their sources and contents (*Evaluation*), interpreted through the intelligence requirements that have been defined (*Analysis*) and eventually merged with other messages in an enriched framework (*Fusion*). This framework will be used for grounding future decisions (*Interpretation*).

In the last stage of "*Dissemination*", a conclusive report is delivered to the decision-maker that fits with the intelligence requirements. Recipients generally conduct further collection and exploitation stages which lead to new iterations of the intelligence cycle.

3.2.2 The Traditional Scale for Information Evaluation

Message Scores		***Content Credibility***					
		1	**2**	**3**	**4**	**5**	**6**
Source Reliability	**A**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**
	B	**B1**	**B2**	**B3**	**B4**	**B5**	**B6**
	C	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
	D	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**
	E	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**
	F	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**

Table 3.1: The 6×6 Traditional Alphanumeric Matrix.

In the present and next chapters, I focus on the *evaluation step* of the *processing stage*. This step proposes to assess "*an item of information in respect of the reliability of the source, and the credibility of the information*" [e.g. STANAG-2511 2003, FM-2-22.3 2003, DIA-2 2010].[2] This evaluation is made along a 6×6 alphanumeric matrix

[2] The doctrine precises that intelligence reports based on credibility and reliability ratings "*must*

Content Credibility	*Source Reliability*
1: Confirmed	**A:** Completely Reliable
2: Probably True	**B:** Usually Reliable
3: Possibly True	**C:** Fairly Reliable
4: Doubtfully True	**D:** Not Usually Reliable
5: Improbable	**E:** Unreliable
6: Cannot Be Judged	**F:** Cannot Be Judged

Table 3.2: Linguistic Labels for Ratings.

which is based on 2 ratings ("*Content Credibility*" vs. "*Source Reliability*) with 6 levels of the rating *Credibility* and 6 levels of the rating *Reliability*. Both credibility and reliability ratings are then crossed (e.g. "**B3**", "**E5**") to give an overall insight into the quality of messages. Visual representations of the matrix and of the distinct levels of evaluation are provided by Tables 3.1 and 3.2.

The conventional reading for score "**E1**" (for instance) is that the source of the message is judged as "Unreliable" while the content the source delivers is "Confirmed" by other sources. It should be noted that, strictly speaking, credibility and reliability ratings range from **1** to **5** and from **A** to **E** (respectively). Ratings **6** and **F** are not evaluatives *stricto sensu* since in these cases, external evidence is lacking to cross-check contents and sources. Exact descriptions of the linguistic labels are given in Tables 4.1 and 4.2 based on the US Army Field Manuals [see FM-2-22.3 2003, Appendix B].

Concerning *credibility*, ratings are captured through different labels that express *decreasing degrees of confirmation* given contextual information consistent or inconsistent with the message (*contextual evidence*). Degree **1** corresponds to an absolute label ("Confirmed") that captures cross-checked certainty. Degrees from **2** to **5** correspond to adverbial modulations ("Probably True", "Possibly True", etc.) that capture *high consistency* ("Probably True"), *moderate consistency* ("Possibly True"), *weak inconsistency* ("Doubtfully True") and *blatant inconsistency* ("Improbable").

have the four basic qualities of relevance, conciseness, clarity and timeliness" [STANAG-2511 2003, 2]. But those qualities should not be taken into account when evaluating credibility and reliability. They are only *writing guidelines* to make the reports valid and actionable for operationals.

Ratings	Linguistic Labels	Descriptions
1	Confirmed	*Confirmed by other independent sources; consistent with other information on the subject*
2	Probably True	*Not confirmed; consistent with other information on the subject*
3	Possibly True	*Not confirmed; agrees with some other information on the subject*
4	Doubtfully True	*Not confirmed; possible; no other information on the subject*
5	Improbable	*Not confirmed; contradicted by other information on the subject*
6	Cannot Be Judged	*No basis exists for evaluating the validity of the information*

Table 3.3: The Credibility of the Message Content.

Ratings	Linguistic Labels	Descriptions
A	Completely Reliable	*No doubt of authenticity, trustworthiness, or competency; has a history of complete reliability*
B	Usually Reliable	*Minor doubt about authenticity, trustworthiness, or competency; has a history of valid information most of the time*
C	Fairly Reliable	*Doubt of authenticity, trustworthiness, or competency but has provided valid information in the past*
D	Not Usually Reliable	*Significant doubt about authenticity, trustworthiness, or competency but has provided valid information in the past*
E	Unreliable	*Lacking in authenticity, trustworthiness, and competency; history of invalid information*
F	Cannot Be Judged	*No basis exists for evaluating the reliability of the source*

Table 3.4: The Reliability of the Message Source.

Degree **6** is ascribed when *no evidential ground* exists for assessing the credibility of the information.

Concerning *reliability*, ratings are captured through adverbial modulations of the evaluative term "reliable" ("Completely Reliable", "Usually Reliable", etc.). They correspond to *decreasing degrees of trustworthiness* based on past experience with the source (*historical evidence*). Assuming that the sources we consider are always *authentic* (because they are competent on the topic they inform about), reliability is a two-dimensional concept that aggregates the sources' propensity to be *truthful* (viz. to regularly provide true information) as well as their propensity to be *honest* (viz. to regularly provide information they believe to be true). In this chapter, analyses of reliability only rely on the sources' *honesty* since their truthfulness is controlled by the truth or falsity of the content they deliver.

3.2.3 Some Virtues of the Alphanumeric Scale

The existing scale is *balanced* since both semantic and pragmatic features of informational messages are taken into account. Semantic aspects are framed through *credibility ratings* associated to the *truth* of the message contents. Pragmatic aspects are covered through *reliability ratings* linked to the sources' *intentions*. This parity conforms with Grice's intention-based semantics according to which both semantic and pragmatic aspects contribute to the *meaning* of messages [see Grice 1957].

But the existing scale is also *relevant* since credibility and reliability are common dimensions one would use to *assess* a given message. Credibility is a crucial parameter for believing a content while reliability is another crucial criterion for trusting a source on a secure basis. Sometimes called "*veracity*" or "*accuracy*", credibility refers to the *plausibility* that a message be true with respect to a given context of evidence. Reliability refers to the officer's relative *certainty* that a given source is honest with respect to their *informational pédigree* (that is to their disposition to tell what they believe to be the truth as well as the objective truth). Defined as such, credibility and reliability should not be confused with *truth* and *honesty*. Credibility and reliability are *subjective dimensions* while truth and

honesty are *objective dimensions* of messages.

But one major issue of the scale is that it does not emphasize these objective dimensions of *truth* and *honesty* on which the subjective dimensions of *credibility* and *reliability* are built. Yet surprisingly, the scale is followed with the assumption that intelligence officers should always make clear the distinction between *objective facts* and *subjective interpretations* in their intelligence reports. Based on past empirical findings, I argue that it is not the case (see section 3.3). The scale comports various dimensions that should be clarified accordingly (see section 3.4).

3.3 Discerning Facts from Interpretations

3.3.1 The Fact vs. Interpretation Assumption

The 6×6 alphanumeric matrix is based on a major assumption we may label "*fact vs. interpretation*" and that is expressed in the NATO Standardization Agreement as follows [see STANAG-2511 2003, 2]:

> "*Intelligence reports transmit facts and/or assessments. The distinction between fact and interpretation must always be clearly indicated.*"

The *fact vs. interpretation* assumption states that intelligence officers must separate intelligence facts from contextual interpretations they make of those facts. In other words, they should clearly distinguish what the scale *objectively* measures from what they think the scale measures from a *subjective* perspective. But does the existing scale allows such a practical distinction? Are the basic features of the scale clear enough to avoid pragmatic equivocation?

In the remaining part of this section, I first argue that ratings labels, as well as their respective levels, are not provided with *clear* definitions and descriptions. For this reason, the distinction between *facts* and *interpretations* cannot be made by officers since they cannot well understand the theoretical notions on which the scale is built. As a matter of fact, experimental results tend to support this

hypothesis. Researchers have observed that confusion in the scale leads to misunderstandings and inconsistencies amongst and between raters. A given officer varies in interpretations he or she makes of the ratings in different contexts while different officers disagree over a single interpretation in a given context. Since officers do not well understand the dimensions of the scale, it is unclear how they could actually comply with the *fact vs. interpretation assumption*.

3.3.2 Identifying Issues in the Assumption

3.3.2.1 Semantic Confusion in the Definition of Ratings

Semantic confusion can be observed at *three different stages* in the *fact vs. interpretation assumption*: *credibility ratings*, *reliability ratings* and *their combination*. First, labels and descriptions associated to *credibility ratings* are conflicting. These labels are based on *plain* adverbial quantifiers ("Probably True", "Possibly True", etc.) while their matching descriptions define them as *conditional* adverbial quantifiers. In fact, the labels "Probably True", "Possibly True", etc. express levels of credibility for intelligence messages *based on a set of consistent/inconsistent evidence*. Furthermore, Capet & Revault d'Allonnes [2013, 115] stress that the highest level of credibility is labelled "Confirmed" as if simple confirmation by a source secured absolute certainty in the message. But without any additional information concerning the source at stake, in fact that the source is "Completely Reliable" in the case in point, no absolute certainty can be ascribed to the message itself [see DIS 2001, Capet & Revault d'Allonnes 2013, for more details on the confirmation issue].

Second, *reliability ratings* are prone to semantic confusion. Their labels and intended descriptions are ambiguous between three senses.[3] It can refer to: (1) the trust the officer puts into the source of the message; (2) the credibility that the source itself attributes to the message; and (3) the capacity of the source to well understand the message he or she delivers. On the other hand, descriptions associated to reliability labels are opaque. They are based on complex notions of *authenticity* and *trustworthiness* without providing minimal definitions of them or

[3] Especially in the case of HUMINT intelligence.

explaining how they combine in reliability ratings.

Third, semantic confusion is inherent to *resultant scores* since they attempt to cross credibility and reliability ratings whose descriptions are confused. Resultant scores naturally inherit from the unclear meanings of the first. It seems that resultant scores are deprived of proper meanings and references since they do not *fuse* but simply *cross* credibility and reliability ratings to give an overall insight into sources and contents. As a consequence, it is difficult, if not impossible, to know whether resultant scores have proper *extensions* and *intensions* on their own.

3.3.2.2 Pragmatic Misunderstandings Intra-Officers

Rather expectedly, pragmatic misunderstandings follow from such semantic confusion. As Phelps, Halpin, Johnson & Moses point out, many authors insisted on the "*loose, ambiguous language used to communicate uncertain intelligence information*" [Phelps *et al.* 1980, 1]. To begin with, they have observed that the existing scale is "*not used at full range*" by intelligence officers, thus indicating that officers have difficulties understanding and interpreting the meaning of credibility ratings, reliability ratings and resultant scores. For instance, Baker *et al.* [1968] analyzed 695 rating reports made by American army officers during field exercises. In all the intelligence reports they examined, only 40 % of the reports contained ratings of both the content credibility and the source reliability. And when both evaluations were made, 74 % of the reports received a score of **B2** (viz. "Usually Reliable"-"Probably True", $N = 518$). Results from Baker & al. were latter replicated by Baker & Mace [1973] who observed no improvement in comprehension when officers were assisted by a decision flow chart (viz. a sequence of basic clear-cut questions) to help them make more appropriate evaluations. Later studies also confirmed these interpretative issues by comparing the current intelligence scale with officers' subjective interpretations of them [see Miron *et al.* 1978, Halpin *et al.* 1978, for details].

Moreover, data collected on the officers' *confidence judgments* also reveal asymmetries in their respective understandings. When officers were asked to express their relative confidence in the scores they provided, Meeland & Rhyne [1967]

observed that the 36 possible scores were *not equally weighted* by intelligence personnels. Officers turned out to be *six times* more confident of a **B1** rating than of a **F3** rating. However, this discrepancy seems intuitive after some clarification. The higher the credibility rating of a message content, the more consistent evidence officers have for cross-checking it and the more confident they are that the message is true. By contrast, the lower the credibility rating of a message content, the less consistent evidence officers have for verifying it and the less confident they are that the message is true (see Table 4.1). Similar arguments can be given for confidence judgments based on reliability ratings [see Peterson 2008, on this point].

Lee & Dry [2006] have shown that *judges*' confidence in ratings of information conveyed by *advisors* were not only determined by the *accuracy* (or *credibility*) of the information at stake but also by the *frequency* of the advisors to convey accurate information over time, namely by their *reliability*. In the field of intelligence, however, officers may be forced to establish their confidence on little feedback concerning the credibility of the message and/or the reliability of the source, — either because external evidence is rare or because the source is new to the officer. In a recent study, Hainguerlot *et al.* [2018] have been interested by the way judges can learn from their confidence judgments in the absence of external feedback. They have observed that the judges' efficiency to make adequate evaluations in the absence of external feedback increased with their metacognitive abilities and that confidence judgments happened to be an adequate proxy for measuring such abilities when feedback is missing.

3.3.2.3 Pragmatic Inconsistencies Inter-Officers

If misunderstandings have been pointed out concerning individual officers (*intra-raters*), inconsistencies have been noticed between officers in their respective interpretations of the ratings (*inter-raters*). Adverbial quantifiers (*probably*, *possibly*, *usually*, *fairly*, etc.) happened to be interpreted differently by intelligence personnels. To begin with, Baker *et al.* [1968] noticed high inconsistencies in the responses provided by different officers in similar contexts. Looking at the ratings made during an intelligence course, they found that those ratings differed

from the instructor solution about 49 % of the time concerning credibility, and 15 % of the time concerning reliability. Based on those observations, researchers have proposed to provide adverbial quantifiers with numerical encodings such as *percentages*, *probabilities*, *odds*, etc. In 1964, the first Director of CIA's Office of National Estimates, Sherman Kent, made a theoretical proposal to match intelligence quantifiers with probabilities and percentages concerning the certainty that events will occur or not. His proposal was intended for the alphanumeric scale in particular [see Kent 1964, 60-61], and to credibility ratings **2** ("Probably True"), **3** ("Possibly True") and **5** ("Improbable") more specifically.

From a probabilistic perspective, a reasonable prediction is that intelligence officers would assign the following probabilities to the credibility ratings ranging from **1** to **5**:

Credibility Ratings	Probability Ranges	Probability Means
1	0.80 - 1.0	0.90
2	0.60 - 0.80	0.70
3	0.40 - 0.60	0.50
4	0.20 - 0.40	0.30
5	0 - 0.20	0.10

Table 3.5: Expected Probability Degrees for Credibility Ratings.

Except for rating **3** ("Possibly True"), these expected probabilities are consistent with the ones Kent [1964] has proposed based on his experience on the field. Kent matched the adjective *probable* with a probability range of 0.63 - 0.87 (mean: 0.75) and the adjective *improbable* with a probability range of 0.20 - 0.40 (mean: 0.30). By contrast, he matched the adjective *possible* with a wider range of probability, that is to more than 0 but less than 1.0, which corresponds to a range of 0 - 1.0 and to an absolute degree of 0.50. But are these degrees and ranges confirmed by empirical findings?

Results that have been collected only concern the assignation of absolute degrees. These results are consistent with expected probabilities but they give rise to high

variability between officers [see Levine & Eldredge 1970, Kelly & Peterson 1971, Samet 1975]. Although Wark [1964] observed that the modal adverb "Probably" corresponds to a probability degree of 0.75 with a very high consensus (90% of inter-agreement), he observed no real consensus that chances are about even (0.50) in case of the modal adverb "Possibly" (only 53% of inter-agreement). This lack of consensus was also observed by Johnson [1973] for the adjective "Possible" whose mean probability was 0.62 but with results varying from 0.04 to 0.80 across officers. The consensus observed by Wark in case of "Probably" was not replicated by Johnson for adjective "Probable" whose results varied from 0.10 to 0.99 across officers (mean: 0.51). Johnson also tested the adjective "Improbable", which was assigned an absolute degree of 0.17 but with results varying from 0 to 0.70.

Such variability in the officers' probabilistic interpretations of adverbs shows the vagueness induced by qualitative vocabulary. Similar results have been obtained in other fields of investigation. In linguistics, for instance, Lichtenstein & Newman [1967] collected results that were consistent with the ones of Wark [see also Budescu & Wallsten 1985]. They observed consistency for the adjective "Probable" (as well as for the adverb "Usually"), and variability for "Possible". In the field of medicine, a comparison of studies from O'Brien [1989] and Bryant & Norman [1980] also show consistency for adjective "Probable" that turned out to be associated to a probability of 0.75 in O'Brien's study and to a probability of 0.77 in Bryant & Norman' study [see also Hobby *et al.* 2000]. But their studies disagree over adjective "Possible" that was associated to a probability degree of 0.25 in O'Brien's study and to a degree of 0.47 in Bryant & Norman's. Besides that, O'Brien and Bryant & Norman obtained similar rates for "Probable" to those of Wark for "Probably" ($p = 0.75$), but results were different for "Possible" in case of O'Brien ($p = 0.25$).

In the intelligence field, proposals for numerical encodings of adverbial quantifiers remain a major challenge due to this high variability in the officers' interpretations. But for the sake of argument, let us suppose that we could finally reach inter-officers agreement concerning the correspondence between intelligence adverbs and probability degrees and ranges. Then, probabilities could be completely

substituted to modal adverbs since officers would agree on a single interpretation for each probability degree or range. However, other issues materialize that have been pointed out in the literature on people's asymmetric perception of risk and probability. Pighin *et al.* [2011] observed that for two distinct numerical probabilities 1/307 < 1/28 such that 1/307 is the probability for a child to have Down Syndrome while 1/28 is the probability for a child to have insomnia, Down Syndrome was interpreted as *more likely* to happen than Insomnia on a 7-point scale ranging from "extremely low" to "extremely high". This effect qualified as a "Severity Bias" [Weber & Hilton 1990, Bonnefon & Villejoubert 2006] shows that the interpretation of probabilities is not purely *extensional* but strongly linked to *expected utilities*, more particularly to the *severity of the outcome* associated to the probability at stake. The more detrimental the outcome associated to the probability, the lower the threshold for qualifying the outcome as being *highly* likely.

Similar observations have been made for *numbers* by Egré [2014] and Egré & Cova [2015].[4] Égré and Égré & Cova showed that the ascription of the vague quantifier *many* is not only determined *extensionally* by the (absolute or relative) number of things at stake but is tied *intensionally* to *moral expectations* associated to these things, that is to what counts as desirable or undesirable in a given context or society. Basically, Egré & Cova [2015] tested people's ascriptions of *many* for pairs of sentences based on the following prediction: participants are more prone to ascribe *many* in case the sentence is linked to a more detrimental outcome. To show this, they asked participants to agree whether "*Many children perished in the fire*" or "*Many children escaped the fire*" in case *exactly* 5 children survived and 5 children perished out of an absolute number of 10. Their prediction was confirmed, thus indicating that cardinals, like probabilities, are not interpreted from a sole *extensional* perspective (based on a given comparative class) but are also interpreted *intensionally* based on moral expectations of desirability. Their findings on judgments about cardinalities parallel those of Knobe' on the increasing effect of moral considerations concerning people's ascriptions of intentions [see Knobe 2003a].

[4] Following Petit & Knobe' results on the asymmetric perception of the gradable adjective *cold* in the context of different liquids [see Pettit & Knobe 2009].

We have seen that the semantic confusion in the labels and descriptions makes officers misunderstand the ratings and scores. This confusion is also apparent when the officers disagree over the correct interpretation of the ratings and scores. Following Kent [1964], some have proposed to substitute numerical encodings to intelligence adverbs to mitigate this confusion. However, results from Pighin *et al.* [2011] and Egré & Cova [2015] show that, even though consensus could be reached on these various encodings, risks associated to probabilities and cardinalities would create new asymetries in the officers' interpretations of the numerals. For this reason, I will continue to use qualitative adverbs for intelligence ratings. But I will clarify the definitions and descriptions of the ratings to mitigate confusion and to help officers better tease apart facts from interpretations in their evaluations.

3.3.3 A Descriptive Proposal for Intelligence Messages

Regarding the intelligence scale, empirical results weaken the hypothesis that officers have a clear comprehension of the scale and can effectively tease apart facts from interpretations on this basis. Of course, semantic confusion in the ratings contribute to this failure so that clarification is needed from a theoretical perspective. But a more practical imperative is to help officers clarify the divide between objective facts and subjective interpretations. In that sense, clarification should be first and foremost a matter of making relevant distinctions.

Basically, *evaluative terms* for contents and sources (viz. *credibility* and *reliability*) should be set aside from the dimensions that are more surely *descriptive terms* for those: *truth* for the content and *honesty* for the source. To use the words of the first assumption, truth and honesty are the *facts* that characterize intelligence messages (in an *objective* sense) while credibility and reliability are subjective *interpretations* of those facts based on contextual and historical evidence. My proposal will consist in trying to understand how intelligence messages can be *defined* and *described* from an objective perspective.

The remaining part of this chapter is devoted to explaining this *descriptive account*. I aim to show that once we set aside the evaluative dimensions of credibility and reliability, truth and honesty can be used to *define* intelligence messages and to

describe the informational type they correspond to in an objective sense (*information*, *misinformation*, *lie*, *omission*, etc.). At the semantic level, various degrees of truth can be distinguished for *characterizing contents*, — either because the content corresponds to objective facts (*truth*) or not (*falsity*), or because its truth status is objectively unclear (*semantic indeterminacy*). At the pragmatic level, various degrees of honesty can also be isolated for *charactering sources* in an objective sense, — either because the source is clearly cooperative (*honest*) or non-cooperative (*dishonest*), or because he or she is less cooperative than they should according to the Gricean principles of communication (*pragmatic imprecision*). Once combined, these semantic and pragmatic aspects lead to isolate distinct *informational categories* for messages, or "*types*". I will also associate these types to more familiar categories based on existing taxonomies (*information*, *misinformation*, *lie*, etc.).

3.4 The Definition of Intelligence Messages

3.4.1 A Matrix for Defining Informational Types

Informational messages m are contents that *specific sources* deliver to *intended addressees*. Determining the type of a given message consists in ascribing a *truth-value* to its content (viz. *true*, *false*, etc.) as well as a *pragmatic intent* to its source (viz. *honest*, *dishonest*, etc.). In doing so, we are led to isolate *pairs* of truth-value and pragmatic intent for all the messages we have. Such pairs, written $\mathbf{t}(m)$, constitute *objective characterizations* of the messages. I call them *informational types*. Thanks to existing taxonomies, more familiar names can be given to those: *information*, *misinformation*, *lie*, *half-truths*, etc.

In this section, I propose a 3×3 matrix to see more clearly into informational types. Consistent with my proposal, the matrix is built on the *descriptive dimensions* of the existing scale: "*Truth of the Content*" vs. "*Honesty of the Source*". I distinguish 3 levels of Truth ("*True*", "*False*", "*Indeterminate*") and 3 levels of Honesty ("*Honest*", "*Dishonest*", "*Imprecise*"). Concerning truth, a content is said to be *true* if it corresponds to objective facts, *false* if it does not and *indeterminate* if its status is semantically unclear. Concerning honesty, sources are said to be *honest* in case they tell what they believe to be the truth, *dishonest* if they tell the opposite of what

they believe and *imprecise* if they partly tell what they believe. Levels of truth and honesty are finally crossed through the following format: [Level of *Honesty*/Level of *Truth*]. For instance, [*Dishonest/True*] means that the source is *dishonest* but that the content he or she delivers is *true*.

3.4.2 The Most Classical Types of Messages

Message Type t		***Truth of the Content***		
		True	*Indeterminate*	*False*
Honesty of the Source	*Honest*	t_1 = [*Honest/True*]		t_2 = [*Honest/False*]
	Imprecise			
	Dishonest	t_3 = [*Dishonest/True*]		t_4 = [*Dishonest/False*]

Table 3.6: The Classical Types of Messages.

Classically, a message content can be either *true* or *false* while a message source can be either *honest* or *dishonest*. According to Table 3.6, we distinguish *four classical categories of messages*: t_1 = [*Honest/True*], t_2 = [*Honest/False*], t_3 = [*Dishonest/True*] and t_4 = [*Dishonest/False*]. I now describe those classical types in detail and determine to which informational category they may correspond according to existing taxonomies.

Let us first consider types t_1 and t_2 in which the source is clearly *honest*. In

both cases, the source is fully cooperative and informative: he or she delivers information they *believe to be true*. But in type $\mathbf{t_1}$, the information is *objectively* true while in type $\mathbf{t_2}$, the information is *objectively* false. So in this second case, the information the source provides is *more misleading* than in the first, even though the source does not have any intention to deceive the officer in both cases. We may call *information*, type $\mathbf{t_1}$, and *misinformation*, type $\mathbf{t_2}$.

From an epistemological perspective, information is usually defined as "*well-formed, meaningful and veridical data*" [Floridi 2007, 31]. *Veridical* must be understood as *truthful*: indeed, information is true data conveyed by honest sources. Following this traditional definition, information necessarily qualifies as *true* [e.g. Dretske 1981, Grice 1989, Floridi 1996]. Accordingly, Dretske [1983, 57] states that "*false information, misinformation (...) are not varieties of information*" while Grice [1989, 371] considers that "*false information is not an inferior kind of information; it just is not information*". Following this traditional view, false information is not information in the proper sense.

But since Fallis [2009a] and Floridi [2011], false information is taken as information *per se*. They call *misinformation* "*well-formed and meaningful data (i.e. semantic content) that are false*" [Floridi 2011, 260]. But though being *false*, misinformation is not falsidical data in the same way information is veridical data. Misinformation is *false* data conveyed *unintentionally*. In case of misinformation, sources are honest since the semantic content is only "*accidentally defective*" [Fallis 2011, 204]. Accordingly, I call *misinformation* type $\mathbf{t_2}$.

Let us now address types $\mathbf{t_3}$ and $\mathbf{t_4}$ in which the source is clearly *dishonest*. In both cases, the source is non-cooperative and even deceitful: her or she delivers information they *believe to be false*. In that respect, the source inevitably flouts Grice's first Maxim of Quality ("*Do not say what you believe to be false*"). But in type $\mathbf{t_3}$, the information he or she provides turns out to be *objectively* true while in type $\mathbf{t_4}$, the information is *objectively* false. Then, the source breaches the Gricean Supermaxim of Quality ("*Try to make your contribution one that is true*") only in the second case. For that reason, information of type $\mathbf{t_3}$ is less misleading than information of type $\mathbf{t_4}$. Even though sources are dishonest in both cases,

information they provide is *epistemically worse* in the second case than in the first. We may call *subjective lie*, type $\mathbf{t_3}$, and *objective lie*, type $\mathbf{t_4}$.

According to the literature on lying, a piece of information qualifies as a *lie* if this piece is a "*believed-false statement*" that is uttered to "*another person with the intention that the other person believe that statement to be true*" [Mahon 2015]. So a content does not need to be *false* to count as a lie but only to be *believed as false*. If dishonesty is a necessary condition for lying, truth or falsity are not. I have shown in Chapter 1 that dishonesty is a sufficient condition for an utterance to count as a *lie* but that traditional epistemologists distinguish between *subjective lies* and *objective lies* along the truth value dimension of the content involved in the speaker's utterance. A dishonest true statement counts as a *subjective lie* while a dishonest false statement counts as an *objective lie*.

Note that type $\mathbf{t_4}$ would be more appropriately called *disinformation* in case the false information is not *uttered* but, more generally, *disseminated* by the dishonest source. Dissemination can be a *linguistic* phenomenon (as in subjective and objective lies made through deliberate utterances) but it can also be a *visual* phenomenon as in falsified maps or forged documents, for instance. In such cases, the addressee has not been stated anything but directed to false information through visual means (such as *pointing* for instance).

We have seen that classical types of messages are based on binary contents, that can be either *true* or *false*, and binary sources, that can be either *honest* or *dishonest*. But a more realistic account of types requires making more fine-grained distinctions between contents and sources by introducing intermediate semantic and pragmatic values. Based on joint work with Paul Egré [see Égré & Icard 2018], I now present several types of messages that are vague for semantic or pragmatic reasons.

3.4.3 Some Types Based on Semantic Vagueness

Linguistic vagueness is a pervasive phenomenon in society as well as in intelligence affairs [see Kent 1964, Capet & Revault d'Allonnes 2013, Barnes 2016]. In a

Message type t		***Truth of the Content***		
		True	*Indeterminate*	*False*
Honesty of the Source	*Honest*		t_5 = *error-avoidance*	
	Imprecise			
	Dishonest		t_6 = *half-truth*	

Table 3.7: Some Types Based on Semantic Vagueness.

seminal article published in 1923, Russell defines vagueness as follows: "*a representation is vague when the relation of the representing system to the represented system is not one-one, but one-many*". According to him, the relation between an expression and a representation is *one-one* in formal languages: "*no two words would have the same meaning*". On the contrary, vagueness naturally arises in natural languages because meaning is no longer one-one but *one-many*. As Russell puts it: "*there is not only one object that a word means, and not only one possible fact that will verify a proposition*" [Russell 1923, 89-90]. Based on Russell's article, two general forms of vagueness can be set apart: *semantic vagueness* vs. *pragmatic vagueness*. I focus on semantic vagueness in the present subsection. Pragmatic vagueness will be addressed in the next one.

According to Table 3.7, types t_5 and t_6 are instances of *semantic vagueness* (or *semantic indeterminacy*). Such vagueness can be observed, for instance, in the

gradable expression *about* (e.g. "*there were about 200 guests at their wedding*") or in the adjective *intelligent* (e.g."*John is intelligent*"). In the first case, *quantitative indeterminacy* is attached to *about* (since the number 200 is compatible with an indeterminate range of numbers between 190 and 210) while *qualitative indeterminacy* is tied to the adjective *intelligent* (since John can be intelligent *in some respect*, i.e. relative to some field of expertise, but not necessarily *in all respects*).

The meaning of expressions that are semantically vague is "*intrinsically uncertain*" according to Peirce [1902, 748]. These expressions have truth conditions that vary from one context to another such that they are true in one sense but false in another. However, type $\mathbf{t_5}$ must be distinguished from type $\mathbf{t_6}$: the source is *honest* in the first case and *dishonest* in the second. In fact, the source flouts the first Maxim of Quality only in type $\mathbf{t_6}$, even though information he or she provides is *indeterminate* in both situations.

In type $\mathbf{t_5}$, sources are clearly *honest* but the message they deliver has *unclear truth conditions*. For a fully cooperative source who turns out to be uncertain whether an event occured or not [see Channell 1994, Frazee & Beaver 2010], indeterminacy is a truthful way for conveying maximum information without risking any falsehoods. Vagueness is indeed an optimal rationale between *honesty* and *truthfulness* in case of uncertainty. If the source was more precise, he or she would either be *dishonest* (by conveying messages they lack evidence for) or *untruthful* (by delivering messages that are blatantly false). Based on Égré & Icard [2018, 10-12], we may label "*error-avoidance*" such message types in which the sources are epistemically cautious towards the officer they will to inform.

In type $\mathbf{t_6}$, the source is clearly *dishonest* and the message he or she delivers is *indeterminate*. Here vagueness is no longer a resource for honesty and truthfulness but a strategic way to mislead the officer. Even though the message may be true in some way of resolving its vagueness, this way is tendencious and biased. Following Égré & Icard [2018, 15-19], I may call "*half-truths*" such message types based on *amphiboly* (or *mental reservation*) [Bok 1979, Mullaney 1980, Adler 1997]. In case of message types $\mathbf{t_5}$, the source expects that the meaning the officer will infer from the message actually differs from the meaning he or she has in mind.

Types $\mathbf{t_5}$ and $\mathbf{t_6}$ could be more finely characterized by distinguishing *two specific cases of semantic indeterminacy*: *degree-vagueness* and *open-texture* [e.g. Waismann 1945, Burks 1946, Alston 1964]. To see more clearly into that, let us consider two examples of Questions & Answers between an intelligence officer and a source belonging to a belligerent country. The source has revelations to make concerning 8 nuclear submarines his or her country is building. Such pieces of information are of a major interest to the intelligence officer due to the devastating power of these nuclear ships.

(1) **Officer:** "*How many submarines did your country build?*

Source: *Many.*"

(2) **Officer:** "*What can you tell me about their submarines?*

Source: *They are powerful.*"

Examples **(1)** and **(2)** are cases of *semantic indeterminacy*: the answers provided by the source are *difficult to interpret* in one way or another. But **(1)** is a case of *degree-vagueness* while **(2)** is a case of *open-texture*.

In **(1)**, the source goes against the officer's expectations for a numerically precise answer. The source gives a response that is consistent with a wide range of states of affairs (5 submarines? 10 submarines? 20 submarines? etc.) and that prevents the officer from representing a precise state of affairs. The use of *many* is misleading here since *many* is not a verbal quantifier to which one can settle precise truth conditions relative to a fixed countable domain because *many* is a context-dependent expression in this case [see Sapir 1944, Partee 1989, Lappin 2000, Greer 2014, Egré & Cova 2015]. Due to strategic issues associated to nuclear submarines, *many* can be interpreted in a multiplicity of ways that have distinct implications in the context at stake.

In **(2)**, the source provides an answer whose semantic meaning is open. The adjective *powerful* is multidimensional [Sassoon 2012] since submarines may be powerful *in some respect* (because they can launch nuclear missiles) but not powerful *in all respects* (because their top speed is limited compared to other nuclear-powered submarines). In the sense of Alxatib & Pelletier [2011], adjective *powerful* can be sub-interpreted (when understood as "*in some respect*") or super-interpreted (when understood as "*in all respects*").

Like for many nominal expressions such as *game* [see Wittgenstein 1953], it seems difficult to frame a set of *necessary* and *sufficient* conditions for applying the adjective *powerful*. For this reason, the officer cannot be sure of what the meaning the source endorses in the context. However, Grice's first Maxim of Quantity ("*Make your contribution as informative as is required for the current purposes of the exchange*") requires that the source gives answers that are *maximally* specific to the officer. Then, it is expected that the source has the *stronger sense* of *powerful* in mind when saying to the officer that the submarines are "*powerful*". The assumption that *powerful* should be interpreted as *powerful in all respects* has also been defended by Dalrymple *et al.* [1998] concerning plural predicates such as *reciprocals*. According to their "*strongest meaning hypothesis*", a sentence expressing reciprocity will be preferably interpreted as expressing *strong* instead of *weak* reciprocity, for instance the sentence "*The girls know each other*" will be preferably interpreted as "*Every girl knows every other girl*" [see also Alxatib & Pelletier 2011, Cobreros *et al.* 2012]. Consistent with this hypothesis, the officer is also expecting that the source will answer along the *stronger sense* of the adjective *powerful*. In other words, if the source answers that the country's submarines are "*powerful*" but means that they are only powerful in the weaker sense, the source is misleading the officer in that respect. Now that we have dealt with cases of semantic vagueness, we can turn to borderline types based on *pragmatic vagueness*.

3.4.4 Some Other Types Based on Pragmatic Vagueness

According to Table 3.8, types $\mathbf{t_7}$ and $\mathbf{t_8}$ are instances of *pragmatic vagueness* (or *imprecision*). Their linguistic contents are *semantically clear-cut* but their sources are *pragmatically imprecise*. In fact, they are less informative than they should according to the first Maxim of Quantity. However, the same sources keep intact all the Maxims of Quality by not telling some content they believe to be false or some content they lack evidence for.

However, type $\mathbf{t_7}$ can be distinguished from type $\mathbf{t_8}$ in terms of pragmatic purposes. Both types involve *concealing information* instead of simply *withholding information* but in distinct ways. Imprecision is used to *hide the truth* in type $\mathbf{t_7}$ and to *hide falsity* in type $\mathbf{t_8}$. Following Chisholm & Feehan [1977] as well as Fallis

Message type $\mathbf{t}$		***Truth of the Content***		
		True	*Indeterminate*	*False*
Honesty of the Source	*Honest*			
	Imprecise	$\mathbf{t_7}$ = *negative omission*	$\mathbf{t_9}$ = *mixed*	$\mathbf{t_8}$ = *positive omission*
	Dishonest			

Table 3.8: Some Types Based on Pragmatic Vagueness.

[2014 2018], both types correspond to different *epistemic goals*. Messages of type $\mathbf{t_7}$ aim to *prevent the officer from acquiring a true belief* (by not unveiling the truth) while messages of type $\mathbf{t_8}$ aim to *maintain an existing false belief in their mind* (by not unveiling falsity). Consistent with those analyses, I label $\mathbf{t_7}$ *negative omission* and $\mathbf{t_8}$, *positive omission*.[5]

As for semantic types $\mathbf{t_5}$ and $\mathbf{t_6}$, pragmatic types $\mathbf{t_7}$ and $\mathbf{t_8}$ can be more finely characterized by differentiating *two specific cases of pragmatic imprecision*: *generality* and *approximation* [e.g. Pinkal 1995, Kennedy 2007, Solt 2015]. To see more clearly into those refinements, let us consider two further variations on the dialogue between the officer and the source.

[5] Those types are named, respectively, "*Negative deception secundum quid*" and "*Positive deception secundum quid*" by Chisholm & Feehan [1977] and then Fallis [2014]. But for simplicity's sake, I prefer to use the more basic labels *negative omission* and *positive omission*.

(3) **Officer:** *"Who is in charge of this new submarine program?*

Source: *Some General at the Admiralty."*

(4) **Officer:** *"What is the shape of the submarines hulls?*

Source: *They are cylindrical."*

Examples **(3)** and **(4)** are cases of *pragmatic imprecision*: the answers provided by the source are *less informative* than it is required by the Gricean Cooperative Principe and the first Maxim of Quantity. But **(3)** is a case of *generality* while **(4)** is one of *approximation*.

In **(3)**, vagueness does not mean any indeterminacy in the truth status of the adverb *some* contrary to **(1)**. Unlike for *many*, *some* can receive determinate truth conditions relative to a fixed countable domain. But in **(1)**, the source fails to be maximally informative by being too general and underspecific in the response he or she gives to the officer. Specifying a precise individual whose role is "General at the Admiralty" would be more informative in that respect.

In **(4)**, the source relies on approximation for answering to the officer's question. Submarines hulls do not have the exact shape of a cylinder but viewing those as *cylinders* is a reasonable coarsening. The shape of submarine hulls is sufficiently close to that of cylinders to see the expression *cylindrical* as acceptable. A well-known geometrical approximation given by Austin [1962] and discussed by Lewis [1979] is that "*France is hexagonal*". Like *cylindrical*, *hexagonal* is semantically determinate but used with slack to approximate the shape of France.

Finally, type $\mathbf{t_9}$ is a (very) borderline type mixing semantic indeterminacy *with* pragmatic imprecision: $\mathbf{t_9}$ = [*Imprecise/Indeterminate*]. In case of $\mathbf{t_9}$, the content of the message has unclear truth conditions and the source of the message is less informative than he or she should according to the Gricean Maxim of Quantity. I will not describe this type in detail since $\mathbf{t_9}$ is a *compound* of the *primitive* borderline types $\mathbf{t_5}$, $\mathbf{t_6}$, $\mathbf{t_7}$ and $\mathbf{t_8}$. I chose to label type $\mathbf{t_9}$ as "*mixed*" to insist on this specificity. Based on the nine types I have identified, I can give a concluding table that wraps up all the informational labels I have proposed for them (see Table 4.10).

Message type t		*Truth of the Content*		
		True	*Indeterminate*	*False*
Honesty of the Source	*Honest*	t_1 = *information*	t_5 = *error-avoidance*	t_2 = *misinformation*
	Imprecise	t_7 = *negative omission*	t_9 = *mixed*	t_8 = *positive omission*
	Dishonest	t_3 = *subjective lie*	t_6 = *half-truth*	t_4 = *objective lie*

Table 3.9: Informational Labels for Message Types.

3.5 Conclusion

The scale commonly used for intelligence evaluation is provided with the assumption that officers should always make a clear distinction between *intelligence facts* and their *subjective interpretations* of these facts [STANAG-2511 2003, 2]. They should indicate when they report objective facts and when they interpret these facts on a more subjective basis. However, empirical results show confusion intra and inter-officers concerning their respective interpretations of the message. Officers misunderstand the credibility and reliability ratings and fail to separate their respective levels. But officers also disagree over a single interpretation of these ratings in different contexts of evaluation. In that sense, it is unclear how officers could comply with the objective-subjective divide since they do not have a clear and consistent understanding of the scale's various dimensions.

Facing those issues, intelligence researchers have proposed to substitute numer-

ical probabilities to the qualitative labels used for the credibility and reliability ratings. However, empirical results show that assignations of degrees of probability give rise to high variability between officers, although probability means are consistent with the probability means we can expect for credibility ratings (see Table 3.5). Supposing that some agreement could be reached in their assignments of probabilities, I also argued that expectations and risks associated to probabilities tend to generate new asymetries in the officers' understandings of the probability rates. For those reasons, I continued to use qualitative adverbs to address the scale's confusion and help officers comply with the distinction between facts and interpretations.

My proposal to mitigate misunderstandings and disagreement has been twofold. To make the *fact vs. interpretation assumption* effective, I have proposed to *separate* the subjective dimensions of credibility and reliability from the objective dimensions of truth and honesty on which the first dimensions rely. That distinction being made, I have proposed to *clarify* the objective dimensions of the scale. I distinguished three degrees of truth for describing contents and three degrees of honesty for characterizing sources. By combining these degrees, I obtained a 3×3 descriptive matrix that isolates nine categories of informational messages: *information*, *misinformation*, *subjective lie*, *objective lie*, *error-avoidance*, *half-truth*, *negative omission*, *positive omission* and *mixed*.

It is important to note that my 3×3 matrix is *not meant as a substitute* for the 6×6 matrix. The 3×3 matrix only gives a *descriptive insight* into the 6×6 evaluative one. However, the 3×3 matrix is meant to *clarify* some aspects that remained implicit in the 6×6 matrix and led to confusion. Let us point out the main efforts we have made in this direction.

Concerning the ratings, the 3×3 matrix helps see more clearly into the descriptive dimensions that determine the 6×6 evaluation of *credibility* and *reliability*. The 3 degrees we consider for truth and honesty offer a clear-cut insight into the objective aspects of contents and sources. We have two absolute degrees of truth (*true vs. false*) and two absolute degrees of honesty (*honest vs. dishonest*), with one fuzzy degree for truth (*indeterminate*) and one fuzzy degree for honesty

(*imprecision*). Though more delineations are possible, these 3 degrees of truth and honesty give a comprehensive perspective on classical and borderline aspects of messages. Being coarser than the 6×6 matrix, the 3×3 matrix is more straightforward and, thus, easier to understand and to handle. This coarser granularity can be seen as positive. By authorizing less degrees of discrimination, the 3×3 matrix forces officers to make more resolute decisions when they categorize messages.

Concerning resultant scores, the 6×6 matrix simply crosses the 6 degrees of credibility with the 6 degrees of reliability to give an overall insight into the quality of messages. Resultant scores are of the form "**RC**" with **R** standing for reliability and **C** standing for credibility. Defined as such, however, resultant scores raise difficulties that the 3×3 matrix addresses. Resultant scores **RC** do not *fuse* information provided by the reliability rating **R** with information provided by the credibility rating **C**. In other words, resultant scores are strictly *redundant* relative to the initial credibility and reliability ratings **R** and**C**. In other words, **RC** do not provide *more information* than **R** and **C**. In contrast, the informational labels I have proposed in the 3×3 matrix (e.g. *misinformation*, *subjective lie*, *objective lie*, etc.) provide *more information* than the initial degrees of truth and honesty they are based on. These qualitative categories transform abstract dimensions of truth and honesty into more familiar categories that are meaningful for intelligence officers. On the contrary, resultant scores are obscure and difficult to interpret meaningfully as it was shown by empirical results [e.g. Baker *et al.* 1968, Baker & Mace 1973, Meeland & Rhyne 1967]. Such instrumental opacity is avoided in the nine categories I have proposed.

Based on the *descriptive account* of Chapter 3, I offer a new *evaluative account* in the next chapter. Consistent with the fact that the 3×3 matrix *does not replace* the 6×6 matrix (descriptive aspects *do not conflate* with evaluative ones), I continue to use six degrees for evaluating the credibility of contents as well as six degrees for evaluating the reliability of sources. However, my proposal is motivated by experimental results showing that contrary to another STANAG assumption [see STANAG-2511 2003, A-2], officers do not put equal weight on the dimensions of credibility and reliability. As we shall see, they perceive the credibility dimension as being *prevalent* over the reliability dimension in their respective evaluations.

Chapter 4

A Dynamic Procedure for Information Evaluation

4.1 Introduction

The doctrine for intelligence evaluation has been critized for wrongly assuming that "*intelligence reports transmit facts and/or assessments*" and that "*the distinction between fact and interpretation [can] always be clearly indicated*" (see Chapter 3). But this doctrine has also been criticized for making the irrealistic assumption that "*reliability and credibility, the two aspects of evaluation, [be] considered independently of each other*" [see STANAG-2511 2003, A-2]. This assumption which we may call *credibility vs. reliability* is actually challenged by empirical results: field officers perceive dimensions of credibility and reliability as being strongly correlated and even redundant [see Baker *et al.* 1968, Samet 1975]. Credibility is seen as the determinant factor for evaluating intelligence messages. The evaluation is made along a credibility continuum on which the reliability of the source helps strike a balance.

The aim of this chapter is to specify a *new evaluative procedure* that conforms with empirical findings. The setting must respect the *overwhelming importance* of the credibility dimension as well as the *balancing role* of the reliability dimension. I propose a syntactic procedure in numerical belief revision whose semantic interpretation complies with those requisits. Syntactically, credibility ratings are

expressed by *degrees operators* indexed from 1 to 6, while reliability ratings are captured by *dynamic operators* indexed from A to F. Semantically, plausibility models help interpret credibility operators for a given message m. States are indexed from 1 to 6 depending on the evidence they satisfy or not, — the aim of which is to provide a *prior* distribution of ratings for m. That being done, reliability ratings act as *scoring functions* on this prior distribution. Based on the reliability of the source, initial distributions are updated in ways that have *positive* or *negative* impacts on the prior credibility of the message.

Let us remind ourselves that the basis for evaluating the truth and honesty of contents and sources through credibility and reliability ratings are relevant *contextual evidence* for truth and relevant *historical evidence* for honesty. A content is said to be "*true*" (or *false*, etc.) depending on the evidence the officer detains in the context at stake. Sources are said to be "*honest*" (or *dishonest*, etc.) depending on the information they have delivered in the past on similar topics. Consistent with those intuitions, levels of truth and/or honesty cannot be ascribed when no contextual and/or historical information is available to the officer.

This chapter is structured as follows. In section 4.2, I briefly review the main notions involved in intelligence evaluation. I present the alphanumeric scale used in this endeavour and put emphasis on its positive features (subsection 4.2.1). I then present the *credibility vs. reliability* assumption on which the scale relies (subsection 4.2.2), and review extant results showing that this assumption is strongly challenged (subsection 4.2.3).

In section 4.3, I explain my *evaluative proposal* in detail. Inspired by previous works from Aucher [2004 2008], van Ditmarsch [2005], van Ditmarsch & Labuschagne [2007], the setting is based on numerical belief revision for expressing a more *subjective characterization* of intelligence messages (subsection 4.3.1). Credibility ratings are captured through a set of *conditional credibility operators* indexed from 1 to 6 (subsection 4.3.2). Reliability ratings are captured through *updates of degrees* labeled from A to F (subsection 4.3.3). Once applied, these updates help get posterior credibility degrees depending on the reliability of the message source.

Section 4.4 is devoted to discussing the scoring operation prescribed by this numerical procedure. I start out by explaining the inner workings of the scoring operation and show that *ratings*, *informational types* and *resultant scores* can be linked further (subsection 4.4.1). The correspondence between *ratings* and *message types* is investigated first (subsection 4.4.2). I argue that dimensions of credibility and reliability can be used for recovering the informational types I have isolated in Chapter 3. I then focus on the correspondence between *scores* and *message types* to put more emphasis on the limitations of the numerical procedure in terms of discriminative power (subsection 4.4.3).

4.2 Some Reminders on Information Evaluation

4.2.1 The 6×6 Matrix for Intelligence Evaluation

Let us briefly review the routine commonly used for intelligence evaluation. A competent authority also known as intelligence officer is charged of assessing informational messages along a 6×6 scale provided with 6 levels for evaluating the credibility of message contents as well as 6 levels for evaluating the reliability of sources [see STANAG-2511 2003, FM-2-22.3 2003, DIA-2 2010]. Both credibility and reliability ratings are then crossed to give an overall insight into the quality of intelligence messages.

Basically, *credibility ratings* are captured through 6 degrees ranging from **1** to **6** (see Table 4.1). The way they are defined, credibility ratings express *decreasing degrees of confirmation* given some *contextual evidence* the officer has *for* or *against* the message being true. In other words, detaining some pieces of evidence in the context is a *precondition* for making an evaluation of the message. Evidence is *consistent* with the message in case some pieces are true when the message is true. Evidence is *inconsistent* with the message when pieces are true but the message is false. Finally, precondition fails when pieces of evidence are missing for making an evaluation of the message. Accordingly, degree **1** corresponds to cross-checked certainty: all the evidence the officer has is consistent with the message being true ("Confirmed"). Degrees **2** to **5** correspond to adverbial modulations that express weaker states of consistency: from *high consistency* ("Probably True") to

Ratings	Linguistic Labels	Descriptions
1	Confirmed	*Confirmed by other independent sources; consistent with other information on the subject*
2	Probably True	*Not confirmed; consistent with other information on the subject*
3	Possibly True	*Not confirmed; agrees with some other information on the subject*
4	Doubtfully True	*Not confirmed; possible; no other information on the subject*
5	Improbable	*Not confirmed; contradicted by other information on the subject*
6	Cannot Be Judged	*No basis exists for evaluating the validity of the information*

Table 4.1: The Credibility of the Message Content.

moderate consistency ("Possibly True"), *weak consistency* ("Doubtfully True") and *blatant inconsistency* ("Improbable"). Degree **6** is ascribed when *no evidence* exists for assessing the credibility of the message ("Cannot Be Judged"). Precondition fails in that latter case.

Reliability ratings are captured through 6 levels ranging from **A** to **F** (see Table 4.2). These ratings express *decreasing degrees of trustworthiness* based on the officer's attitude towards the source. Level **A** corresponds to absence of suspicion ("Completely Reliable"): *no doubt of authenticity*. Levels from **B** to **E** correspond to increasing suspicion againt the source being honest: from *minor doubt* ("Usually Reliable") to *doubt* ("Fairly Reliable"), *significant doubt* ("Not Usually Reliable") and *lack of trustworthiness* ("Unreliable"). Level **F** is ascribed when *no evidence* exists for assessing the reliability of the source ("Cannot Be Judged"), — either because the source is new to the officer or because he or she has not delivered relevant information in the past.

We know from Chapter 3 that the existing setting has strengths. The scale is *balanced* and *relevant* since both semantic and pragmatic dimensions of messages are taken into consideration. Semantic dimensions are covered through credibility ratings associated to the *contents* of messages. Pragmatic aspects are considered through reliability ratings linked to the sources' *intentions* [see Grice 1957]. But the existing scale is also *sensitive* since officers are given 5 levels for evaluating the credibility of the content ("Confirmed", "Probably True", etc.) and 5 levels for evaluating the reliability of the source ("Reliable", "Fairly Reliable", etc.), as well as one extra level on each scale when no evaluation can be made ("Cannot Be Judged"). The intelligence scale being a 6×6 matrix, officers can choose between 36 combinations to evaluate the quality of intelligence messages. That being said, however, the existing procedure is based on a misleading assumption I will present now.

4.2.2 The Credibility vs. Reliability Assumption

In addition to the *fact vs. interpretation* assumption, the scale relies on a second assumption we may call "*credibility vs. reliability*" and such that:

Ratings	Linguistic Labels	Descriptions
A	Completely Reliable	*No doubt of authenticity, trustworthiness, or competency; has a history of complete reliability*
B	Usually Reliable	*Minor doubt about authenticity, trustworthiness, or competency; has a history of valid information most of the time*
C	Fairly Reliable	*Doubt of authenticity, trustworthiness, or competency but has provided valid information in the past*
D	Not Usually Reliable	*Significant doubt about authenticity, trustworthiness, or competency but has provided valid information in the past*
E	Unreliable	*Lacking in authenticity, trustworthiness, and competency; history of invalid information*
F	Cannot Be Judged	*No basis exists for evaluating the reliability of the source*

Table 4.2: The Reliability of the Message Source.

> "*Reliability and credibility, the two aspects of evaluation, must be considered independently of each other.*" [see STANAG-2511 2003, A-2].

This assumption states that intelligence officers must perform their evaluations of credibility and reliability on an independent basis. According to the Field Manual 30-5: "*Although both letters and numerals are used to indicate the evaluation of an item of information, they are independent of each other*" [FM-30-5 1971]. Evaluations of credibility should not interact with evaluations of reliability and, conversely, evaluations of reliability should not interplay with evaluations of credibility. But this presupposes that the credibility and reliability evaluations are indeed perceived as independent and non-overlapping by officers. Is this really the case?

Empirical findings actually challenge the "*credibility vs. reliability*" assumption. As for the distinction between *facts* and *interpretations* (see Chapter 3), the distinction between *credibility* and *reliability* is *not effective* in the alphanumeric scale. Results show that the two dimensions are *not ascribed independently* by intelligence officers. The credibility dimension *prevails over* the reliability dimension in scores they mark.

4.2.3 Identifying Issues with the Assumption

In 1968, Baker, McKendry & Mace analyzed 695 joint ratings obtained from two US intelligence corps during field exercises (Raw Number $N = 716$). Based on the distribution of the ratings, they concluded that the credibility and reliability dimensions were seen as highly *correlated* by officers [see Baker *et al.* 1968]. In fact, Baker & al. observed that 87 % of the scores *fell stricly along the diagonal* of the scale, that is on the continuum **A1-B2-C3-D4-E5-F6** ($N = 608$). Moreover, score **B2** alone comprised 75 % of all the ratings they analyzed ($N = 518$). Results from Baker & al.' experiment are presented in Table 4.3.

In 1975, Samet conducted four experiments to see more clearly into the correlation between the credibility and reliability ratings: Form 1, Form 2, Form 3 and Form 4 [see Samet 1975]. Based on 37 intelligence officers, Samet aimed to confirm the

Distribution of Ratings		*Content Credibility*						Raw
		1	**2**	**3**	**4**	**5**	**6**	**Total**
Source Reliability	**A**	*43*	*11*	*2*	*0*	*0*	*0*	*56*
	B	*11*	*518*	*57*	*2*	*0*	*0*	*588*
	C	*0*	*0*	*5*	*1*	*1*	*0*	*7*
	D	*0*	*0*	*0*	*8*	*0*	*0*	*8*
	E	*0*	*0*	*0*	*0*	*3*	*1*	*4*
	F	*0*	*0*	*1*	*0*	*0*	*31*	*32*
Total		*54*	*529*	*65*	*11*	*4*	*32*	*695*

Table 4.3: Results from Baker, McKendry & Mace' 1968 Experiment.

correlation Baker & al. had observed and to determine which of the credibility and reliability dimensions was prevalent in resultant scores.

The goal of Form 2 was simply to confirm or invalidate the correlation Baker & al. had noticed. The 37 intelligence officers were asked to express the conditional probability that a given report carried a specific *credibility* (or *reliability*) rating provided that the report *already* carried a specific *reliability* (or *credibility*) rating. Results proved to be highly significant 73 % of the time since reponses provided by 27 officers showed a strong interaction between the probabilities the officers assigned to credibility and reliability ratings.[1] This first experiment did confirm Baker & al.' observations. But does one of the evaluative dimensions prevail over the other one? Is credibility or reliability the dominant dimension when evaluating intelligence messages?

Forms 1, 3 and 4 were devoted to answering to these questions. I will review Form 1 in detail but simply present the conclusions Samet drew from Forms 3 and 4 by using techniques of multiple linear regression. In Form 1, officers were given the following scenario:

[1] For 10 intelligence officers, however, the two dimensions were treated independently. But these results have not been included in Samet's analysis for the following reason: officers have different conceptualizations of the ratings in this case. For these officers, the credibility of the content gives no indication of the reliability of the source who delivered that content, — which is counterintuitive according to the doctrinal descriptions (see Table 4.1). Conversely, the reliability of the source is not seen as the determinant parameter for estimating the credibility of the content he or she delivers, — which is also counterintuitive based on the doctrinal descriptions (see Table 4.2).

> *"Suppose that you know that one of two camps, X or Y, is definitely going to be attacked by the enemy. Now suppose that you have two intelligence reports, one saying that camp X will be attacked and the other saying that camp Y will be attacked. The reports differ only in their respective ratings for the reliability of the report's source and the [credibility] of the report's information. Assume further that the given reliability and* [*credibility*] *ratings for each report are correct assessments of their actual reliability and [credibility]. On the basis of this information alone, your task is to decide whether it is more likely that camp X will be attacked or that camp Y will be attacked."*[2]

Officers were then presented a sheet of the following form:

X will be attacked		**Y will be attacked**
	vs.	
$\mathbf{R}_i\mathbf{C}_j$		$\mathbf{R}_k\mathbf{C}_l$

In which **RC** are resultant scores such that $\mathbf{R}_i$, $\mathbf{R}_k$ can be any of the 5 reliability ratings from **A** to **E**, — **F** being excluded, while $\mathbf{C}_j$, $\mathbf{C}_l$ can be any of the 5 credibility ratings ranging from **1** to **5**, — **6** being excluded. Of all the possible 625 combinaisons of joint ratings, Samet excluded all the combinaisons in which reliability or credibility ratings happened to be identical (viz. $i = k$ and/or $j = l$), or in which both the ratings of one score were *higher* than the reliability-credibility ratings of the other score (viz. $i \leq k$ and $j \leq l$ or $k \leq i$ and $l \leq j$). Officers were finally asked to circle camp X or camp Y depending on whether they thought that camp X or that camp Y would be attacked based on their corresponding scores. Basically, each of the 37 officers had to evaluate 100 different combinaisons presented in a random order.

For a given combinaison of scores $\mathbf{R}_i\mathbf{C}_j$ *vs.* $\mathbf{R}_k\mathbf{C}_l$, Samet's prediction was the following: in case $i \leq k$ and the officer elicits the response corresponding to score $\mathbf{R}_k\mathbf{C}_l$, then he or she decides in favor of higher reliability. But in case $j \leq l$ and the

[2] Samet's scenario used the term "*accuracy*" instead of the term "*credibility*" (see intended brackets). But this difference has no impact on our claim since the two evaluative terms received the same scales ranging from **1** to **6** with the same linguistic labels ("Confirmed", "Probably True", "Possibly True", "Doubtfully True", "Improbable" and "Cannot Be Judged") as well as the same descriptions [see Samet 1975, for details].

officer elicits the response corresponding to score $\mathbf{R}_k\mathbf{C}_l$, then he or she decides in favor of higher credibility. Results showed that about 72.1 % of the responses were made from scores based on *higher credibility* but *lower reliability*. In other words, the credibility dimension prevailed over the reliability dimension in resultant scores.

As a matter of fact, Baker & al.' data already gave a slight indication of this prevalence. Setting aside the diagonal continuum which comprised 87 % of joints ratings ($N = 608$), Table 4.3 shows that scores were generally *confined to the high end of the scale*, thus indicating a weak preference for the credibility dimension over the reliability one. In fact, 11 % of the scores fell (stricly) above the diagonal of the scale ($N = 75$) whereas only 2 % of the scores fell (strictly) below the diagonal ($N = 12$). But Samet's results gives a clear-cut insight into the dominance of the credibility dimension.

Samet's further experiments aimed to estimate the relative influence of the individual credibility and reliability ratings in joint ratings. In Form 3, Samet asked the 37 intelligence officers to assign a probability degree to each of the 25 possible scores to express the level of likelihood these degrees correspond to. In Form 4, officers were asked to assign a probability degree to each of the 5 credibility ratings and to each of the 5 reliability ratings. Both experiments revealed high consistency in the responses officers provided. Mean probabilities for ratings and scores are presented in Tables 4.4 and 4.5.

Ratings		**Mean Probabilities**
Reliability	**A**	*.86*
	B	*.73*
	C	*.57*
	D	*.36*
	E	*.18*
Credibility	**1**	*.93*
	2	*.79*
	3	*.61*
	4	*.38*
	5	*.21*

Table 4.4: Samet's Mean Probabilities for each Rating.

Mean Probabilities		*Content Credibility*				
		1	**2**	**3**	**4**	**5**
Source Reliability	**A**	*.96*	*.86*	*.74*	*.55*	*.38*
	B	*.92*	*.81*	*.67*	*.48*	*.31*
	C	*.87*	*.74*	*.60*	*.42*	*.25*
	D	*.81*	*.64*	*.48*	*.32*	*.19*
	E	*.75*	*.56*	*.40*	*.24*	*.14*

Table 4.5: Samet's Mean Probabilities for each Score.

Based on techniques of multiple linear regression, Samet conducted a post-hoc analysis of the results to determine the extent to which the probability degrees of joint ratings (Form 3) could be derived from a linear combination of the probabilities of credibility and reliability ratings (Form 4). The analysis was made from mean probabilities each officer gave to individual ratings and to their combination. For 35 of the 37 offficers, the credibility dimension turned out to account for 76.6 % of resultant scores whereas reliability only accounted for 23.4 % of them [see Samet 1975, 199]. In other words, the credibility dimension was seen *as three times as important as* the reliability dimension by intelligence officers. This strong dominance was later observed by Miron *et al.* [1978] based on 40 different messages and 55 officers enrolled in an intelligence course at the US Army. They found that credibility accounted for 57 % of the quality of intelligence messages, even though other dimensions also played a role of lesser importance such as *relevance* (19 %) and *directness* (6 %) (in particular).

Practically then, these empirical results show that the evaluative procedure needs *revisions*. Following recommendations from Samet [1975] and Phelps *et al.* [1980], intelligence evaluation should be made along a *single credibility scale* on which reliability plays an ancillary role to be determined. In the next section, I propose a new procedure for intelligence evaluation which aims to keep the virtues of the existing scale but to remedy its shortcomings.

4.2.4 A New Proposal for Intelligence Evaluation

I propose to represent the evaluation of intelligence messages in numerical belief revision theory. To conform with empirical findings [see Samet 1975, Phelps *et al.* 1980], the evaluation is made along a single credibility scale on which reliability helps strike resultant scores. Basically, numerical plausibility models are used to define a *prior distribution of credibility ratings* ranging from 1 to 6. In this distribution, each credibility rating is framed through a *specific notion of conditional belief*. Semantically, officers ascribe a credibility rating to a message m at a given state $\boldsymbol{s}$ depending on the pieces of evidence that are *consistent* or *inconsistent* with m at state $\boldsymbol{s}$. Once this prior distribution is determined, officers elicit a reliability rating that corresponds to the level of confidence they put into the source of message m. To this end, reliability ratings are framed by a *set of belief revision operators* labeled from A to F. That being done, the evaluation *per se* consists in marking *posterior* credibility ratings, or *resultant scores*, by updating *prior* distributions of ratings thanks to the reliability operators that may have been elicited. The remaining part of this chapter is devoted to explaining this *evaluative proposal* in detail.

4.3 The Evaluation of Intelligence Messages

4.3.1 A Formal Language $\mathcal{L}_{(\mathtt{intel})}$ for Information Evaluation

Let us define a dynamic-epistemic language $\mathcal{L}_{(\mathtt{intel})}$ for assessing the quality of intelligence messages through credibility and reliability ratings. The setting is inspired by previous works in qualitative belief revision that have a quantitative flavour[3] [see Spohn 1988, Aucher 2004 2008, van Ditmarsch 2005 2008, van Ditmarsch & Labuschagne 2007]. $\mathcal{L}_{(\mathtt{intel})}$ is a propositional syntax in which (conditional) credibility operators $\boldsymbol{C}^n_O$ and dynamic reliability operators $\boldsymbol{R}^{\pm}_O$ are primitives in the language for expressing 6 degrees of credibility for contents as well as 6 degrees of reliability for sources. The set $\mathcal{F}_{(\mathtt{intel})}$ of all $\mathcal{L}_{(\mathtt{intel})}$-*formulas* is given by the following Backus-Naur Form:

[3] To use the words of Baltag & Smets [2008a] and van Ditmarsch [2008].

$\langle$ ***Formulas*** $\rangle$ $\qquad\qquad \varphi, \psi := \top \mid m \mid \neg\varphi \mid \varphi \wedge \psi \mid \boldsymbol{C}^n_O\varphi \mid [\boldsymbol{R}^{\pm}_O\varphi]\psi$

with $n \in \{1,2,3,4,5,6\}$ and $\pm \in$ {A,B,C,D,E,F}. Subscript O refers to the intelligence officer in charge of assessing intelligence messages. The basic features of $\mathcal{L}_{(intel)}$ are the conventional ones: $\top$ is the common abbreviation for tautologies and additional connectives ($\bot$,$\vee$,$\rightarrow$,$\leftrightarrow$) are defined as usual. Atomic propositions stand for *contents* of some informational messages m.

The intuitive interpretation of the (set of) conditional operators $\boldsymbol{C}^n_O\varphi$ is "*Officer O ascribes a degree of credibility n to formula* φ". More precisely, $\boldsymbol{C}^n_O\varphi$ means that O ascribes a degree n to formula φ *based on some piece(s) of evidence* $\varepsilon \in \mathcal{L}_{(intel)}$ *he or she detains in the context* (in a sense I will make precise). The intended reading of the (set of) dynamic operators $[\boldsymbol{R}^{\pm}_O\varphi]\psi$ is the following: "*After officer O performs a reliability update of type* $\pm$ *with formula* φ, ψ *is the case*". I give more precise semantic interpretations to these conditional operators in a doxastic plausibility model:

$$\mathbb{S} = \langle\, \mathcal{S}, \leqslant_O, \| \cdot \| \,\rangle$$

Structure $\mathbb{S}$ consists of a non-empty set $\mathcal{S}$ of *possible states*, a *well-preorder*[4] $\leqslant_O$ for the officer such that $\leqslant_O \subseteq \mathcal{S} \times \mathcal{S}$ and a standard valuation map $\| \cdot \|$: $\mathcal{M} \rightarrow \wp(\mathcal{S})$. The preorder should be interpreted as a (*prior*) *plausibility order*: when $\boldsymbol{s} \leqslant_O \boldsymbol{t}$ (for all $\boldsymbol{s}, \boldsymbol{t} \in \mathcal{S}$), officer O considers that state $\boldsymbol{t}$ is "*at least as plausible as*" state $\boldsymbol{s}$ in model $\mathbb{S}$. I now provide semantic clauses to interpret the non-epistemic formulas of $\mathcal{L}_{(intel)}$. The clauses for the credibility and reliability operations will be defined after some clarifications.

- $\mathbb{S}, \boldsymbol{s} \models \top$ *Always*.
- $\mathbb{S}, \boldsymbol{s} \models m$ *iff* $\boldsymbol{s} \in \| m \|$.
- $\mathbb{S}, \boldsymbol{s} \models \neg\varphi$ *iff* $\mathbb{S}, \boldsymbol{s} \not\models \varphi$.
- $\mathbb{S}, \boldsymbol{s} \models \varphi \wedge \psi$ *iff* $\mathbb{S}, \boldsymbol{s} \models \varphi$ and $\mathbb{S}, \boldsymbol{s} \models \psi$.

[4] Let us remind that a *preorder* over $\mathcal{S}$ is a reflexive and transitive relation over $\mathcal{S}$. A *well-preorder* over $\mathcal{S}$ is a preorder such that every non-empty subset of $\mathcal{S}$ has least elements. Well-foundedness is crucial in our setting: it ensures that non trivial formulas φ are always conditionally believed on some formula ψ [see van Ditmarsch 2008, Baltag & Smets 2006 2008b].

Note that according to the intelligence doctrine, pieces of evidence ε are *preconditions* for evaluating information φ. In that sense, evidence ε can be either *consistent* or *inconsistent* with information φ. Sometimes also, pieces of evidence ε can be all false and fail as preconditions for evaluating the credibility of φ. Semantically, I propose to capture these notions of *consistency*, *inconsistency* and *failure* of adequate precondition through the notion of *conditional plausibility*. The more plausible a message m is relatively to some contextual evidence ε, the more evidence ε can be seen as *consistent* with m being true. On the contrary, the more the negation of message m, namely $\neg m$, is plausible based on some evidence ε, the more ε can be seen as *inconsistent* with m being true. When all pieces ε are false, precondition fails for evaluating the credibility of m. They cannot serve as a basis to adjudicate on m.

4.3.2 Rating Credibility Through Credibility Degrees

4.3.2.1 Expressing Degrees of Credibility

Before matching credibility degrees with credibility ratings, let us define $\| \varepsilon \|$ as the set of states of domain $\mathcal{S}$ that satisfy some formula $\varepsilon \in \mathcal{F}_{(intel)}$: $\| \varepsilon \| := \{\boldsymbol{u} \in \mathcal{S} \mid \mathbb{S}, \boldsymbol{u} \models \varepsilon\}$ for some $\varepsilon \in \mathcal{F}_{(intel)}$. From the order $\leqslant_O$ and set $\| \varepsilon \|$, we can define 6 degrees of *credibility strength* based on formula ε. We write $\boldsymbol{degree}^i(\leqslant_O, \| \varepsilon \|)$ the set of all states of degree i that also satisfy some formula ε. Since $\leqslant_O$ is a well-preorder, every non-empty subset of $\leqslant_O$ has maximal elements. Then, we can derive various degrees of credibility strength relative to formula ε by restricting further and further the set of $\leqslant_O$-maximal states that are ε-consistent. Up to degree $n = 5$, sets $\boldsymbol{degree}^n(\leqslant_O, \| \varepsilon \|)$ are defined by induction:

$$\begin{aligned} \boldsymbol{degree}^1(\leqslant_O, \| \varepsilon \|) &= Max_{\leqslant_O}^{\|\varepsilon\|} \\ \boldsymbol{degree}^n(\leqslant_O, \| \varepsilon \|) &= Max_{\leqslant_O}^{\|\varepsilon\| \setminus \bigcup_{i<n} \boldsymbol{degree}^i(\leqslant_O, \|\varepsilon\|)} \end{aligned}$$

for some $\varepsilon \in \mathcal{F}_{(intel)}$.

For $n = 6$, I define $\boldsymbol{degree}^6(\leqslant_O, \| \neg\varepsilon \|)$ by the clause:

$$\textbf{\textit{degree}}^6(\leqslant_O, \| \neg\varepsilon \|) \quad = \quad Max_{\leqslant_O}^{\|\neg\varepsilon\|}$$

for all $\varepsilon \in \mathcal{F}_{(\texttt{intel})}$.

Here $Max_{\leqslant_O}$ is the set of states that are *maximal* for the plausibility ordering $\leqslant_O$: $Max_{\leqslant_O} := \{\boldsymbol{u} \in \mathcal{S} \mid \forall \boldsymbol{v} \in \mathcal{S}\ \boldsymbol{v} \leqslant_O \boldsymbol{u}\}$. That is: $Max_{\leqslant_O}$ is the set of worlds that officer O considers to be the *most plausible* of the entire ordering. Then, the set $Max_{\leqslant_O}^{\|\varepsilon\|}$ is the set of the most plausible states of the entire ordering that also satisfy formula $\varepsilon \in \mathcal{F}_{(\texttt{intel})}$, that is: $Max_{\leqslant_O}^{\|\varepsilon\|} = Max_{\leqslant_O} \cap \| \varepsilon \|$. Contrariwise, $Max_{\leqslant_O}^{\|\neg\varepsilon\|}$ is the set of the most plausible states of the ordering that do not satisfy any formula $\varepsilon \in \mathcal{F}_{(\texttt{intel})}$.

Intuitively, $\textbf{\textit{degree}}^1(\leqslant_O, \| \varepsilon \|)$ is the set of most plausible states that also satisfy *some formula* $\varepsilon \in \mathcal{F}_{(\texttt{intel})}$: $Max_{\leqslant_O}^{\|\varepsilon\|}$ for some $\varepsilon \in \mathcal{F}_{(\texttt{intel})}$. Then, the sets from $\textbf{\textit{degree}}^2(\leqslant_O, \| \varepsilon \|)$ to $\textbf{\textit{degree}}^5(\leqslant_O, \| \varepsilon \|)$ are obtained by successively *removing* ε-states from the top of the plausibility ordering. In that sense, $\textbf{\textit{degree}}^2(\leqslant_O, \| \varepsilon \|)$ is the set of *most plausible states* that are *not the most plausible* ε*-states of the entire ordering*, namely that are not in $\textbf{\textit{degree}}^1(\leqslant_O, \| \varepsilon \|)$. Set $\textbf{\textit{degree}}^3(\leqslant_O, \| \varepsilon \|)$ is the set of most plausible ε-states that are *neither* in $\textbf{\textit{degree}}^1(\leqslant_O, \| \varepsilon \|)$ *nor* in $\textbf{\textit{degree}}^2(\leqslant_O, \| \varepsilon \|)$, *and so on*. But $\textbf{\textit{degree}}^6(\leqslant_O, \| \neg\varepsilon \|)$ is defined differently: $\textbf{\textit{degree}}^6(\leqslant_O, \| \neg\varepsilon \|)$ is the set of the most plausible states of the entire ordering that do not satisfy *any formula* $\varepsilon \in \mathcal{F}_{(\texttt{intel})}$, namely $Max_{\leqslant_O}^{\|\neg\varepsilon\|}$ for all $\varepsilon \in \mathcal{F}_{(\texttt{intel})}$.

Up to $n = 5$, formula(s) ε can be seen as an *ordering source* in the sense of Kratzer' [see Kratzer 1981 1991, Lassiter 2017]. Definitions of sets $\textbf{\textit{degree}}^1(\leqslant_O, \| \varepsilon \|)$ to $\textbf{\textit{degree}}^5(\leqslant_O, \| \varepsilon \|)$ are based on whether states that belong to them satisfy *at least* one formula $\varepsilon \in \mathcal{F}_{(\texttt{intel})}$. But since $\textbf{\textit{degree}}^1(\leqslant_O, \| \varepsilon \|)$ to $\textbf{\textit{degree}}^5(\leqslant_O, \| \varepsilon \|)$ are obtained by successively removing ε-states from the top of the plausibility ordering, states are finally ordered depending on which formula(s) ε they do satisfy or not. In that sense, formulas ε induce a plausibility ordering over the distinct possible states depending on whether some of these ε turn out to be *true* or *false* at these states. This notion of *ordering source* will become clearer in the evaluative case I will present later.

From those six sets of credibility strength, six degrees of conditional credibility can

be defined for the intelligence officer. Up to degree $n = 5$, the officer's conditional credibility of degree n in formula φ is given by the clause:

- $\mathbb{S}, \boldsymbol{s} \models \boldsymbol{C}^n_O\varphi$ *iff* for all $\boldsymbol{t} \in \boldsymbol{degree}^n(\leqslant_O, \| \varepsilon \|)$: $\mathbb{S}, \boldsymbol{t} \models \varphi$.

for some $\varepsilon \in \mathcal{F}_{(intel)} \backslash \{\varphi, \top\}$.

This semantic clause means that formula φ is judged as credible by officer O at a degree $n \leq 5$ in state $\boldsymbol{s}$ (of model $\mathbb{S}$) *if and only if* formula φ is true in the most plausible states of degree n that also satisfy some other formula ε (distinct from φ itself). In other words, *some precondition* ε exists for evaluating the credibility of φ.

The restriction $\varepsilon \in \mathcal{F}_{(intel)} \backslash \{\varphi, \top\}$ is made to avoid lack of informativity of evidence ε (in case $\varepsilon = \top$) and trivialities (in case $\varepsilon = \varphi$). But this restriction will also prove useful in the next subsection when credibility degrees will be matched with the doctrinal credibility ratings. For degree $n = 6$, conditional credibility is defined by:

- $\mathbb{S}, \boldsymbol{s} \models \boldsymbol{C}^6_O\varphi$ *iff* for all $\boldsymbol{t} \in \boldsymbol{degree}^6(\leqslant_O, \| \neg\varepsilon \|)$: $\mathbb{S}, \boldsymbol{t} \models \varphi$.

for all $\varepsilon \in \mathcal{F}_{(intel)} \backslash \{\varphi, \top\}$.

Here the restriction $\varepsilon \in \mathcal{F}_{(intel)} \backslash \{\varphi, \top\}$ is made to avoid lack of informativity of ε as well as contradictions (in case $\varepsilon = \varphi$). The semantic clause means that formula φ is judged as being credible to degree 6 by officer O at state $\boldsymbol{s}$ (in model $\mathbb{S}$) *if and only if* formula φ is true in the most plausible states of the ordering where no other formula ε (distinct from φ itself) also turns out to be true. In that case, *precondition* fails for making an evaluation of φ since all the pieces of evidence ε are false.

For clarity's sake, let us rewrite the sets $\boldsymbol{degree}^1(\leqslant_O, \| \neg\varepsilon \|)$ to $\boldsymbol{degree}^5(\leqslant_O, \| \neg\varepsilon \|)$, as well as set $\boldsymbol{degree}^6(\leqslant_O, \| \neg\varepsilon \|)$, in a simpler way: $\boldsymbol{degree}^i$ for all $i \in \{1, 2, 3, 4, 5, 6\}$. From the sets $\boldsymbol{degree}^i$, we can define the degree function $\boldsymbol{dg} : \mathcal{S} \to \{1, 2, 3, 4, 5, 6\}$ such that: $\boldsymbol{dg}(\boldsymbol{s}) = i$ *iff* $\boldsymbol{s} \in \boldsymbol{degree}^i$. This means that the degree of state $\boldsymbol{s}$ is equal

to i if and only if state $\boldsymbol{s}$ belongs to the set of states of degree i. Accordingly, the semantic clause for operators $\boldsymbol{C}^n_O\varphi$ can be rewritten in the following way:

- $\mathbb{S}, \boldsymbol{s} \models \boldsymbol{C}^n_O\varphi$ *iff* for all $\boldsymbol{t}$ such that $\boldsymbol{dg}(\boldsymbol{t}) = n : \mathbb{S}, \boldsymbol{t} \models \varphi$.

with $n \in \{1, 2, 3, 4, 5, 6\}$. This clause intuitively means that formula φ is judged as being credible by officer O to a degree $n \leq 6$ in state $\boldsymbol{s}$ of model $\mathbb{S}$ *if and only if* formula φ happens to be true in the states of degree n of model $\mathbb{S}$.

Now I propose to match the various credibility operators with the doctrinal credibility ratings ranging from **1** to **6**. First note that the doctrinal ratings are *not strict* but *all conditional*, — at least when the evaluation is possible. According to Table 4.1, ascribing a level of credibility to a message m consists in determining the *conditional credibility* of message m based on a set of *contextual evidence* ε the officer has for, or against, m. For this reason, doctrinal credibility ratings can be captured by teasing apart sets of consistent evidence in terms of plausibility strength. Let us proceed step-by-step.

4.3.2.2 Matching Credibility Degrees with Credibility Ratings

According to doctrinal descriptions (see Table 4.1), a message m is classified as "Confirmed" and rated **1** if it is *confirmed by independent sources* and *consistent with other information on the subject*. Being *confirmed*, the message is expected to reach the maximum credibility score. In such a case, m is judged as credible for being true in the *best* set of plausible states satisfying some relevant evidence ε consistent with it: $Max^{\|\varepsilon\|}_{\leqslant_O}$ *for some evidence* $\varepsilon \in \mathcal{F}_{(intel)} \backslash \{\varphi, \top\}$. Accordingly, I propose to match this top credibility rating with conditional credibility $\boldsymbol{C}^1_O$ from language $\mathcal{L}_{(intel)}$.

By contrast, a message is classified as "Probably True" and rated **2** if it is *not confirmed* but *consistent with other information on the subject*. Semantically, this rating is weaker than rating **1**. When rated **2**, a message is consistent with *most* but *not all* the highly plausible evidence the officer has. This can be understood as follows: the message m is true in the *second* set of most plausible states satisfying some evidence ε consistent with it. So the best set is no longer $Max^{\|\varepsilon\|}_{\leqslant_O}$ but a

slight restriction of it: $Max_{\leqslant_O}^{\|\varepsilon\|\setminus degree^1}$. I propose to match this credibility rating with conditional credibility $\boldsymbol{C}_O^2 m$.

In the same vein, a content is classified as "Possibly True" and rated **3** if it is *not confirmed* and only *consistent with some other information on the subject*. Now the set of consistent evidence is even more tenuous than for rating **2**: the officer judges that *some* but not *most* of the states that are consistent with m are highly plausible. Based on similar intuitions as before, I propose to match credibility rating **3** with degree $\boldsymbol{C}_O^3$. In that case, the set of most plausible states is the *third best* set overall: $Max_{\leqslant_O}^{\|\varepsilon\|\setminus\{degree^1 \cup degree^2\}}$. Accordingly, rating **3** is matched with conditional credibility $\boldsymbol{C}_O^3 m$.

The description provided with rating "Doubtfully True" is more difficult to interpret: a message m is rated **4** if it is *not confirmed*, *possible* but not conclusive since *no other information on the subject* is available. However, this description cannot be taken at first value because of two difficulties. First, if a message was classified as "Doubtfully True" because no other information on the subject was available, rating **4** could not be distinguished from rating **6**. Second, words used for the label "Doubtfully True" conflict with the description itself: *doubtfully* suggests that the truth status of the message is *uncertain*.

As a matter of fact, the label "Doubtfully True" indicates that there is *doubt* concerning the credibility of message m. The message is only *moderately plausible* based on the contextual evidence the officer has. Accordingly, the set of states in which m is true is weaker than before in terms of plausibility strength. So we can assume that the message m is rated **4** when m turns out to be true in the *fourth best* plausible states satisfying some evidence ε consistent with it: $Max_{\leqslant_O}^{\|\varepsilon\|\setminus\{degree^1 \cup degree^2 \cup degree^3\}}$. The corresponding credibility operator for rating **4** is then $\boldsymbol{C}_O^4 m$.

Concerning rating "Improbable", the doctrine tells that a message is classified **5** if it is *not confirmed* as well as *contradicted by other information on the subject*. In such a case, m is very implausible based on the contextual evidence the officer has. Intuitively then, m is true only in the *fifth* set of most plausible states that are also

the least plausible states of the ordering: $Max_{\leqslant_O}^{\|\varepsilon\|\setminus\{\boldsymbol{degree}^1\cup\boldsymbol{degree}^2\cup\boldsymbol{degree}^3\cup\boldsymbol{degree}^4\}}$. Quite naturally, the credibility operator for rating **5** is $\boldsymbol{C}_O^5 m$.

Finally, rating **6** is ascribed when the message m "Cannot Be Judged": *no basis exists for evaluating the validity of* m. There is no evidence ε, consistent or inconsistent with m, on which the officer can condition his credibility on m. I propose to match rating **6** with credibility operator $\boldsymbol{C}_O^6 m$ for this reason. The clauses I have given for expressing credibility ratings are summed up in Table 4.6.

Rating	**Label**	**Set of States in** $Max_{\leqslant_O}$
$\boldsymbol{C}_O^1 m$	Confirmed	$\|\| \varepsilon \|\|$
$\boldsymbol{C}_O^2 m$	Probably True	$\|\| \varepsilon \|\| \setminus\{\boldsymbol{degree}^1\}$
$\boldsymbol{C}_O^3 m$	Possibly True	$\|\| \varepsilon \|\| \setminus\{\boldsymbol{degree}^1 \cup \boldsymbol{degree}^2\}$
$\boldsymbol{C}_O^4 m$	Doubtfully True	$\|\| \varepsilon \|\| \setminus\{\boldsymbol{degree}^1 \cup \boldsymbol{degree}^2 \cup \boldsymbol{degree}^3\}$
$\boldsymbol{C}_O^5 m$	Improbable	$\|\| \varepsilon \|\| \setminus\{\boldsymbol{degree}^1 \cup \boldsymbol{degree}^2 \cup \boldsymbol{degree}^3 \cup \boldsymbol{degree}^4\}$
$\boldsymbol{C}_O^6 m$	Cannot Be Judged	$\|\| \neg\varepsilon \|\|$

Table 4.6: Semantic Conditions for Credibility Ratings.

4.3.2.3 A General Case Study

Let us model a practical case of evaluation. Graphically, to represent the fact that the state $\boldsymbol{t}$ is *at least as plausible as* the state $\boldsymbol{s}$ for officer O ($\boldsymbol{s} \leqslant_O \boldsymbol{t}$), I draw a right arrow " $\rightarrow$ " from state $\boldsymbol{s}$ to state $\boldsymbol{t}$: $\boldsymbol{s} \rightarrow \boldsymbol{t}$. When states $\boldsymbol{s}$ and $\boldsymbol{t}$ are *equally plausible* for the officer ($\boldsymbol{s} \leqslant_O \boldsymbol{t}$ and $\boldsymbol{t} \leqslant_O \boldsymbol{s}$), I draw a left-right arrow " $\leftrightarrow$ " between $\boldsymbol{s}$ and $\boldsymbol{t}$: $\boldsymbol{s} \leftrightarrow \boldsymbol{t}$. Reflexive arrows are omitted at each state to simplify the models. Each arrow is labelled with a numerical integer $n \in \{1, 2, 3, 4, 5, 6\}$ that indicates the degree of the state located on top of the arrow. For instance, if $\overset{\rightarrow}{}_n \boldsymbol{s}$, then the degree of state $\boldsymbol{s}$ is n. When two states are equally plausible, for instance $\boldsymbol{s} \overset{\leftrightarrow}{}_n \boldsymbol{t}$, they receive the same plausibility degree n.

Suppose the intelligence officer has to evaluate the following message from a given source:

$$m : \text{"}\textit{My country is building 8 nuclear submarines}\text{"}$$

Suppose that for evaluating the credibility of m, the officer detains *three pieces* of contextual evidence e, f and g:

$$e : \text{"}\textit{Imagery shows that the country has been delivered 8 submarine hulls}\text{"}$$

$$f : \text{"}\textit{The country is testing existing torpedos in the bay area}\text{"}$$

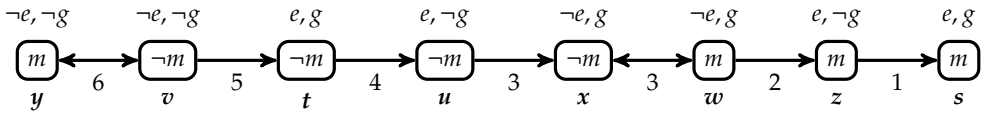
$$g : \text{"}\textit{Sources reported that the country has decided to buy new nuclear reactors}\text{"}$$

We can see that amongst those pieces of evidence, only e and g are *highly relevant* to the evaluation of message m. Evidence f is less relevant with respect to the evaluation of m. Hence the states of contextual evidence on which the officer can define his or her set of prior credibility ratings $\boldsymbol{C}^n_O m$ are based on the set $\boldsymbol{\varepsilon} = \{e, g\}$. From set $\boldsymbol{\varepsilon}$ and message m, eight possible states can be distinguished:

$$\boldsymbol{s}^{\{e,g,m\}}, \quad \boldsymbol{t}^{\{e,g,\neg m\}}, \quad \boldsymbol{u}^{\{e,\neg g,\neg m\}}, \quad \boldsymbol{v}^{\{\neg e,\neg g,\neg m\}}, \quad \boldsymbol{w}^{\{\neg e,g,m\}}, \quad \boldsymbol{x}^{\{\neg e,g,\neg m\}}, \quad \boldsymbol{y}^{\{\neg e,\neg g,m\}}, \quad \boldsymbol{z}^{\{e,\neg g,m\}}$$

In states where at least one piece of evidence amongst evidence e and g turns out to be true, precondition exists for making an evaluation of message m: states $\boldsymbol{s}$, $\boldsymbol{t}$, $\boldsymbol{u}$, $\boldsymbol{w}$, $\boldsymbol{x}$ and $\boldsymbol{z}$. But in states where both pieces e and g turn out to be false,

precondition fails for making such an evaluation of m. That is in states $\boldsymbol{y}$ and $\boldsymbol{v}$. I put states $\boldsymbol{y}$ and $\boldsymbol{v}$ at the bottom of the ordering to set them apart. Let us now suppose that the officer gives the following *prior distribution* of degrees to message m based on evidence from the set $\boldsymbol{\mathcal{E}}$:



Remind that the numerals below each arrow (for instance, $\overset{\rightarrow}{_4}\boldsymbol{u}$) indicate the (prior) plausibility of the state that is located on the top of the arrow (state $\boldsymbol{u}$ in that case). More precisely, these numerals indicate the credibility rating of the message that is *true* or *false* at the state on top of the arrow. The model shows that message m is rated **1** at state $\boldsymbol{s}$ whereas the negation of m, namely $\neg m$, is rated **5** at state $\boldsymbol{t}$. These ratings depend on the pieces of evidence ε that turn out to be *true* (or *false*) when the message itself turns out to be *true* (or *false*).

As I said earlier, the set $\boldsymbol{\mathcal{E}}$ plays the role of a Kratzerian *ordering source* [Kratzer 1981 1991]. Possible states are ordered along the relation $\leqslant_o$ depending on whether or not they satisfy evidence e,$g \in \boldsymbol{\mathcal{E}}$. More precisely, evidence set $\boldsymbol{\mathcal{E}}$ induces a plausibility ordering over the possible states since the more states satisfy pieces of evidence from $\boldsymbol{\mathcal{E}}$ that are consistent with m, the more these state are judged as plausible by the officer. On the contrary, the more the states satisfy pieces of evidence from $\boldsymbol{\mathcal{E}}$ that are consistent with $\neg m$, the more these states are judged as implausible by the officer.[5]

When determining a credibility distribution for message m, it seems reasonable that *if* the officer gives the degree of credibility **1** to message m based on the evidence set $\{e, g\}$, the same officer will give, *by symmetry*, a degree of credibility that is very low to $\neg m$ based on the same evidence set (degree **5** in that case). A similar reasoning applies to all the degrees of credibility. The different ratings and their evidence sets are presented in Table 4.7.

[5] A difference with Kratzer's framework is that in my own proposal, evidence from set $\boldsymbol{\mathcal{E}}$ do not have the same weight. For instance, evidence e has more weight than evidence e since message m gets a credibility degree of **2** based on the evidence set $\{e, \neg g\}$ but a lower, and even uncertain, credibility degree of **3** based on the evidence set $\{\neg e, g\}$.

State(s)	Evidence Set	Prior Rating $C^n_O m$
s	$\{e,g\}$	$C^1_O m$
z	$\{e,\neg g\}$	$C^2_O m$
x,w	$\{\neg e, g\}$	$\neg C^3_O m \wedge \neg C^3_O \neg m$
u	$\{e,\neg g\}$	$C^4_O \neg m$
t	$\{e,g\}$	$C^5_O \neg m$
y,v	$\{\neg e, \neg g\}$	$\neg C^6_O m \wedge \neg C^6_O \neg m$

Table 4.7: Evidence-Based *Prior* Ratings.

Now that I have matched doctrinal credibility ratings with credibility operators in $\mathcal{L}_{(intel)}$, I turn to reliability ratings to propose similar clauses. But the correspondence between reliability ratings and updates of degrees is not as straightforward as before. Unlike for credibility, reliability ratings cannot be defined from literal readings of their intended descriptions.

4.3.3 Rating Reliability Through Degrees Updates

4.3.3.1 Expressing Updates of Credibility Degrees

The general semantic clause for operation $[R^{\pm}_O \varphi]\psi$ in model $\mathbb{S}$ is:

- $\mathbb{S}, s \models [R^{\pm}_O \varphi]\psi$ *iff* $\mathbb{S}^{[R^{\pm}_O \varphi]}, s \models \psi$.

Updated models $\mathbb{S}^{[R^{\pm}_O \varphi]}$ are obtained from $\mathbb{S}$ in the following way:

$$\mathbb{S}^{[R^{\pm}_O \varphi]} = \langle \mathcal{S}^{[R^{\pm}_O \varphi]}, \leqslant_O^{[R^{\pm}_O \varphi]}, \| \cdot \|^{[R^{\pm}_O \varphi]} \rangle$$

In models $\mathbb{S}^{[R^{\pm}_O \varphi]}$, the sets of states $\mathcal{S}^{[R^{\pm}_O \varphi]}$ and valuation maps $\| \cdot \|^{[R^{\pm}_O \varphi]}$ are strictly identical to the initial set of states $\mathcal{S}$ and valuation map $\| \cdot \|$ from $\mathbb{S}$. The crucial aspect of $\mathbb{S}^{[R^{\pm}_O \varphi]}$ are the plausibility preorders $\leqslant_O^{[R^{\pm}_O \varphi]}$. For all $\pm \in$ {A,B,C,D,E,F}, each operation $R^{\pm}_O$ with φ induces a specific change on the ranking of states depending on whether these states satisfy formula φ or not.

To begin with, it is important to distinguish two *kinds of situations*. There are situations in which formula φ is evaluable at the state s under consideration in

the current plausibility ordering $\leqslant_O$: $dg(s) < 6$. But there are also situations in which formula φ is not evaluable at the state s under consideration in the current plausibility ordering $\leqslant_O$: $dg(s) = 6$. Making such a distinction is crucial since in the first case (viz. $dg(s) < 6$), the credibility of φ in the new ordering $\leqslant_O^{[R_O^{\pm}\varphi]}$ will depend *both* on the prior credibility of the formula at state s *and* on the level of reliability $\pm$ of its source at the same state. In the second case (viz. $dg(s) = 6$), the credibility of formula φ in the updated ordering $\leqslant_O^{[R_O^{\pm}\varphi]}$ will depend *only* on the reliability of its source at state s.

In both situations, however, I write $dg(s)^{R_O^{\pm}\varphi}$ for the *posterior degree* of state s obtained by exerting operation $\pm$ with formula φ on the *prior degree* $dg(s)$ of state s. To ensure that new degrees $dg^{R_O^{\pm}\varphi}$ remain within the bounds of set $\{1, 2, 3, 4, 5\}$, we introduce the function $Cut(x)$ that will be applied in the definition of updates.[6]

$$Cut(x) = \begin{cases} x & \text{if } 1 \leqslant x \leqslant 5 \\ 1 & \text{if } x < 1 \\ 5 & \text{if } x > 5 \end{cases}$$

Let us start out by defining the operations $R_O^{\pm}$ when formula φ is evaluable at the state s under consideration: $dg(s) < 6$. Operation $R_O^{A}\varphi$ is defined as the update in which all φ-states of the initial plausibility ordering $\leqslant_O$ gain 3 ranks in the new ordering $\leqslant_O^{[R_O^{A}\varphi]}$ while all $\neg\varphi$-states of the initial plausibility ordering $\leqslant_O$ loose 3 ranks in the new ordering $\leqslant_O^{[R_O^{A}\varphi]}$. In terms of credibility degrees, this means that degrees of φ-states *decrease by* 3 while degrees of $\neg\varphi$-states *increase by* 3. Accordingly, operation $R_O^{A}\varphi$ is defined by:

$$dg(s)^{R_O^{A}\varphi} = \begin{cases} Cut(dg(s) - 3) & \text{if } \mathbb{S}, s \models \varphi. \\ Cut(dg(s) + 3) & \text{if } \mathbb{S}, s \models \neg\varphi. \end{cases}$$

Operation $R_O^{B}\varphi$ is the update according to which all φ-states of the initial plausibility ordering gain 2 ranks in the new ordering $\leqslant_O^{[R_O^{B}\varphi]}$ (their degrees *decrease by* 2) while all $\neg\varphi$-states of the initial plausibility ordering loose 2 ranks in the new ordering $\leqslant_O^{[R_O^{B}\varphi]}$ (their degrees *increase by* 2). Numerically, $R_O^{B}\varphi$ is defined by:

[6] This technical device is inspired by Aucher [2004].

$$dg(s)^{\boldsymbol{R}^{\mathrm{B}}_{o}\varphi} \quad = \quad \begin{cases} Cut(dg(s)-2) & \text{if } \mathbb{S}, s \models \varphi. \\ Cut(dg(s)+2) & \text{if } \mathbb{S}, s \models \neg\varphi. \end{cases}$$

Operation $\boldsymbol{R}^{\mathrm{C}}_{O}\varphi$ is the update according to which all φ-states of the initial ordering gain 1 rank in the new ordering $\leqslant_{O}^{[\boldsymbol{R}^{\mathrm{C}}_{o}\varphi]}$ while all $\neg\varphi$-states of the initial ordering loose 1 rank in the new ordering $\leqslant_{O}^{[\boldsymbol{R}^{\mathrm{C}}_{o}\varphi]}$. Numerically, operation $\boldsymbol{R}^{\mathrm{C}}_{O}\varphi$ is given by:

$$dg(s)^{\boldsymbol{R}^{\mathrm{C}}_{o}\varphi} \quad = \quad \begin{cases} Cut(dg(s)-1) & \text{if } \mathbb{S}, s \models \varphi. \\ Cut(dg(s)+1) & \text{if } \mathbb{S}, s \models \neg\varphi. \end{cases}$$

Contrary to the operations $\boldsymbol{R}^{\mathrm{A}}_{O}\varphi$ and $\boldsymbol{R}^{\mathrm{C}}_{O}\varphi$ that have a promoting effect on φ-states, operation $\boldsymbol{R}^{\mathrm{D}}_{O}\varphi$ is the update according to which all φ-states of the initial ordering loose 1 rank in the new ordering $\leqslant_{O}^{[\boldsymbol{R}^{\mathrm{D}}_{o}\varphi]}$ (their degrees *increase by* 1) while all $\neg\varphi$-states of the initial ordering gain 1 rank in the new ordering $\leqslant_{O}^{[\boldsymbol{R}^{\mathrm{D}}_{o}\varphi]}$ (their degrees *decrease by* 1). In terms of plausibility degrees, operation $\boldsymbol{R}^{\mathrm{D}}_{O}\varphi$ is defined by:

$$dg(s)^{\boldsymbol{R}^{\mathrm{D}}_{o}\varphi} \quad = \quad \begin{cases} Cut(dg(s)+1) & \text{if } \mathbb{S}, s \models \varphi. \\ Cut(dg(s)-1) & \text{if } \mathbb{S}, s \models \neg\varphi. \end{cases}$$

Operation $\boldsymbol{R}^{\mathrm{E}}_{O}\varphi$ is the update according to which all φ-states of the initial ordering loose 2 ranks in the new ordering $\leqslant_{O}^{[\boldsymbol{R}^{\mathrm{E}}_{o}\varphi]}$ while all $\neg\varphi$-states of the initial ordering gain 2 ranks in the new ordering $\leqslant_{O}^{[\boldsymbol{R}^{\mathrm{E}}_{o}\varphi]}$. Numerically, operation $\boldsymbol{R}^{\mathrm{E}}_{O}\varphi$ is defined by:

$$dg(s)^{\boldsymbol{R}^{\mathrm{E}}_{o}\varphi} \quad = \quad \begin{cases} Cut(dg(s)+2) & \text{if } \mathbb{S}, s \models \varphi. \\ Cut(dg(s)-2) & \text{if } \mathbb{S}, s \models \neg\varphi. \end{cases}$$

Contrary to all the operations above, operation $\boldsymbol{R}^{\mathrm{F}}_{O}\varphi$ leaves the initial plausibility ordering $\leqslant_{O}$ as it is, — no matter if formula φ is true or false at the state where the operation is performed. Accordingly, operation $\boldsymbol{R}^{\mathrm{F}}_{O}\varphi$ is defined by:

$$dg(s)^{R_o^F\varphi} \quad = \quad dg(s) \quad \textit{if } \mathbb{S}, s \models \varphi \textit{ or if } \mathbb{S}, s \models \neg\varphi.$$

So far, I have defined the $R_O^{\pm}$-operations in case the degree of the evaluation state s is less than 6. I have focused on cases in which true evidence already exists at state s for making an evaluation of formula φ. But what if all the evidence turns out to be false at the state s such that the credibility of formula φ cannot be judged? In this case, the officer's evaluation will be based *only* on the reliability of the source. In other words, posterior credibility degrees $dg(s)^{R_o^{\pm}\varphi}$ will stricly reduce to the (level of) reliability $R_O^{\pm}$ of the source at state s [see Samet 1975, 200]. For this reason, I associate a *fixed degree* to all the $R_O^{\pm}$-operations performed at states of degree 6. These degrees will reflect the *gain* or *loss* of credibility implied by the intended operation.

In case formula φ is not evaluable at the state s under consideration ($dg(s) = 6$), operation $R_O^A\varphi$ is the update according to which all φ-states of the initial ordering $\leqslant_O$ go to the *first* rank in the new ordering $\leqslant_O^{[R_o^A\varphi]}$ while all $\neg\varphi$-states of the initial ordering $\leqslant_O$ go to the *fifth* rank in the new ordering $\leqslant_O^{[R_o^A\varphi]}$. Numerically, operation $R_O^A\varphi$ is defined by:

$$dg(s)^{R_o^A\varphi} \quad = \quad \begin{cases} 1 & \textit{if } \mathbb{S}, s \models \varphi. \\ 5 & \textit{if } \mathbb{S}, s \models \neg\varphi. \end{cases}$$

When formula φ is not assessable at state s, operation $R_O^B\varphi$ makes the states of the initial ordering $\leqslant_O$ go to the *second* rank in the new ordering $\leqslant_O^{[R_o^B\varphi]}$ if these states satisfy formula φ, and to the *fourth* rank in the new ordering $\leqslant_O^{[R_o^B\varphi]}$ if these states satisfy $\neg\varphi$. In terms of credibility degrees, operation $R_O^B\varphi$ is defined by:

$$dg(s)^{R_o^B\varphi} \quad = \quad \begin{cases} 2 & \textit{if } \mathbb{S}, s \models \varphi. \\ 4 & \textit{if } \mathbb{S}, s \models \neg\varphi. \end{cases}$$

Operation $R_O^C\varphi$ is the update according to which all the states of the initial ordering $\leqslant_O$ go to the *third* rank in the new ordering $\leqslant_O^{[R_o^C\varphi]}$ no matter whether these states

satisfy formula φ or satisfy formula $\neg\varphi$. Numerically, operation $\boldsymbol{R}^{\mathrm{C}}_{O}\varphi$ is defined by:

$$\boldsymbol{dg}(s)^{\boldsymbol{R}^{\mathrm{C}}_{o}\varphi} \quad = \quad 3 \;\; \textit{if } \mathbb{S}, s \models \varphi \textit{ or if } \;\mathbb{S}, s \models \neg\varphi.$$

Operation $\boldsymbol{R}^{\mathrm{D}}_{O}\varphi$ makes the states of the initial ordering $\leqslant_{O}$ go to the *fourth* rank in the new ordering $\leqslant_{O}^{[\boldsymbol{R}^{\mathrm{D}}_{o}\varphi]}$ if these states satisfy formula φ, and to the *second* rank in the new ordering $\leqslant_{O}^{[\boldsymbol{R}^{\mathrm{D}}_{o}\varphi]}$ if these states satisfy $\neg\varphi$. In terms of credibility degrees, operation $\boldsymbol{R}^{\mathrm{D}}_{O}\varphi$ is defined by:

$$\boldsymbol{dg}(s)^{\boldsymbol{R}^{\mathrm{D}}_{o}\varphi} \quad = \quad \begin{cases} 4 & \textit{if } \mathbb{S}, s \models \varphi. \\ 2 & \textit{if } \mathbb{S}, s \models \neg\varphi. \end{cases}$$

Operation $\boldsymbol{R}^{\mathrm{E}}_{O}\varphi$ is the update according to which all φ-states of the initial ordering $\leqslant_{O}$ go to the *fifth* rank in the new ordering $\leqslant_{O}^{[\boldsymbol{R}^{\mathrm{E}}_{o}\varphi]}$ while all $\neg\varphi$-states of the initial ordering $\leqslant_{O}$ go to the *first* rank in the new ordering $\leqslant_{O}^{[\boldsymbol{R}^{\mathrm{E}}_{o}\varphi]}$. In terms of degrees, operation $\boldsymbol{R}^{\mathrm{E}}_{O}\varphi$ is defined by:

$$\boldsymbol{dg}(s)^{\boldsymbol{R}^{\mathrm{E}}_{o}\varphi} \quad = \quad \begin{cases} 5 & \textit{if } \mathbb{S}, s \models \varphi. \\ 1 & \textit{if } \mathbb{S}, s \models \neg\varphi. \end{cases}$$

Finally, operation $\boldsymbol{R}^{\mathrm{F}}_{O}\varphi$ is the update according to which all the states of the initial ordering $\leqslant_{O}$ go to the *sixth* rank in the new ordering $\leqslant_{O}^{[\boldsymbol{R}^{\mathrm{F}}_{o}\varphi]}$, — no matter if these states satisfy formula φ or satisfy formula $\neg\varphi$. Numerically, operation $\boldsymbol{R}^{\mathrm{F}}_{O}\varphi$ is defined by:

$$\boldsymbol{dg}(s)^{\boldsymbol{R}^{\mathrm{F}}_{o}\varphi} \quad = \quad 6 \;\; \textit{if } \mathbb{S}, s \models \varphi \textit{ or if } \;\mathbb{S}, s \models \neg\varphi.$$

4.3.3.2 Matching Updates of Degrees with Reliability Ratings

Now I propose to match the various reliability operators with the doctrinal reliability ratings ranging from **A** to **F**. Intuitively, the more reliable officers judge a source of information to be, the more they will favor contextual evidence that is

consistent with the message from this source. On the contrary, the less reliable officers judge a source to be, the more they will favor contextual evidence that is inconsistent with the message from this source. When officers are unable to assess the reliability of a source, they keep credibility ratings as they are. My proposal to match reliability operators $\boldsymbol{R}^{\pm}_{O}$ with doctrinal reliability ratings is based on those intuitions. In model $\mathbb{S}$, *favoring* or *dismissing* pieces of evidence that are consistent or inconsistent with a message amounts to *decreasing* or *increasing* the rank of their corresponding possible state.

The criteria for deciding whether sources are *reliable*, *unreliable* or *unassessable*, depends on their *informational pedigree*. Levels of reliability officers may put into sources are determined by the *truth* or *falsity* of all the information they have provided in the past. That being determined, officers know how much they should promote or dismiss pieces of evidence that back up or contradict the message m they are evaluating.

Following doctrinal descriptions (see Table 4.2), sources are classified as "Completely Reliable" and rated **A** if they *have a history of complete reliability* such that there is *no doubt of authenticity and trustworthiness* towards them. In such conditions, the officer is *strongly justified* to promote the states $\boldsymbol{s}$ satisfying some evidence ε that is consistent with message m being true and to dismiss the states $\boldsymbol{s}$ satisfying some evidence ε that are consistent with m being false. Accordingly, I offer to match rating **A** with reliability operator $\boldsymbol{R}^{\mathrm{A}}_{O}$ from $\mathcal{L}_{(\mathtt{intel})}$. In case message m was already evaluable at state $\boldsymbol{s}$ (because *some distinct evidence* ε exists for assessing m, i.e. $\boldsymbol{dg}(\boldsymbol{s}) < 6$), the degree of state $\boldsymbol{s}$ *decreases by* 3 if $\boldsymbol{s}$ is consistent with m being true: $\boldsymbol{dg}(\boldsymbol{s})^{\boldsymbol{R}^{\mathrm{A}}_{o}\varphi} = \mathtt{Cut}(\boldsymbol{dg}(\boldsymbol{s}) - 3)$ *if* $\mathbb{S}, \boldsymbol{s} \models m$. On the contrary, the degree of state $\boldsymbol{s}$ *increases by* 3 if $\boldsymbol{s}$ is consistent with m being false: $\boldsymbol{dg}(\boldsymbol{s})^{\boldsymbol{R}^{\mathrm{A}}_{o}\varphi} = \mathtt{Cut}(\boldsymbol{dg}(\boldsymbol{s}) + 3)$ *if* $\mathbb{S}, \boldsymbol{s} \models \neg m$. In case message m was not already evaluable at state $\boldsymbol{s}$ (because *no distinct piece of evidence* ε exists for assessing m, i.e. $\boldsymbol{dg}(\boldsymbol{s}) = 6$), the degree of state $\boldsymbol{s}$ becomes 1 if $\boldsymbol{s}$ satisfies m: $\boldsymbol{dg}(\boldsymbol{s})^{\boldsymbol{R}^{\mathrm{A}}_{o}\varphi} = 1$ *if* $\mathbb{S}, \boldsymbol{s} \models m$, and becomes 5 if $\boldsymbol{s}$ satisfies $\neg m$: $\boldsymbol{dg}(\boldsymbol{s})^{\boldsymbol{R}^{\mathrm{A}}_{o}\varphi} = 5$ *if* $\mathbb{S}, \boldsymbol{s} \models \neg m$.

This clause expresses the strongest trusting attitude the officer may have towards

the intelligence source. But this strength can be relaxed to express weaker trusting attitudes. We know that officers classify sources as "Usually Reliable" and rate them **B** if they *have a history of valid information most of the time* such that there is *minor doubt about their authenticity and trustworthiness*. Now officers are *moderately justified* to promote m-consistent states on top of their plausibility ordering. This level can be captured by operator $\boldsymbol{R}^{\mathrm{B}}_{O}$ from $\mathcal{L}_{(intel)}$. In case message m was already evaluable at state $\boldsymbol{s}$ (i.e. $\boldsymbol{dg}(\boldsymbol{s}) < 6$), the degree of state $\boldsymbol{s}$ *decreases by* 2 if $\boldsymbol{s}$ is consistent with m being true: $\boldsymbol{dg}(\boldsymbol{s})^{\boldsymbol{R}^{\mathrm{B}}_{o}\varphi} = \mathtt{Cut}(\boldsymbol{dg}(\boldsymbol{s}) - 2)$ *if* $\mathbb{S}, \boldsymbol{s} \models m$, and *increases by* 2 if $\boldsymbol{s}$ is consistent with m being false: $\boldsymbol{dg}(\boldsymbol{s})^{\boldsymbol{R}^{\mathrm{B}}_{o}\varphi} = \mathtt{Cut}(\boldsymbol{dg}(\boldsymbol{s}) + 2)$ *if* $\mathbb{S}, \boldsymbol{s} \models \neg m$. But in case message m was not already evaluable at state $\boldsymbol{s}$ (i.e. $\boldsymbol{dg}(\boldsymbol{s}) = 6$), the degree of state $\boldsymbol{s}$ becomes 2 if it satisfies m: $\boldsymbol{dg}(\boldsymbol{s})^{\boldsymbol{R}^{\mathrm{B}}_{o}\varphi} = 2$ *if* $\mathbb{S}, \boldsymbol{s} \models m$, and becomes 4 if it satisfies $\neg m$: $\boldsymbol{dg}(\boldsymbol{s})^{\boldsymbol{R}^{\mathrm{B}}_{o}\varphi} = 4$ *if* $\mathbb{S}, \boldsymbol{s} \models \neg m$.

By contrast, the description provided with rating "Fairly Reliable" is more mixed: sources are rated **C** if they *have provided valid information in the past* but there is some *doubt concerning their authenticity and trustworthiness*. Because of this, officers should make a *minor revision* of the plausibility ordering. This corresponds to reliability operator $\boldsymbol{R}^{\mathrm{C}}_{O}$ in $\mathcal{L}_{(intel)}$. In case message m was already evaluable at the state $\boldsymbol{s}$ under consideration (i.e. $\boldsymbol{dg}(\boldsymbol{s}) < 6$), the degree of state $\boldsymbol{s}$ *decreases by* 1 if $\boldsymbol{s}$ is consistent with m being true: $\boldsymbol{dg}(\boldsymbol{s})^{\boldsymbol{R}^{\mathrm{C}}_{o}\varphi} = \mathtt{Cut}(\boldsymbol{dg}(\boldsymbol{s}) - 1)$ *if* $\mathbb{S}, \boldsymbol{s} \models m$, and *increases by* 1 if $\boldsymbol{s}$ is consistent with m being false: $\boldsymbol{dg}(\boldsymbol{s})^{\boldsymbol{R}^{\mathrm{C}}_{o}\varphi} = \mathtt{Cut}(\boldsymbol{dg}(\boldsymbol{s}) + 1)$ *if* $\mathbb{S}, \boldsymbol{s} \models \neg m$. In case message m was not already evaluable at state $\boldsymbol{s}$ (i.e. $\boldsymbol{dg}(\boldsymbol{s}) = 6$), the degree of state $\boldsymbol{s}$ becomes 3 no matter whether it satisfies m or satisfies $\neg m$: $\boldsymbol{dg}(\boldsymbol{s})^{\boldsymbol{R}^{\mathrm{C}}_{o}\varphi} = 3$ *if* $\mathbb{S}, \boldsymbol{s} \models m$ *or if* $\mathbb{S}, \boldsymbol{s} \models \neg m$.

According to the doctrine, sources are classified as "Not Usually Reliable" and rated **D** if there is *significant doubt concerning their authenticity and trustworthiness* but they *have provided valid information in the past*. Expression "*significant doubt*" clearly indicates that states which satisfy inconsistent evidence with m become more plausible than states satisfying evidence consistent with m. But this promotion is minimized by the fact that the sources delivered valid information in the past. So the officer is *barely justified* in promoting m-inconsistent states on top of his or her plausibility ordering. The corresponding operation is $\boldsymbol{R}^{\mathrm{D}}_{O}$ in

language $\mathcal{L}_{(intel)}$. In case message m was already evaluable at state s, this operator makes the degree of s *decreases by* 1 if s is consistent with m being false: $dg(s)^{R_o^D\varphi} = Cut(dg(s) - 1)$ *if* $\mathbb{S}, s \models \neg m$, and *increases by* 1 if s is consistent with m being true: $dg(s)^{R_o^D\varphi} = Cut(dg(s) + 1)$ *if* $\mathbb{S}, s \models m$. In case message m was not already evaluable at state s, the degree of state s becomes 1 if it satisfies $\neg m$: $dg(s)^{R_o^D\varphi} = 1$ *if* $\mathbb{S}, s \models \neg m$, and becomes 5 if it satisfies m: $dg(s)^{R_o^D\varphi} = 5$ *if* $\mathbb{S}, s \models m$.

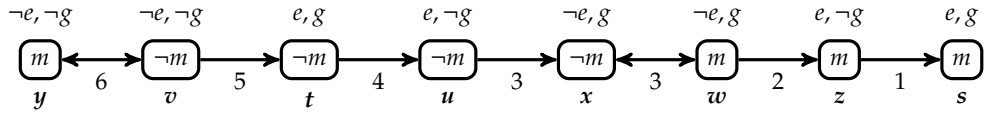
Concerning the rating "Unreliable", the doctrine states that sources are classified **E** in case they are *lacking in authenticity and trustworthiness* and have an *history of invalid information*. Now officers are *highly justified* to promote m-inconsistent states on top of their plausibility ordering. But they are not as justified as they were to promote m-consistent states in case of rating **A**. In the latter case, the doctrinal description indicates sources have an history of *complete* reliability. In case of rating **E**, the description is weaker: sources are not trustworthy and have provided invalid information in the past but they are not described as *completely* unreliable. For this reason, officers are not strongly justified to promote m-inconsistent states on top of the plausibility ordering. This moderate attitude can be captured by operator R_O^E in $\mathcal{L}_{(intel)}$. In case message m was already evaluable at state s, this operator makes the degree of s *decreases by* 2 if s is consistent with m being false: $dg(s)^{R_o^E\varphi} = Cut(dg(s) - 2)$ *if* $\mathbb{S}, s \models \neg m$, and *increases by* 2 if s is consistent with m being true: $dg(s)^{R_o^E\varphi} = Cut(dg(s) + 2)$ *if* $\mathbb{S}, s \models m$. In case message m was not already evaluable at state s, the degree of state s becomes 2 if it satisfies $\neg m$: $dg(s)^{R_o^E\varphi} = 2$ *if* $\mathbb{S}, s \models \neg m$, and becomes 4 if it satisfies m: $dg(s)^{R_o^E\varphi} = 4$ *if* $\mathbb{S}, s \models m$.

Eventually, rating **F** is ascribed when the reliability of the source "Cannot Be Judged". That is: when *no basis exists for evaluating the reliability of the source*. In such a case, the evidence the officer has are *all false* and for this reason, they do not provide a reasonable ground for adjudicating on the truth value of message m. A corresponding operator in $\mathcal{L}_{(intel)}$ is R_O^F. No matter if the message was already evaluable at state s, this operator keeps the degree of s as it was if s satisfies m or satisfies $\neg m$: $dg(s)^{R_o^F\varphi} = dg(s)$ *if* $\mathbb{S}, s \models m$ *or if* $\mathbb{S}, s \models \neg m$. If m was not evaluable

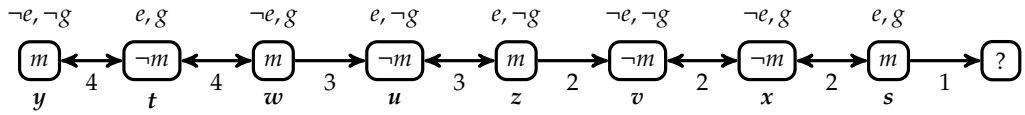
at state s (i.e. $dg(s) = 6$), the degree of s remains equal to 6 after operation R^{F}_{O} is performed in that case. Table 4.8 sums up all the semantic clauses I have given for expressing reliability ratings *when the message is evaluable.*

4.3.3.3 General Case Study: *a follow-up*

Let us remind the model expressing a prior distribution of ratings for message m based on the evidence set $\mathcal{E} = \{e, g\}$ (see subsection 4.3.2):



Suppose now that the officer gives rating **D** to the source of the message. The source is *not usually reliable*, there is *significant doubt* concerning his or her honesty even though they have provided valid information about submarines in the past. Based on rating **D**, the officer can mark a *posterior credibility distribution* for message m. Once applied, the scoring rules that correspond to relability rating R^{D}_{O} return the following updated model:



For instance, the (prior) credibility rating of message m was **2** at state z in the initial model. Now, the *posterior* credibility score of m is **3** at state z in the updated model. This suggests that in state z, m is not "Probably True" as it seemed initially but more certainly "Possibly True". Accordingly, the officer should be more careful than he or she was concerning the truth of message m.

Table 4.9 sums up the update operation. However, two issues can be observed after the computation. Although posterior scores are mutually consistent, the rating **1** is now *undefined*. Besides that, some states, namely $u^{\{e,\neg g,\neg m\}}$ and $z^{\{e,\neg g,m\}}$, are no longer distinguishable. They receive the same credibility score of **3** that make them indistinguishable. In other words, some credibility information has been lost during the process. But in fact some credibility information has also been gained during the operation. States $x^{\{\neg e,g,\neg m\}}$ and $w^{\{\neg e,g,m\}}$ that were indistinguishable

Rating	Label	Operations on Degrees
$\boldsymbol{R}_o^{\mathrm{A}}\varphi$	Completely Reliable	$\boldsymbol{Cut}(\boldsymbol{dg}(s) - 3)$ *if* $\mathbb{S}, s \models \varphi$ $\boldsymbol{Cut}(\boldsymbol{dg}(s) + 3)$ *if* $\mathbb{S}, s \models \neg\varphi$
$\boldsymbol{R}_o^{\mathrm{B}}\varphi$	Usually Reliable	$\boldsymbol{Cut}(\boldsymbol{dg}(s) - 2)$ *if* $\mathbb{S}, s \models \varphi$ $\boldsymbol{Cut}(\boldsymbol{dg}(s) + 2)$ *if* $\mathbb{S}, s \models \neg\varphi$
$\boldsymbol{R}_o^{\mathrm{C}}\varphi$	Fairly Reliable	$\boldsymbol{Cut}(\boldsymbol{dg}(s) - 1)$ *if* $\mathbb{S}, s \models \varphi$ $\boldsymbol{Cut}(\boldsymbol{dg}(s) + 1)$ *if* $\mathbb{S}, s \models \neg\varphi$
$\boldsymbol{R}_o^{\mathrm{D}}\varphi$	Not Usually Reliable	$\boldsymbol{Cut}(\boldsymbol{dg}(s) + 1)$ *if* $\mathbb{S}, s \models \varphi$ $\boldsymbol{Cut}(\boldsymbol{dg}(s) - 1)$ *if* $\mathbb{S}, s \models \neg\varphi$
$\boldsymbol{R}_o^{\mathrm{E}}\varphi$	Unreliable	$\boldsymbol{Cut}(\boldsymbol{dg}(s) + 2)$ *if* $\mathbb{S}, s \models \varphi$ $\boldsymbol{Cut}(\boldsymbol{dg}(s) - 2)$ *if* $\mathbb{S}, s \models \neg\varphi$
$\boldsymbol{R}_o^{\mathrm{F}}\varphi$	Cannot Be Judged	$\boldsymbol{dg}(s)$ *if* $\mathbb{S}, s \models m$ *or if* $\mathbb{S}, s \models \neg m$

Table 4.8: Semantic Conditions for Reliability Ratings.

beforehand can now be differentiated. Furthermore, states $v^{\{\neg e,\neg g,\neg m\}}$ and $y^{\{\neg e,\neg g,m\}}$ are no longer evaluative blindspots for the officer.

States	Evidence Set	Posterior Rating $C^n_O m$
x,s,v	$\{\neg e,g\}, \{e,g\}, \{\neg e, \neg g\}$	$\neg C^2_O m \wedge \neg C^2_O \neg m$
u,z	$\{e,\neg g\}$	$\neg C^3_O m \wedge \neg C^3_O \neg m$
t,w,y	$\{e,g\}, \{\neg e, g\}, \{\neg e, \neg g\}$	$\neg C^4_O m \wedge \neg C^4_O \neg m$

Table 4.9: Evidence-Based *Posterior* Ratings.

4.4 Discussion

4.4.1 The Scoring Operation: *Before* and *After*

The procedure I have defined in numerical belief revision is consistent with experimental results on intelligence evaluation. On the one hand, credibility is the *crucial dimension* for evaluating messages. For a given message m as well as a set of evidence ε for or against this message, plausibility orderings help determine a *prior credibility distribution* for m. On the other hand, reliability ratings play a *balancing role* on this distribution. They help mark posterior ratings, or *resultant scores*, for message m. Let $C_O(m)$ and $R_O(m)$ stand for specific *credibility* and *reliability ratings* for m, and $Sc_O(m)$ stands for the *resultant score* that corresponds to m.

Basically, *two steps* can be distinguished along the scoring operation. The first step occurs *before marking a score*: the officer elicits a credibility rating $C_O(m)$ as well as a reliability rating $R_O(m)$ for m. I have explained that determining the credibility and reliability of message m is determining its *degree of truth* as well as the *degree of honesty* of its source (see Chapter 3). Contextual evidence is used in the first case while historical evidence serves in the second case. But a question that we can naturally ask about the numerical procedure is as follows: *what is the correspondence between the evaluative dimensions of credibility and reliability and*

the underlying descriptive dimensions of truth and honesty? How do these dimensions combine with respect to informational types?

The second step of the scoring operation occurs *once a score has been marked* on message *m*. Resultant scores mirror the different ways in which some officer may change his or her confidence in *m* with respect to its *posterior* credibility score. In fact, this score expresses a *weaker* or a *stronger* credibility rating for the message based on the reliability of its source. Another question we can naturally ask is therefore: *what is the correspondence between resultant scores and informational types? What kinds of information each provide and how do these pieces of information correlate?*

The remaining part of this section addresses these questions. I first investigate the correspondence between ratings and message types. I then discuss the correspondence between resultant scores and types.

4.4.2 The Correspondence between Ratings and Types

The way they are defined, credibility ratings express *decreasing degrees of confirmation* for the intelligence message based on the evidence the officer has for evaluating its *degree of truth* or *falsity* (see Table 4.1). Similarly, reliability ratings express *decreasing degrees of trustworthiness* regarding the source based on growing suspicion concerning her *degree of honesty* or *dishonesty* (see Table 4.2). Looking at the descriptions given by the intelligence doctrine, credibility and reliability ratings can be distinguished along a 3×3 matrix in the same way degrees of truth and honesty were combined for giving a 3×3 matrix that described message types. Let us start out by reminding the 3×3 matrix we proposed in Chapter 3 for classifying message types:

Based on doctrinal descriptions, a similar matrix can be provided that *partitions* the evaluative space into 3 categories of credibility ratings and 3 categories of reliability ratings. Regarding credibility, a first category concerns messages that are judged as *true* by officers: rating **1** ("Confirmed"), rating **2** ("Probably True") and rating **3** ("Possibly True"). In these cases, the evidence officers have is suf-

Message type t		***Truth of the Content***		
		True	*Indeterminate*	*False*
Honesty of the Source	*Honest*	t_1 = *information*	t_5 = *error-avoidance*	t_2 = *misinformation*
	Imprecise	t_7 = *negative omission*	t_9 = *mixed*	t_8 = *positive omission*
	Dishonest	t_3 = *subjective lie*	t_6 = *half-truth*	t_4 = *objective lie*

Table 4.10: Informational Labels for Message Types.

ficient for making them believe that message contents are more clearly *true* than *false*, or even *uncertain*. By contrast, a second category can be teased apart in which contents are judged as clearly *false* by officers: rating **5** ("Improbable"). In rating **5**, the message is contradicted by the evidence officers have on the subject of interest. Finally, a third category can be isolated in which contents are judged as being neither *true* nor *false*, but clearly *uncertain*. This is rating **4** ("Doubtfully True"). I have insisted on the specificity of rating **4** by comparing the discrepancy between the label "Doubtfully True" and its intended description by the intelligence doctrine (see section 4.1 and subsection 4.3.2). Despite this description, the label "Doubtfully True" clearly indicates that the message may be *true* but may also be *false*, so that that its truth status is *indeterminate* in that case.

Regarding reliability ratings, three categories can also be distinguished depending on the degree of honesty of the source. The first category concerns rating **A** ("Com-

pletely Reliable"), rating **B** ("Usually Reliable") and rating **C** ("Fairly Reliable"). Even though these ratings express increasing suspicion towards sources being honest, this suspicion is very limited. Sources are considered as being clearly *honest* for having provided valid information in the past. But as for credibility, this category can be opposed to a second category in which sources are clearly *dishonest*: this is rating **E** ("Unreliable"). In this situation, sources lack in trustworthiness and have a history of invalid information such that there is no doubt of their dishonesty. Finally, an intermediary category can be isolated in which sources are *imprecise* regarding the quality of the information they provide: rating **D** ("Not Usually Reliable"). When classified as **D**, sources are clearly less cooperative than they should according to Gricean principles.

These three categories of credibility ratings $\boldsymbol{C}_o(m)$ and three categories of reliability ratings $\boldsymbol{R}_o(m)$ *partition* the evaluative space into nine zones. But since eliciting ratings for m is determining the *degree of truth* of m as well as the *degree of honesty* of its source, these elicitations are elicitations of informational types. More precisely, pairs of ratings, written $\langle \boldsymbol{C}_o, \boldsymbol{R}_o \rangle$, can be seen as *subjective assessments* of message types based on contextual and historical evidence. Accordingly, the nine zones I have identified in the evaluative space can be matched by the nine informational types I have identified in the descriptive space. The correspondence between rating pairs and informational types is given by Table 4.11. For simplicity's sake, I leave aside rating **6** and rating **F** in which contents and/or sources cannot be judged. In these cases, types cannot be judged either.

4.4.3 The Correspondence between Types and Scores

Once credibility and reliability ratings have been elicited by the officer, the latter marks a resultant score for message m. Table 4.12 indicates all the *posterior* scores that can be derived by applying reliability updates $\boldsymbol{R}_o$ on *prior* distributions of ratings $\boldsymbol{C}_o$. Consider for instance update **D**. Once applied on prior credibility rating **2**, rating **D** returns a posterior credibility score of **3**. But once applied on rating **4**, **D** returns a posterior credibility rating of **5**, *etc*.

As constructs, resultant scores differ from types. Types are *descriptive accounts*

Correspondence *Ratings & Types*		***Content Credibility*** C_O		
		1 | 2 | 3	**4**	**5**
Source Reliability R_O	**A** **B** **C**	*information*	*error-avoidance*	*misinformation*
	D	*negative omission*	*mixed*	*positive omission*
	E	*subjective lie*	*half-truth*	*objective lie*

Table 4.11: The Correspondence between Ratings and Types.

Correspondence *Scores & Types*		***Content Credibility*** C_O		
		1 | 2 | 3	**4**	**5**
Source Reliability R_O	**A**	**1 | 1 | 1**	**1**	**2**
	B	**1 | 1 | 1**	**2**	**3**
	C	**1 | 1 | 2**	**3**	**4**
		information	*error-avoidance*	*misinformation*
	D	**2 | 3 | 4**	**5**	**5**
		negative omission	*mixed*	*positive omission*
	E	**3 | 4 | 5**	**5**	**5**
		subjective lie	*half-truth*	*objective lie*

Table 4.12: The Correspondence between Scores and Types.

of intelligence message based on dimensions of *truth* and *honesty* that are not conflated with each other. Indeed, dimensions of *truth* and *honesty* are not reduced to one another or confounded into a single dimension. On the contrary, resultant scores are *evaluative accounts* of messages based on a *credibility score* that also integrates the *reliability of the source*. In that case, reliability is an adjustment of credibility.

Despite being different constructs, scores and types converge. They provide complementary information concerning the quality of intelligence messages. I have shown that pairs of credibility and reliability ratings for m, written $\langle \boldsymbol{C}_{\text{O}}(m), \boldsymbol{R}_{\text{O}}(m) \rangle$, are assessments of messages types $\boldsymbol{t}(m)$. But on the other hand, pairs of ratings for m also determine the resultant score of m: $\boldsymbol{Sc}_{\text{O}}(m)$. From these pairs, we can finally associate resultant scores with types (see Table 4.12). To see more clearly into the link between scores and types, I propose to calculate the *average posterior credibility scores* of the 9 evaluative zones. For instance, the average score of the type *error-avoidance* is **2**: $\frac{1+2+3}{3} = 2$. The average score of the type *subjective-lie* is **4** ($\frac{3+4+5}{3} = 4$) while the average score of the type *positive omission* is simply **5**, etc. All the different calculations are summed up in Table 4.13.

Looking at average scores, we observe that posterior credibility scores are consistent with informational types. The more qualitative the type, the higher its corresponding posterior score. On the contrary, the less qualitative the type, the lower its posterior score. This correspondence can be easily explained: resultant scores indicate the rational course of action to follow after having considered the ratings of the messages. In that sense, the more beneficial the message is based on its informational type, the better its resultant score. See for instance the types *information* and *error-avoidance*: their corresponding scores are **1.1** and **2**. By contrast, the more harmful or detrimental the message is based on its type, the worse its resultant score. See for instance the types *half-truth* and *objective lie*: both have corresponding scores of **5**. Accordingly, scores can be seen as a rational basis for action based on the message being *classified* as *such* or *such*.

However, Table 4.13 also shows that quantitative scores happen to be *less sensitive*

Average Scores / Types		*Content Credibility* C_o		
		1 | 2 | 3	**4**	**5**
Source Reliability R_o	**A** **B** **C**	**1.1** *information*	**2** *error-avoidance*	**3** *misinformation*
	D	**3** *negative omission*	**5** *mixed*	**5** *positive omission*
	E	**4** *subjective lie*	**5** *half-truth*	**5** *objective lie*

Table 4.13: Average Scores for Informational Types.

than qualitative types when credibility ratings are higher than **4** and reliability ratings higher than **D**. From a qualitative perspective, pairs of ratings $\langle \mathbf{4}, \mathbf{D} \rangle$, $\langle \mathbf{5}, \mathbf{D} \rangle$, $\langle \mathbf{4}, \mathbf{E} \rangle$ and $\langle \mathbf{5}, \mathbf{E} \rangle$ can be teased apart since they correspond to clear-cut informational types: *mixed*, *positive omission*, *half-truth* and *objective lie* respectively. But these pairs can no longer be teased apart from a quantitative perspective since the messages they correspond to all get a posterior score of **5**. This qualitative/quantitative discrepancy is illustrated by Figure 4.1 in which all the informational types are ranked from *best to worst* based on their average scores.

Figure 4.1 shows that negative types can no longer be teased apart. This reflects the loss of credibility information we observed in the practical case study: states that were distinguishable before the scoring operation become indistinguishable afterwards. This issue is caused by the recipes I have proposed for reliability updates. The way I define them, they return less plausibility degrees than the prior distributions they are applied to. Similar difficulties were identified by

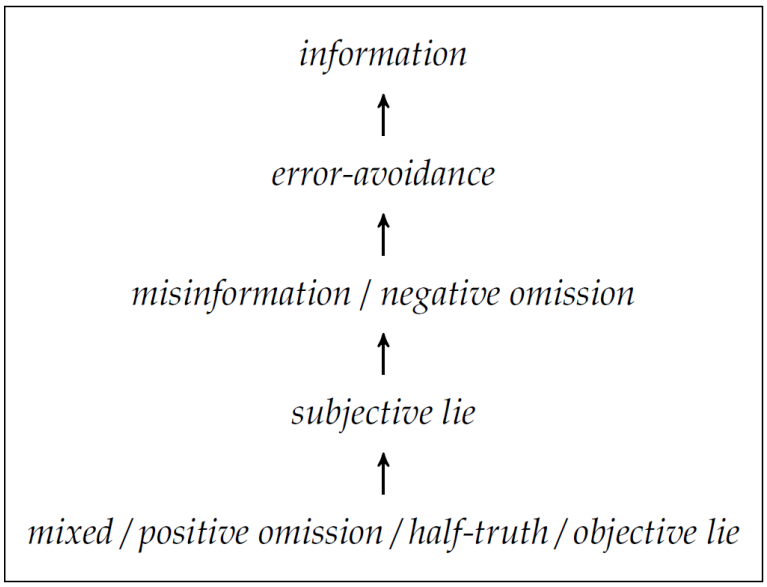


Figure 4.1: Qualitative Ranking Based on Average Scores.

Aucher [2008] and van Ditmarsch [2008] concerning quantitatively formulated proposals for belief revision. Finer-grained recipes are needed to increase the discriminative power of the scoring procedure.

4.5 Conclusion

We have seen that the doctrinal procedure for intelligence evaluation is not adequately specified. Researchers have identified two fallacious assumptions in the existing procedure. One of these assumptions is that "*intelligence reports should transmit facts and/or assessments*" and that the "*the distinction between fact and interpretation [should] always be clearly indicated*" [see STANAG-2511 2003, 2]. But another incorrect assumption is that "*reliability and credibility, the two aspects of evaluation, must be considered independently of each other*" [see STANAG-2511 2003, A-2]. However, empirical results show that this independence is not effective. Field officers perceive dimensions of credibility and reliability as being strongly correlated. More specifically, credibility is seen as *prevalent* over reliability in the resultant scores.

This chapter aimed to make the evaluative procedure compatible with this prevalence. I proposed to formalize information evaluation in a numerical setting in-

spired by previous works in quantitative belief revision [e.g. Aucher 2004, van Ditmarsch 2005, van Ditmarsch & Labuschagne 2007]. Most of the extant settings for evaluating data mix qualitative features with quantitative ones. In particular, symbolic procedures based on many-valued logics always have a quantitative flavour [e.g. Akdag *et al.* 1992, Revault d'Allonnes *et al.* 2007, Revault d'Allonnes & Lesot 2014]. By contrast, symbolic procedures based on modal logics do not have a quantitative flavour [see Demolombe 2004, Cholvy 2013], even though new steps have been made in this direction based on extensions of modal logics [e.g. Legastelois *et al.* 2017ab].

The goal of this chapter was to bridge the gap between *modal logic proposals* and *quantitative approaches*. In the proposed setting, dimensions of credibility and reliability are captured syntactically through static and dynamic modal operators. Credibility operators are indexed by numerical degrees ranging from 1 to 6, which correspond to the distinct levels of the STANAG scale. These degrees are created semantically by putting different constraints on the set of the most plausible states. Reliability operators are numerical rules to update these degrees depending of the reliability of the source. With respect to credibility operators, my setting is essentially *qualitative* since numerical degrees are derived from a qualitative ordering. But with respect to reliability, my setting is more *quantitative* than *qualitative* since numerical features dominate in the updates. Let us now compare my proposal with extant quantitative and qualitative proposals.

At the *quantitative* level, my setting is based on absolute degrees. Since experiments revealed disagreement concerning probability degrees, I have continued to use plain degrees, contrary to most of the quantitative proposals, especially those based on Bayesian analysis and networks [e.g. Zlotnick 1972, Fisk 1972, Schweitzer 1978, Schum 1987, Barbieri 2013, Blasch *et al.* 2013], or on the Dempster-Shafer' theory of evidence [e.g. Nimier & Appriou 1995, Nimier 2005, Cholvy 2004 2010, Pichon *et al.* 2012]. The loss of expressivity I observed in my proposal may have resulted from keeping plain degrees. Without moving to probability degrees, adding new discrete values or moving to fuzzy degrees may solve the issue. I also showed that the loss of information was limited since my setting eventually

leads to recover informational types.

At the *qualitative* level, my proposal has the advantage of being of practical use to intelligence personnel. Officers can better appreciate the quality of messages by putting them into actionable categories (viz. *information*, *misinformation*, *lies*, etc.). But let us draw a closer comparison between my proposal and the existing qualitative ones.

As in *multi-valued logic* proposals [e.g. Seridi & Akdag 2001, Revault d'Allonnes *et al.* 2007, Revault d'Allonnes & Lesot 2014 2015], I define credibility degrees on a discrete scale to match the extant doctrine. Similarly, the credibility dimension is assumed as the *dominant* dimension, but multi-valued proposals differ with respect to the subdimensions they take into account (such as the *reliability* of the source, his or her *competence*, *sincerity*, etc.), and with respect to the sequential procedure which integrates those dimensions [e.g. Revault d'Allonnes & Lesot 2014]. In multi-valued proposals, semantic degrees encode the epistemic uncertainty attached to *message contents*, *sources*, *context* or *semantic formulations* of messages,[7] whereas I use graded operators to express the uncertainty tied to the credibility of messages. In both proposals, however, combination rules return a *unique* posterior credibility degree. In both cases also, *prior* and *posterior* degrees are defined in semantic structures but degrees of plain certainty/uncertainty, as well as extra degrees of indifference and ignorance, are primitive in multi-valued proposals whereas they are derived from the plausibility ordering in my own. To end with, numerical rules are generalised from logical operations in multi-valued proposals (such as *conjunction*, *disjunction* or *implication*) while numerical rules are updates of the plausibility ordering in mine.

Compared to extant proposals in *modal logics* [e.g. Demolombe 2004, Cholvy 2013], my setting is directly intended to capture the credibility and reliability dimensions of the STANAG doctrine. Existing proposals are more focused on giving *conjunctive definitions* of these dimensions, as well as to subdimensions used for charaterizing them, — such as the ability of sources to be *truthful* (or *valid*) as well as *maximally informative* (or *complete*). By duality, sources are characterized as *misinformers* or as *disinformers* when they fail to meet those requisits. Contrary

[7] See Lesot & Revault d'Allonnes [2017] on these different aspects of uncertainty in many-valued proposals.

to these proposals, I have offered a logical syntax that includes static operators for credibility and dynamic operators for reliability. The dynamic part of the evaluation is mirrored directly in the syntax and not performed through syntactic derivations in axiomatic systems provided with rules.

Conclusion

This thesis has been investigating various deceptive attitudes that range from the standard case of *lies* to the non-standard cases of *misleading defaults* and *strategic omissions*. I have combined conceptual, formal and experimental resources to see more clearly into the *definition* as well as the *evaluation* of these deceptive attitudes.

1. Summary of the Chapters

Chapter 1 was entitled *Two Definitions of Lying*. Following common intuitions, a statement is *definitely*, or *typically*, a lie when this statement is not only intended-to-deceive and believed-to-be-false, but also objectively *false*. Then, a dishonest statement must not be *objectively* false to be a lie. However, a typical lie is *more usually* false than true because falsity makes it deceptive in that respect. For this reason, some epistemologists have claimed that the correct definition of lying is the "objective definition" whereby the speaker's utterance is not only subjectively false — as in the traditional, "subjective definition" — but also *objectively* false [e.g. Grimaltos & Rosell 2013, Turri & Turri 2015]. The main lesson of Chapter 1 is that falsity is not mandatory for lying. My experiments have shown that a dishonest utterance is a lie even if it turns out to be objectively *true*. The French expression "*mentir vrai*" can be used in that sense since in true lies, the speaker commits an act of lying by trying to say something *false* (this is the meaning of "*mentir*"), but in fact the outcome turns out to be *true* (this is the meaning of "*vrai*").[8] Accordingly,

[8] This expression was made popular by the French poet and novelist Louis Aragon who used it for describing a fictional method whereby writers transform exact facts to express higher-order truths about the world and human matters [see Aragon 1980]. Here the expression "*mentir vrai*" is used in a different sense for qualifying lies that are intended to be deceitful but happen

dishonest speakers fail to deceive when they say something true. However, they succeed at lying since their deceptive intentions are sufficient in that respect.

Chapter 2 was entitled *The Surprise Deception Paradox*. In Smullyan's story, a question we may ask is whether *Emile* actually lied to *Raymond*. In fact, *Emile*'s announcement is a statement that intends to deceive *Raymond*. In addition to that, this statement is objectively true (since *Raymond* is deceived at the end of the day), even though it triggers a default conclusion that is false (*Raymond* is not deceived by commission). Consistent with the definition of Chapter 1, *Emile*'s announcement meets two of the critera required for the announcement to qualify as a subjective lie. However, dishonesty (or untruthfulness) is not met by *Emile* since he believes his utterance to be true. For this reason, his announcement does not qualify as a subjective lie. The main lesson of Chapter 2 is that, like other pragmatic phenomena such as *false implicatures*, *misleading defaults* are not lies since the speaker's utterance is not insincere although the default conclusion is false or believed-as-false.

If *Emile*'s announcement is not a lie, what is the meaning of the *deception by omission* involved in Smullyan's story? *Emile*'s announcement does not fall into the category of *pragmatic imprecision* since *Emile* does not hide information from *Raymond*. But nor is *Emile*'s announcement *semantically indeterminate* since his statement is not borderline between truth and falsity but clearly true. Then, another lesson of Chapter 2 is that there may be an intermediary category between pragmatic imprecision and semantic indeterminacy we may call "*pragmatic indeterminacy*" and whereby it is pragmatically uncertain whether the speaker's assertion should be interpreted in one way or another. In this regard, it is uncertain whether *Emile*'s announcement should be interpreted as "I will deceive you by making *some action* you do not expect" or as "I will deceive you by *no action* for the reason that you do not expect it". Unfortunately for him, *Raymond* elicits the first interpretation and is deceived.

Chapter 3 was entitled *The Definition of Intelligence Messages* and proposed a taxonomy of deceptive strategies that includes *standard cases*, in particular *objective* and

to be true.

subjective lies, with *non-standard cases*, in particular *omissions*. As in Chapters 1 and 2, dimensions of truth/falsity and of honesty/dishonesty have been considered for characterizing contents and sources. For simplicity's sake, I did not represent the source's intention-to-deceive and the source's utterance in this chapter. Honesty was assumed as a reasonable proxy for representing the source's intention while utterances were taken for granted in the elicited contents. The main lesson of Chapter 3 is that to be integrative of non-standard instances of deception, a taxonomy must considers sources that are borderline between honesty and dishonesty, and contents that are borderline between truth and falsity. So contrary to Chapters 1 and 2, Chapter 3 considers speakers and contents along intermediary dimensions of indeterminacy and imprecision. Imprecision is interpreted as either hiding the truth (*negative omissions*) or as hiding falsity (*positive omissions*), while indeterminacy is interpreted as either preventing false beliefs (*error-avoidance*) or as preventing calibrated beliefs (*half-truths*). Most importantly, this comprehensive account was intended for practical use. I wanted to help intelligence officers give more fine-grained estimates of the messages they were evaluating by putting them into actionable categories. The extant setting for information evaluation did not authorize such epistemological distinctions that turns out to be useful from an instrumental perspective.

Chapter 3 was closely related to Chapter 4. Entitled *A Dynamic Procedure for Information Evaluation*, Chapter 4 proposed a new procedure for assessing intelligence data. This proposal was motivated by empirical results showing the prevalence of the credibility dimension in evaluative aspects. This chapter also offered to bridge the gap between *quantitative* and *qualitative approaches* on information evaluation by proposing a new setting based on numerical plausibility degrees. The main lesson of Chapter 4 is that quality and quantity are complementary and not substituable. Quantitative approaches need qualitative expertise on the epistemic, semantic and pragmatic dimensions of messages. But qualitative approaches also need numerical degrees for capturing the existing features of the intelligence scale.

These four chapters dealt with complementary aspects that are the definition

and evaluation of information, whether this information is deceptive or not. A joint lesson is that these aspects are two sides of the same coin when we aim to understand informational dynamics. Another lesson is that *non-standard cases* are as important as *standard cases* in the study of deception, whether these cases are borderline, like *half-truths* and *omissions*, or uncommon and peripheral, like *misleading defaults*. That being said, many relevant aspects have necessarily been left aside or simply touched upon in this integrative ambition. In addition to that, some of the aspects that have been studied need further analysis. I would like to end this thesis by drawing some perspectives for future work.

2. Future Perspectives

Chapter 1 presented new empirical data showing that falsity is not necessary for lying. In contrast, the speaker's intention to deceive, namely his or her intention to say something false, is necessary for lying. One dimension this chapter did not question is the *statement condition*. I have not investigated whether *attempting* to say something false was also a necessary condition for lying. Is a dishonest speaker who *intends* but *does not attempt* to say something false lying or not in that sense? It seems intuitive that when the intended omission leads the addressee to hold false beliefs, the speaker's omission is indeed a lie. Recently, Wiegmann & Willemsen [2017] and Wiegmann *et al.* [2017] have collected results that are consistent with this intuition. They have observed that intended and dishonest omissions were qualified as lies in case they lead addressees to false implicatures they would have avoided otherwise. This implies that so-called "*lies of omission*" are *lies* if the pragmatic imprecision of the speak leads the addressee to hold false beliefs. One future perspective would be to conduct new experiments to confirm whether utterances are necessary for lying.

Chapter 2 put emphasis on *misleading default inferences* and *deception by omission* through the lenses of the surprise deception paradox. We have seen that *omissions of information* have been related to *false implicatures* by empirical philosophers [see Wiegmann & Willemsen 2017]. Whether the omission is total or partial, omitting information makes the addressee conversationally implicate false information by

hiding the most relevant part of the truth. In that respect, the addressee derives a wrong conclusion because of incomplete beliefs. From a theoretical perspective, further work would help make a closer comparison between *false implicatures* and *misleading defaults* since false default conclusions are also triggered by incomplete beliefs resulting from omissions. Following Caminada [2009] and Sakama *et al.* [2010], the speaker needs to hide relevant information to make the hearer jumps to false conclusions. If the hearer were better informed about the world, he or she would not draw this mistaken conclusion by default reasoning. The link between default inferences and false implicatures is debated in the literature, and so is the difference between misleading defaults and false implicatures. More research is needed to figure out the relations, and differences, between those two.

Chapter 3 proposed a descriptive matrix for defining intelligence messages. Though this matrix is more sensitive than the existing one from a qualitative perspective, there is room for discrimination concerning borderline categories of *indeterminacy* and *imprecision*. We have seen that semantic indeterminacy splits into subcategories of *degree-vagueness vs. open-texture*, whereas pragmatic imprecision splits into subcategories of *generality vs. approximation*. Further discrimination is then possible with respect to the indeterminacy of contents and imprecision of sources. A 4×4 matrix could be proposed based on 4 levels for truth and 4 levels for honesty, with two classical levels (*True vs. False*, *Honest vs. Dishonest*) as well as two borderline levels in both cases (*Degree-vagueness vs. Open-texture*, *Generality vs. Approximation*). That way, officers would be provided with additional categories to make even more specific appreciations of intelligence messages.

But a complementary direction could be to consider further dimensions than truth and honesty for categorizing messages. Quite naturally, two dimensions could be integrated to truth and honesty for giving finer characterizations of lies [following Marsili 2014 2018] but also of non-standard cases of deception. First, the *untruthfulness* of sources, namely the degree to which sources are sincere or insincere, should be considered in pair with their degree of honesty and dishonesty. The more sources believe the messages they deliver to be *false* (*true*), the more they are *untruthful* (*truthful*) and thus, *dishonest* (*honest*) when they deliver

these messages.[9] But considering the untruthfulness of sources implies to move from plain beliefs to graded beliefs in order to express various degrees of certainty for truth and falsity. Another dimension that should be considered for greater expressivity is *assertivity*: how assertive are sources when they deliver the messages they deliver? What is the assertoric force of their utterances in that case? Distinguishing various degrees of intensity in the sources' assertions would also lead to higher discrimination in qualitative aspects.

Chapter 4 presented a procedure for intelligence evaluation that complies with empirical findings on the *prevalence* of the credibility dimension. However, I observed that the quantitative aspects of the procedure were less sensitive than its qualitative features. A discrepancy could be observed between qualitative types and quantitative scores in *posterior* credibility ratings. In future work, I plan to define finer-grained recipes to restore the equilibrium between these two perspectives. Softer calculation rules shall be devised for numerical operations such that *prior* credibility degrees of states that were distinguishable are not made identical once updates are performed.

As a second step, I aim at implementing my proposal as a software that intelligence officers could use to assess new messages. The program should be written in a language rich enough to express the numerical plausibility features of my procedure, and to give a *unique* credibility measure to messages based on two modules, one linked to the credibility of message contents, the other linked to the reliability of their sources. These modules should be supplied by evolving knowledge bases: one containing contextual evidence for cross-checking contents and the other containing information on sources to determine their informational pedigree. However, one major challenge is that the credibility and reliability of messages are interdependent, and so are the resultant scores. In order to succeed, the software should keep constant track of the evolutions of the knowledge bases in order to provide up-to-date ratings and scores.

That being done, It will remain to test the efficiency of the software. Experimental protocols could be divided in relation to the following questions: (1) Does the program give calibrated measures of the dimensions it aims to capture and

[9] On the contrary, the less sources believe the messages they deliver to be *false* (*true*), the more they are *truthful* (*untruthful*) and thus, *honest* (*dishonest*) when they deliver these messages.

aggregate? (2) Is the software tractable for intelligence professionals to use it as a common basis, or should it be simplified? (3) Does the procedure increase the performance of officers making intelligence evaluation? In other words, does recovering informational types help them better appreciate the messages they are evaluating and make sounder decisions? To answer these questions, specific test protocols will be needed. I leave the elaboration of these issues for further work.

List of Figures

List of Tables

Appendix:
Lying and Vagueness[§§]

Paul Égré Benjamin Icard

Abstract

Vagueness is a double-edged sword in relation to lying and truthfulness. In situations of in which a cooperative speaker is uncertain about the world, vagueness offers a resource for truthfulness: it avoids committing oneself to more precise utterances that would be either false or unjustifiably true, and it is arguably an optimal solution to satisfy the Gricean maxims of Quality and Quantity. In situations in which a non-cooperative speaker is well-informed about the world, on the other hand, vagueness can be a deception mechanism. We distinguish two cases of that sort: cases in which the speaker is deliberately imprecise in order to hide information from the hearer; and cases in which the speaker exploits the semantic indeterminacy of vague predicates to produce utterances that are true in one sense, but false in another. Should such utterances, which we call half-truths, be considered lies? The answer, we argue, depends on the context: the lack of unequivocal truth is not always sufficient to declare falsity.

Keywords: lying; deception; vagueness; imprecision; indeterminacy; generality; approximation; open-texture; supervaluations; half-truth; informativeness; subjectivity; equivocation

[§§]Published in *Oxford Handbook of Lying*, J. Meibauer ed., Oxford University Press, 2018.

Lying may be defined as the deliberate utterance of a false sentence (or thought to be false), generally with the aim of misleading the hearer into thinking that the sentence is true. Paradigmatic examples of lies involve sentences expressing *incontrovertibly false* propositions. For example, when former French Minister of Budget Jérôme Cahuzac solemnly declared on December 5, 2012: "*I do not have, Mr Deputee, I never had, any account in a foreign country, neither now nor before*", he made an outright false assertion whose falsity he could no longer deny after investigations found evidence that he had held bank accounts in Switzerland, Singapore, and the Isle of Man. Those investigations quickly led to Cahuzac admitting his lie and resignating.

The Cahuzac scandal is the example of a *blatant lie*: an utterance whose falsity is clear and eventually beyond doubt, including to the speaker. For many of our utterances, however, it is not so clear-cut whether they should be considered a lie or not, even after all the evidence has been collected, and even to the speaker. This happens when utterances are vague. The point is that vague sentences have unclear truth-conditions, and cannot easily be nailed down as false for that matter. Consider horoscopes, and the question of whether they are truthful or not. Suppose my horoscope tells me (an example found on the internet):

(1) Overall, you will feel rather good, physically and morally. You won't be too inhibited this time and...

Is the sentence true or false? The answer is unclear. Characteristic of horoscopes is the exploitation of vagueness. In an expression like *overall*, it is left underspecified exactly what proportion of the time. Similarly, *feel good* is a qualitative predicate for which there is no absolute criterion of application, and likewise for *inhibited*. Further exploitation of vagueness can be found in the use of degree modifiers like *rather* or *too*, whose interpretation is characteristically speaker- and listener-dependent [see Lakoff 1973, Wright 1995]. Moreover, the indexical *this time* leaves its temporal reference open, making simply unclear which context is targeted by the sentence to be counted as true or false.

To philosophers of science, horoscopes are deceitful precisely because they make exploitation of vagueness on such a large scale [Popper 1963]. To casual readers, on the other hand, they are often pleasant to read, because it is easy to

find confirming instances of their truth (horoscopes can be thus argued to exploit a well-documented psychological phenomenon by means of semantic vagueness, the phenomenon of *confirmation bias*, [see Wason 1966]. Such ambivalence suggests that vagueness can be a convenient way of calibrating the truth of an assertion. There is a tradeoff between informativeness and truth, or dually, between vagueness and falsity. That is, the more vague an utterance, the more likely it is to be true relative to some contexts of interpretation, and the less likely it is to be false as a result. The more precise an utterance, on the other hand, the narrower the range of contexts relative to which it can be true. By decreasing informativeness, a vague sentence thus increases its chances of being true [Russell 1923].[1]

This inverse relationship between informativeness and truth is exploited not just by horoscopes, it is a pervasive feature of everyday conversations and exchanges, and it concerns commercial, moral and legal transactions. Consider sales and advertising: like horoscopes, ads generally use vague vocabulary to sell their products. A famous case concerns the firm Ferrero, who used to advertise for its star product as see *healthy*. The firm was sued by a Californian customer on the grounds of making a false claim, considering the high rate of sugar in its product, but the company retorted that "there are health benefits associated with eating chocolate" (more on such moves below). In ordinary exchanges, however, vagueness is not necessarily used to deceive, but simply to avoid making claims that are too committal. Vagueness in that sense is not confined to horoscopes, but concerns predictive utterances quite generally (as in medical communication, see van Deemter 2009 and below). Vagueness is a feature of language that is used to avoid flouting Grice's first Maxim of Quality ("Do not say what you believe to be false / that for which you lack adequate evidence") while exploiting Grice's second Maxim of Quantity ("Don't make your contribution more informative than is required" [see Grice 1975, 45-46].

The goal of this chapter is to clarify the ways in which the use of vague language relates to both of those maxims. Vagueness is a multifaceted notion, however. In the first part of this chapter, we start out by distinguish two main manifestations of vagueness in language: pragmatic *imprecision*, and semantic *indeterminacy*, each of which with more specific varieties. We then go on to explain in what sense

[1] "A vague belief has a much better chance of being true than a precise one, because there are more possible facts that would verify it" [Russell 1923, 91].

vague language is a double-edged sword in relation to lying and truthfulness. First, we show that in situations in which a cooperative speaker wishes to inform about a state of affairs about which she is uncertain, vagueness offers a resource for truthfulness: it avoids making more precise utterances which may be either false or unjustifiably true (section 6.2). In situations in which a non-cooperative speaker is perfectly informed about the world, on the other hand, vagueness can be a deception mechanism. We distinguish two cases of that sort: cases in which the speaker is deliberately imprecise in order to hide information from the hearer, but remains literally truthful (section 6.3); and cases in which the speaker exploits the semantic indeterminacy of vague predicates to make utterances that are true in one sense, but false in another, what we call half-truths (section 6.4). The question is whether such half-truths should be counted as lies. The answer, we suggest, depends on the context: the lack of unequivocal truth is not always sufficient to declare falsity (section 6.5).

6.1 Varieties of vagueness

Russell [1923] offered as a general definition that "a representation is vague when the relation of the representing system to the represented system is not one-one, but one-many". In the linguistic case, an expression is vague according to him if "there is not only one object that a word means, and not only one possible fact that will verify a proposition"[89-90]. That is, the same utterance is compatible with several distinct meanings. This one-many relationship can be realized in several ways, and a merit of Russell's definition is that it covers a range of phenomena associated to linguistic vagueness. In what follows we distinguish four main manifestations: *generality, approximation, degree-vagueness, and open-texture*. Following several authors [see Pinkal 1995, Kennedy 2007, Solt 2015], we argue that generality and approximation are fundamentally cases of pragmatic imprecision, whereas degree-vagueness and open-texture are semantic phenomena, directly affecting the truth-conditions of expressions (see Figure 6.1).

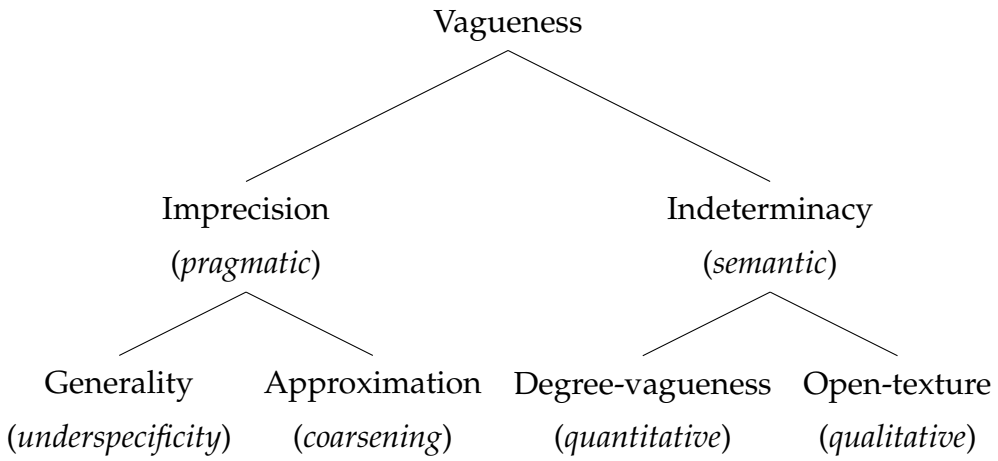


Figure 6.1: Varieties of Linguistic Vagueness

6.1.1 Generality

A first instance of Russell's definition concerns the phenomenon of *generality* in language (being underspecific). Consider the following dialog between a father and his son:

(2) Q. Who did you see at the party?
A. Some friends.

Let us assume that the son has ten friends, all known to the father. The son's answer is general in the sense that it is compatible with several more specific answers being true. The father can't infer from the answer which exact number of friends was seen. An answer like *two friends* would be more informative in that respect, but it would still be general, leaving the father uncertain as to which two-membered subset of the relevant set includes the friends seen by his son.

Important to note is that in this context a sentence like *I saw some friends* has completely clear truth-conditions, it simply means that the number of friends seen by the speaker is greater than zero. Vagueness in that case does not mean any indeterminacy in the statement of the truth-conditions of the sentence, but simply refers to the fact that the response to the question fails to be maximally informative.

Theorists of vagueness often dismiss generality as a central aspect for that matter [see Fine 1975, Keefe 2000]. We find important to keep it into consideration

here, for the underspecificity of answers, although relative to the question under discussion, is a very common aspect of language use of particular relevance in relation to lying.

6.1.2 Approximation

A second illustration of Russell's definition of vagueness pertains to approximation. In cases of approximation an expression with precise truth conditions is used to convey a meaning that differs from its literal meaning, but that is close enough. As a result, the same expression is used with a coarser meaning (larger range of interpretations than its literal meaning). Consider the following dialogues:

(3) Q. What do you make a month?
A. 3,000 euros.

(4) Q. What time did John arrive?
A. He arrived at 3 o'clock.

(5) Q. How old is she?
A. She is 40.

In (3), the answer may be asserted by someone who knows the precise answer to actually be 3287,23 euros. This is a case in which the speaker rounds off the actual number to a lower number, relying on the fact that it is more relevant to set the standard of precision to multiples of 1000 euros than to a multiples of a single euro, let alone cents [see Krifka 2007]. The same often happens with the other two examples: *3 o-clock* can be used when John in fact arrived at five past or five to [van der Henst *et al.* 2002], and *she is 40* could be used to refer to someone whose age is within a few months or even a few years around 40, depending on the context.

Approximation is not limited to numbering, but is also found in other domains, as exemplified in Austin's geometrical example [Austin 1962, Lewis 1979]:

(6) France is hexagonal.

The latter sentence would be false if taken to mean that France has precisely the

shape of a hexagon, but we understand it to mean that it can be circumscribed to a reasonable approximation by a hexagon.

Cases of approximation are cases in which a semantically precise expression is used with slack [Lasersohn 1999]. Importantly, there may not be an absolutely precise convention as to the range of meanings that are compatible with the use of an expression. When is it no longer fine to say *John is 40 years old*? What if John is 35 years old? Approximation is always relative to explicit or implicit standards of precision and to rounding rules, and how close a value needs to be to the literal meaning will often be at the speaker's discretion.

6.1.3 Degree-vagueness

The third aspect of vagueness we isolate concerns the quantitative indeterminacy attached to gradable expressions in particular (which we call degree-vagueness, following Alston 1964; Burks 1946 talks of linear vagueness). Consider the following variation on the dialogue between a father and his son:

(7) Q. How many people were at the party?
A. Many people.

Here again, the answer is imprecise because compatible with a multiplicity of states of affairs obtaining (maybe 25 people were at the party, 50, or 100). Unlike for *some*, however, *many* is not an expression for which we can state determinate truth conditions relative to a fixed countable domain. One way of viewing that phenomenon is as a form of context-dependence [see Sapir 1944, Partee 1989, Lappin 2000, Greer 2014, Egré & Cova 2015]: whereas *some As are Bs* is true exactly if the number of As that are Bs is nonzero,*many As are Bs* would be true if the number of As that are Bs exceeds a context-sensitive number n (cardinal reading), or possibly if the number of As that are Bs exceeds a context-sensitive proportion α of the As or of some other comparison class (proportional reading). The setting of such parameters is problematic: assuming such threshold values, did the son intend *many* to mean *more than 5*, *more than a third*, or some other number? A remarkable fact about vague expressions such as *many* is that the speaker himself or herself need not have a precise idea of the values of such

thresholds in order to apply the expression and to convey meaning.

Beside *many*, paradigmatic examples of vague expressions in that sense include gradable adjectives like *tall*, *long*, *expensive*, *healthy*, etc., all of which accept degree-modification (as in *taller*) or modification by intensifiers (*very tall*) [see Kennedy 2007]. Gradable adjectives give rise to familiar symptoms, in particular the admission of borderline cases of application and the susceptibility to sorites-reasoning [see Keefe 2000, Egré & Klinedinst 2011, Burnett 2016] for a more specific typology of gradable expressions). Borderline cases of application are cases for which it is unclear to the speaker whether the expression should apply or not: for example, it may be unclear whether a man of 178cm should be counted as *tall* or not. An important fact about borderline cases is moreover that they give rise to inconsistent verdicts both between- and within-subjects [see McCloskey & Glucksberg 1978]. Cases of between-subject inconsistencies are often viewed as manifestations of the subjectivity and evaluativity of vague expressions: *many*, *tall*, *healthy*, *beautiful*, could mean different things without error depending on the speaker [see Parikh 1994, Wright 1995, Kölbel 2004, Fara 2000, Raffman 2013, Kennedy 2013, Egré 2016, Verheyen *et al.* 2017]. This subjectivity is important for an assessment of the falsity of vague sentences: the same vague sentence could be used truly relative to one speaker, but be viewed as false by another, depending on their context, interests and evaluative standards (see in particular Kölbel 2004, McNally & Stojanovic 2017 on predicates of personal taste).

6.1.4 Open-texture

The fourth illustration of Russell's definition we single out concerns the openness of the respects constitutive of the meaning of an expression, what we call open-texture (following Waismann 1945; Burks 1946 talks of multidimensional vagueness, and Alston 1964 of combinatorial vagueness). This openness is found at different levels, and it has to do with polysemy and multidimensionality.

Already in the case of dimensional adjectives (like *tall*), the selection of a comparison class is fundamental for the application of the adjective, but it can vary without limit, and it will impact the setting of a boundary between tall and not tall objects (*tall* for a building, for a basketball player, or for a fifth-grader, will mean different things, [see Kamp 1975, Klein 1980].

For a number of gradable adjectives, moreover, several dimensions of comparison interact, and their number and structure is generally indeterminate, even when a comparison class has been fixed. Consider the adjective *healthy*. An indication that *healthy* is multidimensional is the occurrence of adjuncts such as *healthy in some respect*, *healthy in all respects* [Sassoon 2012]. For example, *healthy* as applied to a meal could be predicated based on whether it provides vitamins, or based on whether it has a particular effect on blood pressure, or based on some way of integrating of those respects, and no definitive list of respects appears to be forthcoming.

The phenomenon of open texture is not limited to gradable adjectives, but it concerns the difficulty of providing necessary and sufficient conditions of applications for a vast number of expressions, including nominal expressions (Wittgenstein 1953 famously used the example of the word *game* to show the difficulty in providing a consistent and exhaustive list of defining criteria for that notion).

6.1.5 Representing vagueness

Degree-vagueness and open-texture can be thought of as a forms of "referential multiplicity" [Raffman 2013]. A convenient way of representing the meaning of a vague expression, following the supervaluationist tradition, is thus in terms of a set of admissible sharpenings or precisifications [Mehlberg 1958, Lewis 1979, Fine 1975, Kamp 1975]. For an expression like *tall*, for example, given a comparison class, the meaning can be represented by a set of precise intervals above a variable threshold; for an expression like *healthy*, given a comparison class again, it may be thought of as a set of tuples consisting of variable respects and intervals along a common dimension set by those respects. Similarly for approximation: the meaning of *hundred* as used approximately can be represented by a set of numbers around 100 [Lasersohn 1999]. Depending on the speaker, however, the range of such admissible sharpenings may differ.[2] Different speakers may also assign different weights to different sharpenings depending on the context (see

[2] This makes semantic vagueness close to lexical ambiguity, except that in the case of lexical ambiguity the meanings are supposed to be mentally far apart or disjoint [Keefe 2000, Pinkal 1995]. Logically speaking, however, it is relevant to compare vagueness with ambiguity, since precisifications play the same role as disambiguations [see Lewis 1982]. In the next section, we will see that vagueness, like ambiguity, can give rise to pragmatic equivocation.

Lassiter 2011, Lassiter & Goodman 2015 on probabilistic representations of vague meaning).

In this regard, the main difference between expressions like *hundred* or *some students* on the one hand, and *tall* or *game* on the other, is that the former have determinate truth conditions. Because of that, generality and approximation are cases of *pragmatic vagueness*: by being general rather than more specific a speaker chooses to be less informative than she could be, and by being approximate she gives less information than what the expression literally means. Degree-vagueness and open-texture on the other hand are cases of *semantic vagueness*: the meaning of expressions like *many*, *healthy* or *game* is "intrinsically uncertain" (in the words of Peirce 1902, 748), that is those expressions do not have constant truth conditions across contexts and speakers.

With these distinctions in mind, we are now in a position to examine the ways in which vagueness interacts with the Gricean maxims. The Gricean maxims assume that conversation fundamentally rests on cooperation. As we know from game theory, however, speaker and hearer need not have their interests perfectly aligned, and sometimes they can diverge dramatically. It may be costly to reveal the truth, or to reveal the *whole* truth. Most of the time, however, making an assertion that the listener would recognize as false can be even more costly: if a false claim is exposed, the speaker incurs the risk of losing credibility, or greater costs [see Asher & Lascarides 2013]. In the rest of this chapter, we distinguish two main classes of situations that motivate the use of vague language. On the one hand, there are situations where the speaker is *imperfectly informed* about the facts, and may simply wish to avoid speaking falsely by speaking too precisely. On the other hand, there are situations where the speaker is *perfectly informed* about the facts, but has an interest to hide information from the hearer, and potentially to take advantage of the indeterminacy of vague expressions to bias or mislead.

6.2 Avoiding error

Grice's Maxim of Quality enjoins one not to speak falsely, but also not to say things for which one lacks adequate evidence. One aspect in which the Maxim of Quality justifies the use of vague language concerns cases where the speaker is

uncertain about which precise state of affairs obtains [see Channell 1994, Frazee & Beaver 2010] or will obtain in the future [Channell 1985, van Deemter 2009].

Consider a situation in which you return from a party and are a fully cooperative speaker trying to convey maximum information. The party was attended by a group of people, but you do not know exactly how many there were, because you could not count them. Upon returning from the party, you're asked how many people were there. In this case, there is no number *n* for which you can truly and justifiably say: *there were exactly n people*. In order to respond truly and justifiably, the next option would be to specify an exact interval. Suppose you are sure that there were more than 20 people, and fewer than 200 hundreds, but are uncertain in between. Then you may say:

(8) There were between 20 and 200 people.

The response is general in this case, but little informative. It would be more informative to give your best estimate of a lower bound:

(9) At least 100 people.

But suppose there were in fact 93 persons attending. The answer would be literally false, despite coming close to your assessment. On the other hand, you would not be wrong if you said:

(10) a. About 100 people.
b. Many people.

Semantic expressions like *about* and *many* allow you to convey information truly in this case, compatibly with an indeterminate range of states of affairs obtaining. They allow you to avoid error, but also, somewhat surprisingly, to be more informative than you would if you tried to specify exact intervals without error.

Importantly, the hearer may have a different understanding of what to count as *many* than you. Suppose you understand *many* to denote a range of sharp intervals (using the supervaluationist picture), with a probability distribution on them (some precisifications are more likely to you than other; [see Lassiter 2011, Lassiter & Goodman 2015]. The hearer may have a different probability distribution

that rules out some of the intervals you consider possible, but you would still communicate successfully if the hearer ends up with a posterior distribution that includes the value you actually observed, and if it makes that value more likely than before you answered [see Parikh 1994, Lassiter 2011, Lassiter & Goodman 2015].

The point of the previous example is that vague language, in situations of uncertainty, may accomplish an optimal tradeoff between the need to be truthful and the need to be informative frazee2010vagueness. Use of vague language in situations of uncertainty is also modulated by the cost of speaking falsely, compared to the benefits of speaking accurately. An example discussed by van Deemter [2009] concerns cases of medical communication. Van Deemter points out that "a doctor who says "These symptoms will disappear fairly soon" is less likely to get complaints, and to be sued, than one who says "These symptoms will have disappeared by midnight"" [8].

Vague language, in summary, is a way of speaking truly and informatively in situations of uncertainty. This does not mean that vagueness is immune to falsity: suppose the symptoms disappear only after a month, then the patient may charge the doctor of incompetence, or even of having lied. The patient could complain that *fairly soon* was, in her perspective, incompatible with a time interval of a month. The doctor could deny having spoken falsely, on the other hand, by defending her own perspective. The relativity of vague interpretations to speakers makes charges of lies, as we will see, a delicate matter (see section 6.5).

6.3 Hiding Information

Let us now turn to cases where the speaker has no uncertainty about the world, but has an incentive to be noncooperative. Grice's Maxim of Quantity is twofold: it asks one to be as informative as required for the purpose of the conversation, but also to not be more informative than required. What counts as "required for the purpose of the conversation" is itself vague and heavily depends on the interests that speaker and hearer have in sharing information [Asher & Lascarides 2013]. For a range of situations, a well-informed speaker can legitimately wish to retain

information from the hearer, and so to be vague in order to limit cooperation. Cases of what we called *generality* in the previous section are very common in that regard. Consider the dialogue in (2), repeated here.

(11) Q. Who did you see at the party?
A. Some friends.

Let us assume that the father is actually interested in knowing whether his son saw a particular person, say Ann, whom he suspects his son of dating. The son, on the other hand, wishes to keep his privacy. Assume the son saw Ann indeed, but also Don and Eli, two other friends known to the father. In this case, the son is giving a perfectly true answer, but he is not allowing the father to identify whom exactly he saw.

Compare with the example of the previous section. Assume you know this time that exactly 63 people attended the party, but have an interest not to reveal the exact number. You may choose to be underinformative by responding:

(12) Q. How many people were at the party?
A. Fewer than a hundred.

The answer is literally true, but *partial* in the sense of Groenendijk & Stokhof [1982]: it leaves possibilities open and fails to completely settle the question. Potentially, it is also misleading: for it triggers the implicature that it is compatible with your knowledge that there could have been 90 people or more attending [see Spector 2013]. Such cases, in which a speaker is literally truthful but uses misleading implicatures are called cases of misdirection by Asher & Lascarides [2013], who characterize them as instances of *rhetorical* as opposed to genuine Gricean cooperativity.[3]

Neither of the previous examples relies on utterances that are vague semantically, but we can find similar cases where a semantically vague expression is used to withhold information. Imagine nosey neighbours asking how much you paid

[3] See in particular their discussion of *Bronston vs. United States* as an exploitation of literal truth to refute perjury, as well as the presentation of the case in Tiersma [2004]. See Ransom *et al.* [2017] for a recent study comparing cases in which a truthful speaker may have an incentive to be completely uninformative to where they may choose to be partially informative depending on the level of trust in the hearer.

for your apartment. Assume you know the exact price you paid, but don't want to reveal it:

(13) Q. How much did you buy your apartment?
A. It was not too expensive.

An incentive to avoid being precise in this case is that you may want to avoid appearing either lucky (in case you paid less than your neighbors for the same size) or stupid (in case you paid more), or you may just want to give no indication of your assets. Use of a qualitative expression like *expensive* is advantageous here because it avoids specifying a definite number, and it remains compatible with the preservation of truthfulness: we may assume that you are sincere in thinking that the price you paid was not expensive, even ahead of the dialogue (that assumption is not always warranted, see the next section).

Consider for comparison the following alternative answers, assuming the exact price you paid for your apartment is 220,000 euros:

(14) a. I paid 200,000 euros.
b. I paid around 200,000 euros.
c. I paid between 50,000 euros and 300,000 euros.

Answer (14a) is approximate in this case, but it does not signal that it is approximate. As pointed out by Meibauer [2014], it may be truthfully asserted if the standard of precision in the context of the conversation is such that a difference of 20,000 euros would not be relevant. But the answer could be misleading, instead of just imprecise, if uttered with the intention of making your neighbors believe that you paid less than you actually did. For instance it would count as false in a context in which the standard of precision needs to be maximal (say in the context of declaring taxes).

Answer (14b) makes the approximation explicit, and it is also semantically vague, due to the use of the vague modifier *around*. Despite that, the answer remains more informative than the one in (13), for it lets your neighbors infer the actual price with less uncertainty than based on hearing *not too expensive*.

Answer (14c), finally, is neither approximate nor semantically vague: it states

an exact interval but to create uncertainty. Like (13), it signals either that you do not know the price you paid, or that you don't want to answer the question precisely; however, the interval specified is so large here that the hearers would be better-founded to think you do not want to answer the question. Also, the answer in (13) may end up being more informative than the one in (14c) despite relying on semantic vagueness, because upon hearing "not too expensive" the hearer is likely to narrow down the range of prices you potentially paid to a smaller interval than the one specified in (14c).[4]

6.4 Making half-truths

Beside cases in which a speaker is imprecise to hide information, there is a class of cases where the speaker can exploit the semantic indeterminacy of vague expressions to produce utterances whose truth status in unclear: they are true under some way of resolving their vagueness, but that way can be tendentious or biased.[5]

Consider the following example (from C. List and L. Valentini, p.c.) where you receive an invitation for dinner. As a matter of fact, you would be free to go to that dinner, but have no inclination for it. Imagine the following dialogue:

(15) Q. Are you free to come for supper tomorrow?
A. Sorry, I have an engagement.

(16) Q. Are you free to come for supper tomorrow?
A. Sorry, I am busy.

In (15), your response ought to qualify as a lie. In the case of (16), the answer does not obviously count as a lie, but it does not clearly count as true either. One way of explaining the contrast is in terms of supervaluations [Fine 1975, Kamp 1975]. On all admissible ways of sharpening the meaning of *I have an engagement*, the sentence

[4] This is because *I paid between 50,000 and 300,000 euros* scalarly implicates that it is possible you paid 51,000. With *not too expensive* this inference is not mandated at all. On the mechanism of such implicatures, see Fox [2014].

[5] The term *half-truth* is used in a number of different senses in the literature. Our use is broadly compatible with Carson [2010]'s, who defines a half-truth to be a true statement that "selectively emphasize[s] facts that tend to support a particular interpretation or assessment of an issue" [57-58]. We use *half-true* in the sense of *borderline true*.

would come out false (i.e. super-false). On the other hand, there are admissible ways of sharpening the meaning of *busy* for the sentence to count as true. *I am busy* may even be deemed super-true, that is true literally on all admissible ways of sharpening the meaning of *busy*, but this is moot: it depends on what to count as an admissible precisification (see below). If you end up watching TV, you would obviously be *busy watching TV*, but at the time of utterance *busy* appears to convey that you have some obligation.

In our view the answer in (16) is a half-truth, precisely because it is not clearly false, but not clearly true either. Concretely, *I am busy* offers a polite a way of declining the invitation. A more informative alternative about the speaker's motives would be to say: *I am not very inclined*, but it would be clearly offending. The intent of *I am busy* is partly to mislead, therefore, but consistently with satisfying a norm of politeness.[6]

A more extreme case of exploitation of semantic vagueness concerns President Bill Clinton's declarations about the nature of his relationship with Monica Lewinsky:

(17) I have never had sexual relations with Monica Lewinsky.

This case, importantly, is one where all parties had been fully informed of the relevant facts. To justify his claim without perjury, Bill Clinton took advantage of the open texture of the expression *sexual relations*, that is of the lack of a clear definition. However, he did it not by making up a definition, but by exploiting an attempt made by his opponents to provide an explicit definition of the term "[engaging in] sexual relations" [see Tiersma 2004, for details]. [7] Pressed to explain himself, Clinton's defense was:

(18) "I thought the definition [of sexual relations, as read by Judge Wright] included any activity by the person being deposed, where the person was the actor and came in contact with those parts of the bodies with the purpose or intent of gratification, and excluded any other activity".

[6] Thanks to C. List and L. Valentini for discussion of that aspect.

[7] The explicit definition in question is: "a person engages in "sexual relations" when the person knowingly engages in or causes contact with the genitalia, anus, groin, breast, inner thigh, or buttocks of any person with an intent to gratify or arouse the sexual desire of any person'Contact" means intentional touching, either directly or through clothing."

The way Bill Clinton defended himself can be put in supervaluationist terms again: it is not the case that on all ways of further precisifying the explicit definition proposed by his opponents, receiving oral sex counts as engaging in a sexual relation. Interestingly, in an earlier statement Bill Clinton commented about whether Monica Lewinsky had had "a sexual affair" with him as follows:

(19) Q. If she told someone that she had a sexual affair with you beginning in November of 1995, would that be a lie?
A. It's certainly not the truth. It would not be the truth.

In this occurrence, Clinton appeared to concede that the allegation would not necessarily be false, but without counting as true. In supervaluationist terms again, there are some admissible ways of precisifying *sexual affair* that would make Lewinsky's supposed statement true, yet *not all* ways of precisifying *sexual affair* would make it true. Overall, Bill Clinton was able to exploit the semantic indeterminacy of those expressions in order to avoid the charge of perjury. He would have been convicted if, from the jury's perspective, all admissible ways of precisifying the meaning had led to the sentence being false, but the jury in that case failed to rule out Clinton's way from being admissible.

6.5 Are half-truths lies?

Let us take stock. In section 6.3 we saw that in response to a question, a speaker can be underspecific without committing any lie. In section 6.4, however, we saw that semantic indeterminacy can be used to produce sentences whose truth status is unclear, what we called half-truths. Shouldn't half-truths be considered lies, however, given that those utterances fail to be clearly true?

First of all, utterances like (16) or (17) may typically be uttered insincerely. In the case of (16), I may think to myself *in petto* "well, I am not really busy..." or "well, I am busy watching TV", and Clinton may have silently thought to himself "well, except for an oral sexual relation". Those utterances then may be viewed as cases of *amphiboly* or *mental reservation* [Bok 1979, Mullaney 1980, Adler 1997], whereby the actual meaning that the speaker has in mind is in fact different from the meaning the hearer can reasonably infer.

To avoid that complication, let us assume that each utterance is made sincerely at the time it is uttered, and without mental reservation (without the speaker making any silent addition). In supervaluationist terms, the question we are asking is whether an utterance that fails to be super-true (true on all admissible precisifications) ought to be considered false on normative grounds. We think the answer to this question is nonobvious, for it depends on two parameters: the definition of what to count as an admissible precisification, and the choice of a standard for truth.

Regarding the first issue, most people would agree that Clinton's utterance is false *simpliciter*, despite being true under some very specific sharpening of the meaning of *sexual relation*, for they would deem that particular precisification to be inadmissible in an ordinary conversational context. In the legal context, however, Clinton was successful in making that sharpening relevant, and since it was incumbent on the jury to show that his statement was unequivocally false, it allowed for his sentence not to qualify as a lie, despite the sentence not qualifying as a clear truth either.

This brings us to the second issue. Theories of vagueness differ on the standards whereby a sentence can be truthfully uttered. Supervaluationism treats sentences as true *simpliciter* if they are true on all admissible precisifications, but there is a dual theory, subvaluationism, which treats sentences as true *simpliciter* when true under some precisification [Hyde 1997]. Subvaluationism is very liberal in that it predicts that a sentence and its negation can both be true then.[8]

In practice, the standards for truth and falsity appear to depend on the context. In the Clinton lawsuit, it was sufficient for the sentence to be true under some sharpening to not be considered a lie by the jury. In the class-action lawsuit that opposed Athena Hohenberg to the Ferrero company, on the other hand, the complaint was that *healthy* was used misleadingly for a product containing too much fat and sugar. Ferrero's defense was based on the fact that *healthy* is multidimensional, and that their product was at least healthy in the respects of bringing chocolate, containing low sodium, and so on.[9]. Despite that, the court

[8] This implies that *I am busy* and *I am not busy* would both be true in a context in which either is true under some admissible sharpening. But each of them would also be false, since false under some sharpening. The upshot would be that the sentence both is a lie, and fails to be a lie.

[9] See http://www.scpr.org/news/2011/02/10/23912/a-mom-sues-nutella-maker-for-deceptive-advertising/

eventually forbade Ferrero from advertising the product as *healthy*. The court agreed that it is not enough for a sentence like *this product is healthy* to be true on just some ways of precisifying *healthy* in order for the sentence to avoid being misleading or to count as a lie, presumably in this case because the ways in which the sentence is false outweigh those in which it is true (Ferrero's use would in fact violate the Gricean Maxim of Relevance).

In general, however, the Ferrero example may be more emblematic of the ways in which vague language is interpreted. Grice's Maxim of Manner recommends avoiding ambiguity [see Grice 1975, 46]. There is evidence, however, that in cases in which a vague predicate is used without qualification, and where two interpretations are available for the predicate, a weak one and a strong one, the stronger interpretation will be the default [see Dalrymple *et al.* 1998]. Upon hearing *this person is tall*, the default is to get that the person is clearly tall, rather than borderline tall [Alxatib & Pelletier 2011, Cobreros *et al.* 2012 2015]. Likewise, when saying *this product is healthy*, the default is likely to hear *this product is healthy in most respects*, or even in *all respects* [see Sassoon 2012], rather than just *some respects*. As a result, to say of a product that it is *healthy* without qualification would suggest that the product is more healthy than unhealthy: in the Ferrero case, this pragmatic enrichment is deceptive, and can legitimately be considered a lie.

We see, in summary, that often an utterance will be deemed a lie if it fails to be unambiguously true. But sometimes, as the Clinton case shows us, it might fail to be deemed a lie if it is not unambiguously false. Whichever of those two will prevail appears to depend not just on the *existence* of ways for a sentence to be true, but also on how *relevant* those ways are to the parties involved in the conversation.

6.6 Conclusion

Let us recapitulate the main lessons of our discussion of the relation between lying and vagueness. To begin with, we have seen that vagueness provides a way for a cooperative speaker to remain truthful in situations in which she is trying to communicate information about which she is uncertain. Vagueness may

then be described as a way of avoiding error and therefore lies. This concerns all cases in which the use of qualitative but vague vocabulary (as in *many*, *long*, *expensive*) avoids committing oneself to precise quantitative expressions for which one fails to have adequate evidence. As opposed to that, we have highlighted two kinds of cases in which vagueness can be used deceptively. The first are cases in which a well-informed speaker has motives to hide or retain information. In such cases the speaker is deliberately imprecise and partial, but need not commit lies in the strict sense of the term. She may however be misleading if the partial information given triggers false implicatures. The second kind of cases concern what we have called half-truths, utterances whose status is borderline between true and false, depending on how vague expressions in them are interpreted. Such cases are more problematic. An utterance will be misleading if it is true only under some very peculiar precisification. On the other hand, the indeterminacy of vague expressions can make it difficult to prove that a vague utterance is a lie, as opposed to an expression whose intended meaning was misunderstood.

Acknowledgments: Thanks to Sam Alxatib, Nick Asher, Hans van Ditmarsch, Rae Langton, Christian List, Neri Marsali, Yael Sharvit, Stephanie Solt, Benjamin Spector, Laura Valentini, Steven Verheyen, for helpful conversations on the topics of this paper, and to Jörg Meibauer for helpful comments and for his editorial assistance. We also thank Emar Maier for valuable advice regarding how to convert our paper from Latex to MSWord, as well as the Lorentz Center and the organizers of the workshop "The Invention of Lying: Language, Logic, and Cognition", held in Leiden in January 2017, where Jörg Meibauer invited us to contribute this chapter. Paul Égré thanks the ANR Program TriLogMean ANR-14-CE30-0010-01 for funding. Benjamin Icard thanks the Direction Générale de l'Armement for doctoral funding. Both researchers acknowledge grants ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL* for research carried out at the Department of Cognitive Studies of ENS.

Bibliography

Abe, Nobuhito. 2009. The neurobiology of deception: evidence from neuroimaging and loss-of-function studies. *Current opinion in neurology*, **22**(6), 594–600.

Adler, Jonathan E. 1997. Lying, deceiving, or falsely implicating. *The Journal of Philosophy*, **94**(9), 435–452.

Agotnes, Thomas, van Ditmarsch, Hans P., & Wang, Yanjing. 2016. True lies. *Synthese*, 1–35.

Akdag, Herman, de Glas, Michel, & Pacholczyk, Daniel. 1992. A qualitative theory of uncertainty. *Fundamenta Informaticae*, **17**(4), 333–362.

Alston, William P. 1964. *Philosophy of Language*. Prentice Hall.

Alxatib, Sam, & Pelletier, Francis Jeffry. 2011. The psychology of vagueness: Borderline cases and contradictions. *Mind & Language*, **26**(3), 287–326.

Aragon, Louis. 1980. *Le mentir-vrai*. Editions Gallimard.

Arico, Adam J., & Fallis, Don. 2013. Lies, damned lies, and statistics: An empirical investigation of the concept of lying. *Philosophical Psychology*, **26**(6), 790–816.

Asher, Nicholas, & Lascarides, Alex. 2013. Strategic conversation. *Semantics and Pragmatics*, **6(2)**, 1–62.

Aucher, Guillaume. 2004. A Combined System for Update Logic and Belief Revision. *Pages 1–17 of:* Barley, Mike, & Kasabov, Nikola (eds), *PRIMA*, vol. 3371. Springer. Revised Selected Papers.

Aucher, Guillaume. 2008. *Perspectives on Belief and Change*. Ph.D. thesis, University of Otago, New Zealand Institut de Recherche en Informatique de Toulouse.

Augustine, Saint. 395. On lying. *St. Augustine: Treatises on Various Subjects. New York: The Fathers of the Church*, 47–109.

Austin, John L. 1962. *Sense and Sensibilia*. Oxford University Press.

Bach, Kent. 1994. Semantic slack: What is said and more. *Foundations of speech act theory: Philosophical and linguistic perspectives*, 267–291.

Bach, Kent. 2008. Applying pragmatics to epistemology. *Philosophical issues*, **18**(1), 68–88.

Baker, James D., & Mace, Douglas J. 1973. Certitude Judgments Revisited. *Unpublished manuscript, US Army Research Institute for the Behavior and Social Sciences*.

Baker, James D., McKendry, James M., & Mace, Douglas J. 1968. Certitude Judgments in an Operational Environment. *US Army Technical Research Note*, **200**.

Baltag, Alexandru, & Smets, Sonja. 2006. Dynamic belief revision over multi-agent plausibility models. *Pages 11–24 of: Proceedings of LOFT*, vol. 6.

Baltag, Alexandru, & Smets, Sonja. 2008a. The Logic of Conditional Doxastic Actions. *Pages 9–31 of:* Apt, Krzysztof R., & van Rooij, Robert (eds), *New Perspectives on Games and Interaction, Texts in Logic and Games*, vol. 4. Amsterdam University Press.

Baltag, Alexandru, & Smets, Sonja. 2008b. A qualitative theory of dynamic interactive belief revision. *Logic and the foundations of game and decision theory (LOFT 7)*, **3**, 9–58.

Baltag, Alexandru, & Smets, Sonja. 2011. Keep changing your beliefs, aiming for the truth. *Erkenntnis*, **75**(2), 255.

Baltag, Alexandru, Moss, Lawrence S., & Solecki, Slawomir. 1998. The logic of public announcements, common knowledge, and private suspicions. *Proceeding TARK '98 Proceedings of the 7th conference on Theoretical aspects of rationality and knowledge*.

Baltag, Alexandru, Rodenhäuser, Ben, & Smets, Sonja. 2012. Doxastic attitudes as belief-revision policies. *In: Proceedings of the ESSLLI workshop on strategies for learning, belief revision and preference change*.

Barbieri, Davide. 2013. Bayesian Intelligence Analysis. *9th Conference on Intelligence in the Knowledge Society, Bucharest*.

Barnes, Alan. 2016. Making Intelligence Analysis More Intelligent: Using Numeric Probabilities. *Intelligence and National Security*, **31 (3)**, 327–344.

Barnes, Annette. 2007. *Seeing through self-deception*. Cambridge University Press.

Barnes, John A. 1994. *A pack of lies: Towards a sociology of lying*. Cambridge University Press.

Batini, Carlo, & Scannapieco, Monica. 2006. *Data Quality: Concepts, Methodologies and Techniques*. Data-Centric Systems and Applications. Springer.

Bayes, Thomas. 1763. An Essay Toward Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society*, **53**, 370–418.

van Benthem, Johan. 2004. What One May Come to Know. *Analysis*, **64**(282), 95–105.

van Benthem, Johan. 2007. Dynamic logic for belief revision. *Journal of applied non-classical logics*, **17**(2), 129–155.

van Benthem, Johan. 2014. Two logical faces of belief revision. *Pages 281–300 of: Krister Segerberg on Logic of Actions*. Springer.

van Benthem, Johan, & Liu, Fenrong. 2004. Diversity of logical agents in games. *Philosophia Scientiae*, **8**(2), 163–178.

van Benthem, Johan, & Liu, Fenrong. 2007. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logics*, **17**(2), 157–182.

van Benthem, Johan, & Smets, Sonja. 2015. Dynamic Logics of Belief Change. *Pages 299–368 of: Handbook of Logics for Knowledge and Belief*. College Publications.

Blasch, Erik, Laskey, Kathryn B., Jousselme, Anne-Laure, Dragos, Valentina, Costa, Paulo C., & Dezert, Jean. 2013. URREF reliability versus credibility in information fusion (STANAG 2511). *In: Advances and Applications of DSmT for Information Fusion*, vol. 4.

Bok, Sissela. 1979. *Lying: Moral Choice in Public and Private Life*. Vintage paperback editions.

Bok, Sissela. 1983. *Secrets: On the ethics of concealment and revelation*. New York: Random House.

Bonnay, Denis, & Égré, Paul. 2011. Knowing One's Limits: An Analysis in Centered Dynamic Epistemic Logic. *Pages 103–126 of:* Girard, Patrick, Roy, Olivier, & Marion, Mathieu (eds), *Dynamic Formal Epistemology*. Springer.

Bonnefon, Jean-François, & Villejoubert, Gaëlle. 2006. Tactful or Doubtful? Expectations of Politeness Explain the Severity Bias in the Interpretation of Probability Phrases. *Psychological Science*, **17**(9), 747–751.

Boolos, George. 1996. The Hardest Logic Puzzle Ever. *The Harvard Review of Philosophy*, **6**(1), 62–65.

Bryant, Geoffrey D., & Norman, Geoffrey R. 1980. Expressions of probability: words and numbers. *The New England Journal of Medicine*, **302**(7), 411.

Budescu, David V, & Wallsten, Thomas S. 1985. Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes*, **36**(3), 391–405.

Burks, Arthur W. 1946. Empiricism and vagueness. *The Journal of Philosophy*, **43(18)**, 477–486.

Burnett, Heather. 2016. *Gradability in Natural Language: Logical and Grammatical Foundations*. Oxford University Press.

Bussey, Kay. 1999. Children's categorization and evaluation of different types of lies and truths. *Child Development*, **70**(6), 1338–1347.

Caminada, Martin. 2009. Lies and Bullshit: distinguishing classes of dishonesty. *In: Social Simulation Workshop at the International Joint Conference on Artificial Intelligence*.

Capet, Philippe. 2006. *Logique du mensonge*. Ph.D. thesis, Université Paris III Sorbonne Nouvelle.

Capet, Philippe, & Delavallade, Thomas. 2013. *L'évaluation de l'information: confiance et défiance*. Hermès.

Capet, Philippe, & Revault d'Allonnes, Adrien. 2013. La cotation dans le domaine militaire: doctrines, pratiques et insuffisances. *In:* Capet, Philippe, & Delavallade, Thomas (eds), *L'évaluation de l'information: confiance et défiance*. Hermès.

Carson, Thomas L. 2006. The definition of lying. *Noûs*, **40**(2), 284.

Carson, Thomas L. 2010. *Lying and deception: Theory and practice*. Oxford University Press.

Channell, Joanna. 1985. Vagueness as a conversational strategy. *Nottingham Linguistic Circular*, **14**, 3–24.

Channell, Joanna. 1994. *Vague Language*. Oxford University Press.

Chisholm, Roderick M., & Feehan, Thomas D. 1977. The Intent to Deceive. *Journal of Philosophy*, **74**(3), 143–159.

Cholvy, Laurence. 2004. Information evaluation in fusion: a case study. *Pages 993–1000 of: Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*.

Cholvy, Laurence. 2010. Evaluation of information reported: a model in the theory of evidence. *Pages 258–267 of: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer.

Cholvy, Laurence. 2013. Lorsque l'information est de seconde main. *In:* Capet, Philippe, & Delavallade, Thomas (eds), *L'évaluation de l'information: confiance et défiance*. Lavoisier.

Christ, Shawn E., van Essen, David C., Watson, Jason M., Brubaker, Lindsay E., & McDermott, Kathleen B. 2008. The contributions of prefrontal cortex and executive control to deception: evidence from activation likelihood estimate meta-analyses. *Cerebral cortex*, **19**(7), 1557–1566.

Cobreros, Pablo, Egré, Paul, Ripley, David, & van Rooij, Robert. 2012. Tolerant, Classical, Strict. *The Journal of Philosophical Logic*, **41**(2), 347–385.

Cobreros, Pablo, Egré, Paul, Ripley, David, & van Rooij, Robert. 2015. Pragmatic interpretations of vague expressions: strongest meaning and nonmonotonic consequence. *Journal of Philosophical Logic*, **44**(4), 375–393.

Coleman, Linda, & Kay, Paul. 1981. Prototype Semantics: The English Word Lie. *Language*, **57**(1), 26–44.

Dalrymple, Mary, Kanazawa, Makoto, Kim, Yookyung, Mchombo, Sam, & Peters, Stanley. 1998. Reciprocal expressions and the concept of reciprocity. *Linguistics and Philosophy*, **21**(2), 159–210.

Danziger, Eve. 2010. On trying and lying: Cultural configurations of Grice's Maxim of Quality. *Intercultural Pragmatics*, **7**(2), 199–219.

Davidson, Donald. 1985. Deception and division. *The multiple self*, **79**.

Debey, Evelyne, Verschuere, Bruno, & Crombez, Geert. 2012. Lying and executive control: An experimental investigation using ego depletion and goal neglect. *Acta Psychologica*, **140**(2), 133–141.

van Deemter, Kees. 2009. Utility and language generation: The case of vagueness. *Journal of Philosophical Logic*, **38**(6), 607–632.

Demolombe, Robert. 2004. Reasoning about trust: A formal logical framework. *Pages 291–303 of: International Conference on Trust Management*. Springer.

Demolombe, Robert, & Lorini, Emiliano. 2008. A logical account of trust in information sources. *In: Proceedings of the 11th International Workshop on Trust in Agent Societies*.

Demos, Raphael. 1960. Lying to oneself. *The Journal of Philosophy*, **57**(18), 588–595.

Dempster, Arthur P. 1967. Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics*, 325–339.

DeRose, Keith. 2002. Assertion, knowledge, and context. *The Philosophical Review*, **111**(2), 167–203.

DIA-2. 2010. Doctrine interarmées. *CICDE, Renseignement d'intérêt militaire et contre-ingérence (RIM & CI)*, **2**.

DIS. 2001. Intelligence Wing Student Précis. *Defence Intelligence, Security Center*.

van Ditmarsch, Hans P. 2005. Prolegomena to dynamic logic for belief revision. *Synthese*, **147**(2), 229–275.

van Ditmarsch, Hans P. 2008. Comments on 'The Logic of Conditional Doxastic Actions'. *Pages 33–34 of:* Apt, Krzysztof R., & van Rooij, Robert (eds), *New Perspectives on Games and Interaction, Texts in Logic and Games*, vol. 4. Amsterdam University Press.

van Ditmarsch, Hans P. 2014. Dynamics of lying. *Synthese*, **191**(5), 745–777.

van Ditmarsch, Hans P., & Kooi, Barteld. 2006. The secret of my success. *Synthese*, **153**(2), 339–339.

van Ditmarsch, Hans P., & Labuschagne, Willem A. 2007. My beliefs about your beliefs: a case study in theory of mind and epistemic logic. *Synthese*, **155**(2), 191–209.

van Ditmarsch, Hans P., van Eijck, Jan, Sietsma, Floor, & Wang, Yanjing. 2012. On the Logic of Lying. *Pages 41–72 of:* van Eijck, Jan, & Verbrugge, Rineke (eds), *Games, Actions and Social Software: Multidisciplinary Aspects*. Springer Berlin Heidelberg.

Douven, Igor. 2006. Assertion, knowledge, and rational credibility. *The Philosophical Review*, **115**(4), 449–485.

Douven, Igor. 2009. Assertion, moore, and bayes. *Philosophical Studies*, **144**(3), 361–375.

Dretske, Fred I. 1981. *Knowledge and the Flow of Information*. MIT Press.

Dretske, Fred I. 1983. Précis of Knowledge and the Flow of Information. *Behavioral and Brain Sciences*, **6**(1), 55–90.

Dubois, Didier, & Prade, Henri. 1990. An introduction to possibilistic and fuzzy logics. *Pages 742–761 of: Readings in uncertain reasoning*. Morgan Kaufmann Publishers Inc.

Dynel, Marta. 2011. A web of deceit: A neo-Gricean view on types of verbal deception. *International Review of Pragmatics*, **3**(2), 139–167.

Dynel, Marta. 2015. Intention to deceive, bald-faced lies, and deceptive implicature: Insights into Lying at the semantics-pragmatics interface. *Intercultural Pragmatics*, **12**(3), 309–332.

Dynel, Marta. 2016. Comparing and combining covert and overt untruthfulness. *Pragmatics & Cognition*, **23**(1), 174–208.

Egré, Paul. 2014. Intentional action and the semantics of gradable expressions: (on the Knobe effect). *Pages 176–205 of:* Copley, Bridget, & Martin, Fabienne (eds), *Causation in Grammatical Structures*. Oxford University Press.

Egré, Paul. 2016. Vague judgment: a probabilistic account. *Synthese*. DOI: https://doi.org/10.1007/s11229-016-1092-2.

Egré, Paul, & Cova, Florian. 2015. Moral asymmetries and the semantics of many. *Semantics and Pragmatics*, **8 (13)**, 1–45.

Egré, Paul, & Klinedinst, Nathan. 2011. *Vagueness and Language Use*. Palgrave Macmillan.

Égré, Paul, & Icard, Benjamin. 2018. Lying and Vagueness. *In:* Meibauer, Jörg (ed), *Oxford Handbook of Lying*. Oxford University Press.

Ekman, Paul. 1985. *Telling lies: Clues to deceit in the marketplace, marriage, and politics*. New York: Norton.

Engel, Pascal. 2008. In what sense is knowledge the norm of assertion? *Grazer Philosophische Studien*, **77**(1), 45–59.

Engel, Pascal. 2016. Demi-vérités et demi-mensonges, de Pinocchio à Polichinelle. *In:* Wieviorka, Michel (ed), *Mensonges et vérités*. Sciences Humaines Éditions, Auxerre.

Fallis, Don. 2009a. A Conceptual Analysis of Disinformation. *iConference 2009 Papers*.

Fallis, Don. 2009b. What is lying? *The Journal of Philosophy*, **106**(1), 29–56.

Fallis, Don. 2010. Lying and deception. *Unpublished*.

Fallis, Don. 2011. Floridi on Disinformation. *Etica and Politica / Ethics and Politics*, 201–214.

Fallis, Don. 2014. The Varieties of Disinformation. *Pages 135–161 of: The Philosophy of Information Quality*. Springer.

Fallis, Don. 2015. What is disinformation? *Library Trends*, **63**(3), 401–426.

Fallis, Don. 2018. Lying and Omissions. *In:* Meibauer, Jörg (ed), *Oxford Handbook of Lying*. Oxford University Press.

Fara, Delia. 2000. Shifting Sands: an Interest-Relative Theory of Vagueness. *Philosophical Topics*, **28(1)**, 45–81. Originally published under the name "Delia Graff".

Faulkner, Paul. 2007. What is wrong with lying? *Philosophy and Phenomenological Research*, **75**(3), 535–557.

Fetzer, James H. 2004. Information: Does it have to be true? *Minds and Machines*, **14**(2), 223–229.

Fillmore, Charles J. 1975. An alternative to checklist theories of meaning. *Pages 123–131 of: Annual Meeting of the Berkeley Linguistics Society*, vol. 1.

Fine, Kit. 1975. Vagueness, Truth, and Logic. *Synthese*, **30**, 265–300.

Fisk, Charles E. 1972. The Sino-Soviet border dispute: A comparison of the conventional and Bayesian methods for intelligence warning. *Studies in intelligence*, **16**(2), 53–62.

Floridi, Luciano. 1996. Brave.Net.World: the Internet as a disinformation superhighway? *The Electronic Library*, **14(6)**, 509–514.

Floridi, Luciano. 2007. In defence of the Veridical Nature of Semantic Information. *European Journal of Analytic Philosophy*, **3**(1).

Floridi, Luciano. 2008. *The Blackwell guide to the philosophy of computing and information*. John Wiley & Sons.

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

Floridi, Luciano, & Illari, Phyllis. 2014. *The philosophy of information quality*. Vol. 358. Springer.

FM-2-22.3. 2003. *Human Intelligence: collector operations*. Department of the United States Army.

FM-30-5. 1971. *Combat Intelligence*. Department of the United States Army.

Fox, Danny. 2014. Cancelling the Maxim of Quantity: Another challenge for a Gricean theory of scalar implicatures. *Semantics and Pragmatics*, **7(5)**, 1–20.

Frankfurt, Harry G. 1999. *Necessity, volition, and love*. Cambridge University Press.

Frazee, Joey, & Beaver, David. 2010. Vagueness Is Rational under Uncertainty. *In:* Aloni, M., Bastiaanse, H., de Jager, T., & Schulz, K. (eds), *Logic, Language and Meaning. Lecture Notes in Computer Science*. Springer.

Freud, Sigmund. 1960. Jokes and their Relation to the Unconscious (J. Strachey, Trans.). *New York: Holt, Rinehart, and Winston*.

Fried, Charles. 1978. *Right and Wrong*. Vol. 153. Harvard University Press.

Gamer, Matthias. 2011. Detecting of deception and concealed information using neuroimaging techniques. *Page 90–113 of:* Verschuere, Bruno, Ben-Shakhar, Gershon, & Meijer, Ewout (eds), *Memory Detection: Theory and Application of the Concealed Information Test*. Cambridge University Press.

Ganis, Giorgio, & Keenan, Julian P. 2009. The cognitive neuroscience of deception. *Social Neuroscience*, **4**(6), 465–472.

Gerbrandy, Jelle. 1999. *Bisimulations on planet Kripke*. ILLC Dissertation Series.

Gerbrandy, Jelle. 2007. The surprise examination in dynamic epistemic logic. *Synthese*, **155**(1), 21–33.

Gerbrandy, Jelle, & Groeneveld, Willem. 1997. Reasoning about information change. *Journal of logic, language and information*, **6**(2), 147–169.

Greer, Kristen A. 2014. Extensionality in natural language quantification: the case of many and few. *Linguistics and Philosophy*, **37(4)**, 315–351.

Grice, Paul H. 1957. Meaning. *Philosophical Review*, **66**(3), 377–388.

Grice, Paul H. 1975. Logic and Conversation. *Pages 41–58 of:* Cole, Peter, & Morgan, Jerry L. (eds), *Syntax and Semantics: Speech Acts*. New York: Academic Press.

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

Griffiths, Paul J. 2010. *Lying: An Augustinian Theology of Duplicity*. Wipf and Stock Publishers.

Grimaltos, Tobies, & Rosell, Sergi. 2013. On Lying: A Conceptual Argument for the Falsity Condition. *Unpublished*.

Groenendijk, Joroen, & Stokhof, Martin. 1982. Semantic analysis of wh-complements. *Linguistics and Philosophy*, **5(2)**, 175–233.

Grotius, Hugo. 2005. The Rights of War and Peace, 3 Vols. *Indianapolis: Liberty Fund*.

Hainguerlot, Marine, Vergnaud, Jean-Christophe, & de Gardelle, Vincent. 2018. Metacognitive ability predicts learning cue-stimulus associations in the absence of external feedback. *Scientific reports*, **8**(1), 5602.

Halpin, Stanley M., Moses, Franklin L., & Johnson, Edgar M. 1978. A Validation of the Structure of Combat Intelligence Ratings. *US Army Technical Paper*, **302**.

Harder, Peter, & Kock, Christian Erik J. 1976. *The theory of presupposition failure*. Akademisk forlag.

Hawthorne, John. 2003. *Knowledge and Lotteries*. Oxford University Press.

van der Henst, Jean-Baptiste, Carles, Laure, & Sperber, Dan. 2002. Truthfulness and Relevance in Telling the Time. *Mind & Language*, **17(5)**, 457–466.

Herzig, Andreas, Lorini, Emiliano, Hübner, Jomi F., & Vercouter, Laurent. 2010. A logic of trust and reputation. *Logic Journal of the IGPL*, **18**(1), 214–244.

Hobby, Jonathan L., Tom, BD, Todd, C, Bearcroft, PW, & Dixon, Adrian K. 2000. Communication of doubt and certainty in radiological reports. *The British Journal of Radiology*, **73**(873), 999–1001.

Horn, Laurence R. 2004. Implicature. *In:* Horn, Laurence R, & Ward, Gergory (eds), *The Handbook of Pragmatics*. Oxford: Blackwell.

Horn, Laurence R. 2006. The border wars: A neo-Gricean perspective. *Where semantics meets pragmatics*, **16**, 21–48.

Hyde, Dominic. 1997. From heaps and gaps to heaps of gluts. *Mind*, **106**(424), 641–660.

Isenberg, Arnold. 1965. Conditions for Lying. *Ethical theory and business*, 466–468.

Jeffreys, Harold. 1939. *Theory of probability*. Oxford: Clarendon Press.

Jeffreys, Harold. 1946. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **186**(1007), 453–461.

Johnson, Edgar M. 1973. Numerical Encoding of Qualitative Expressions of Uncertainty. *ARI Technical Paper*, **250**.

Kamp, Hans. 1975. Two theories about adjectives. *In:* Keenan, E. (ed), *Formal Semantics of Natural Language*. Cambridge University Press.

Kamp, Hans, & Partee, Barbara. 1995. Prototype theory and compositionality. *Cognition*, **57**(2), 129–191.

Keefe, Rosanna. 2000. *Theories of Vagueness*. Cambridge University Press.

Keiser, Jessica. 2016. Bald-faced lies: how to make a move in a language game without making a move in a conversation. *Philosophical Studies*, **173**(2), 461–477.

Kelly, C. W., & Peterson, C.R. 1971. Probability Estimates and Probabilistic Procedures in Current Intelligence Analysis. *Report, International Business Machines Corporation*, **71-5047**.

Kennedy, Christopher. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, **30(1)**, 1–45.

Kennedy, Christopher. 2013. Two sources of subjectivity: Qualitative assessment and dimensional uncertainty. *Inquiry*, **56**(2-3), 258–277.

Kent, Sherman. 1964. Words of Estimative Probability. *Studies in Intelligence*, **8**.

Klein, Ewan. 1980. A semantics for positive and comparative adjectives. *Linguistics and philosophy*, **4**(1), 1–45.

Knobe, Joshua. 2003a. Intentional action and side effects in ordinary language. *Analysis*, **63**(279), 190–194.

Knobe, Joshua. 2003b. Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, **16**(2), 309–324.

Knobe, Joshua. 2006. The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*, **130**(2), 309–324.

Kölbel, Max. 2004. Faultless disagreement. *Pages 53–73 of: Proceedings of the Aristotelian Society*, vol. 104.

Kratzer, Angelika. 1981. The notional category of modality. *In:* Eikmeyer, Hans-Jürgen, & Rieser, Hannes (eds), *Words, Worlds, and Contexts*. De Gruyter.

Kratzer, Angelika. 1991. Modality. *In:* von Stechow, Arnim, & Wunderlich, Dieter (eds), *Semantics: An international handbook of contemporary research*. De Gruyter.

Krifka, Manfred. 2007. Approximate interpretation of number words. *Pages 111–126 of:* Bouma, G., Krämer, I., & Zwarts, J. (eds), *Cognitive Foundations of Interpretation*. Humboldt-Universität zu Berlin, Philosophische Fakultät II.

Krishna, Daya. 1961. 'Lying'and the Compleat Robot. *The British Journal for the Philosophy of Science*, **12**(46), 146–149.

Kvanvig, Jonathan. 2009. Assertion, knowledge, and lotteries. *In:* Pritchard, Duncan, & Greenough, Patrick (eds), *Williamson on Knowledge*. Oxford University Press.

Kvanvig, Jonathan. 2011. Norms of assertion. *In:* Brown, Jessica, & Cappelen, Herman (eds), *Assertion: New philosophical essays*. Oxford University Press.

Lackey, Jennifer. 2007. Norms of assertion. *Noûs*, **41**(4), 594–626.

Lackey, Jennifer. 2013. Lies and deception: an unhappy divorce. *Analysis*, **73**(2), 236–248.

Lakoff, George. 1973. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, **2**(4), 458–508.

Lappin, Shalom. 2000. An intensional parametric semantics for vague quantifiers. *Linguistics and philosophy*, **23(6)**, 599–620.

Lasersohn, Peter. 1999. Pragmatic halos. *Language*, 522–551.

Lassiter, Daniel. 2011. Vagueness as Probabilistic Linguistic Knowledge. *Pages 127–150 of: Vagueness in Communication*. Springer.

Lassiter, Daniel. 2017. *Graded modality: Qualitative and quantitative perspectives*. Oxford University Press.

Lassiter, Daniel, & Goodman, Noah D. 2015. Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, 1–36.

Lee, Michael D., & Dry, Matthew J. 2006. Decision making and confidence given uncertain advice. *Cognitive Science*, **30**(6), 1081–1095.

Lee, Yang, Strong, Diane, Kahn, Beverly, & Wang, Richard. 2002. AIMQ: a Methodology for Information Quality Assessment. *Information & management*, **40**(2), 133–146.

Legastelois, Bénédicte, Lesot, Marie-Jeanne, & Revault d'Allonnes, Adrien. 2017a. A Fuzzy Take on Graded Beliefs. *Pages 392–404 of: Advances in Fuzzy Logic and Technology 2017*. Springer.

Legastelois, Bénédicte, Lesot, Marie-Jeanne, & Revault d'Allonnes, Adrien. 2017b. Typology of axioms for a weighted modal logic. *International Journal of Approximate Reasoning*, **90**, 341–358.

Leland, Patrick R. 2015. Rational responsibility and the assertoric character of bald-faced lies. *Analysis*, **75**(4), 550–554.

Lesot, Marie-Jeanne, & Revault d'Allonnes, Adrien. 2017. Information quality and uncertainty. *Pages 135–146 of: Uncertainty Modeling*. Springer.

Lesot, Marie-Jeanne, Delavallade, Thomas, Pichon, Frédéric, Akdag, Herman, Bouchon-Meunier, Bernadette, & Capet, Philippe. 2011. Proposition of a semi-automatic possibilistic information scoring process. *Pages 949–956 of: The 7th Conf. of the European Society for Fuzzy Logic and Technology (EUSFLAT-2011) and LFA-2011*. Atlantis Press.

Lesot, Marie-Jeanne, Pichon, Frédéric, & Delavallade, Thomas. 2013. Cotation quantitative de l'information : modélisation et évaluation expérimentale. *In:* Capet, Philippe, & Delavallade, Thomas (eds), *L'évaluation de l'information: confiance et défiance*. Hermès.

Levine, Jerrold M., & Eldredge, Donald. 1970. The Effects of Ancillary Information upon Photointerpreter Performance. *Report, American Institutes for Research*, **70-14**.

Levinson, Stephen C. 1995. Three levels of meaning. *Pages 90–115 of: Grammar and meaning: Essays in honour of Sir John Lyons*. Cambridge University Press.

Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Lewis, David. 1979. General semantics. *Synthese*, **22(1)**, 18–67.

Lewis, David. 1982. Logic for equivocators. *Noûs*, **16**, 431–441.

Lichtenstein, Sarah, & Newman, Robert J. 1967. Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science*, **9**(10), 563–564.

Liu, Fenrong. 2009. Diversity of agents and their interaction. *Journal of Logic, Language and information*, **18**(1), 23–53.

Liu, Fenrong, & Wang, Yanjing. 2013. Reasoning about agent types and the hardest logic puzzle ever. *Minds and Machines*, **23**(1), 123–161.

Lorini, Emiliano, & Castelfranchi, Cristiano. 2006. The unexpected aspects of surprise. *International Journal of Pattern Recognition and Artificial Intelligence*, **20**(06), 817–833.

Lykken, David T. 1959. The GSR in the detection of guilt. *Journal of Applied Psychology*, **43**(6), 385.

Lykken, David T. 1998. *A tremor in the blood: Uses and abuses of the lie detector*. Plenum Press.

MacFarlane, John. 2014. *Assessment sensitivity: Relative truth and its applications*. OUP Oxford.

Mahon, James E. 2015. The definition of lying and deception. *The Stanford Encyclopedia of Philosophy*.

Marsili, Neri. 2014. Lying as a scalar phenomenon. *Certainty-uncertainty and the Attitudinal Space in Between*, **165**, 153.

Marsili, Neri. 2018. Lying and Certainty. *In:* Meibauer, Jörg (ed), *Oxford Handbook of Lying*. Oxford University Press.

McCloskey, Michael E., & Glucksberg, Sam. 1978. Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, **6(4)**, 462–472.

McKinnon, Rachel. 2016. *The norms of assertion: Truth, lies, and warrant*. Springer.

McNally, Louise, & Stojanovic, Isidora. 2017. Aesthetic adjectives. *Pages 17–37 of:* Young, James O. (ed), *The Semantics of Aesthetic Judgment*. Oxford University Press.

Meeland, T., & Rhyne, R. F. 1967. A Confidence Scale for Intelligence Reports: an application of magnitude estimation scaling. *Technical Note, Stanford Research Institute*, **4923-31**.

Mehlberg, Henry. 1958. *The reach of science*. University of Toronto Press.

Meibauer, Jörg. 2005. Lying and falsely implicating. *Journal of Pragmatics*, **37**(9), 1373–1399.

Meibauer, Jörg. 2011. *On lying: intentionality, implicature, and imprecision*.

Meibauer, Jörg. 2014. *Lying at the semantics-pragmatics interface*. Walter de Gruyter GmbH & Co KG.

Meibauer, Jörg. 2016a. Topics in the linguistics of lying: A reply to Marta Dynel. *Intercultural Pragmatics*, **13**(1), 107–123.

Meibauer, Jörg. 2016b. Understanding bald-faced lies. *International Review of Pragmatics*, **8**(2), 247–270.

Meijer, Ewout H., & Verschuere, Bruno. 2015. The polygraph: Current practice and new approaches. *Detecting deception: Current challenges and cognitive approaches*, 59–80.

Meijer, Ewout H., Verschuere, Bruno, Gamer, Matthias, Merckelbach, Harald, & Ben-Shakhar, Gershon. 2016. Deception detection with behavioral, autonomic, and neural measures: Conceptual and methodological considerations that warrant modesty. *Psychophysiology*, **53**(5), 593–604.

Meyer, Wulf-Uwe, Reisenzein, Rainer, & Schützwohl, Achim. 1997. Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion*, **21**(3), 251–274.

Miron, Murray S., Patten, Samuel, & Halpin, Stanley M. 1978. The Structure of Combat Intelligence Ratings. *US Army Technical Paper*, **286**.

Mullaney, Steven. 1980. Lying like truth: riddle, representation and treason in Renaissance England. *ELH*, **47(1)**, 32–47.

Nimier, Vincent. 2005. Information evaluation: A formalisation of operational recommendations. *NATO Science Series, Sub Series III - Computer and Systems Sciences*, **198**, 81.

Nimier, Vincent, & Appriou, Alain. 1995. Utilisation de la Théorie de Dempster-Shafer pour la Fusion d'Informations. *In: 15th Colloque sur le traitement du signal et des images, France*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images.

Noveck, Ira A., & Reboul, Anne. 2008. Experimental pragmatics: A Gricean turn in the study of language. *Trends in Cognitive Sciences*, **12**(11), 425–431.

O'Brien, Bernie J. 1989. Words or numbers? The evaluation of probability expressions in general practice. *JR Coll Gen Pract*, **39**(320), 98–100.

O'Connor, Donald J. 1948. Pragmatic paradoxes. *Mind*, **57**(227), 358–359.

ODN. 2011. National Intelligence: A consumer's guide. *Office of the Director of National Intelligence*.

Ortony, Andrew, & Partridge, Derek. 1987. Surprisingness and expectation failure: what's the difference? *Pages 106–108 of: IJCAI*.

Parikh, Rohit. 1994. Vagueness and utility: The semantics of common nouns. *Linguistics and Philosophy*, **17(6)**, 521–535.

Partee, Barbara. 1989. Binding Implicit Variables in Quantified Contexts. *In:* Wiltshire, C., Graczyk, R., & Music, B. (eds), *Papers from the 25th Annual Regional Meeting of the Chicago Linguistic Society*. Chicago Linguistic Society.

Pavlick, Ellie, & Callison-Burch, Chris. 2016. So-called non-subsective adjectives. *Pages 114–119 of: Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*.

Pearl, Judea. 2014. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.

Peirce, Charles S. 1902. Vague. *In:* Baldwin, J. M. (ed), *Dictionary of Philosophy and Psychology*. Macmillan.

Peterson, Candida C., Peterson, James L., & Seeto, Diane. 1983. Developmental Changes in Ideas about Lying. *Child Development*, **54**(6), 1529–1535.

Peterson, Joshua J. 2008. Appropriate factors to consider when assessing analytic confidence in intelligence analysis. *Master of Science Thesis, Department of Intelligence Studies, Mercyhurst College, Erie, Pennsylvania*.

Pettit, Dean, & Knobe, Joshua. 2009. The pervasive impact of moral judgment. *Mind & Language*, **24**(5), 586–604.

Phelps, Ruth H., Halpin, Stanley M., Johnson, Edgar M., & Moses, Franklin L. 1980. Implementation of Subjective Probability Estimates In Army Intelligence Procedures: A Critical Review of Research Findings. *Research Report US Army*, **1242**.

Piaget, Jean. 1932. The Moral Judgment of the Child. *Kegan Paul, London*.

Pichon, Frédéric, Dubois, Didier, & Denoeux, Thierry. 2012. Relevance and truthfulness in information correction and fusion. *International Journal of Approximate Reasoning*, **53**(2), 159–175.

Pighin, Stefania, Bonnefon, Jean-François, & Savadori, Lucia. 2011. Overcoming number numbness in prenatal risk communication. *Prenatal diagnosis*, **31**(8), 809–813.

Pinkal, Manfred. 1995. *Logic and lexicon: the semantics of the indefinite*. Kluwer Academics Publishers.

Plaza, Jan. 2007. Logics of public communications. *Synthese*, **158**(2), 165.

Popper, Karl R. 1963. *Conjectures and refutations: the growth of scientific knowledge*. New York: Basic Books.

Posner, Roland. 1980. Semantics and pragmatics of sentence connectives in natural language. *Pages 169–203 of: Speech act theory and pragmatics*. Springer.

Primoratz, Igor. 1984. Lying and the "Methods of Ethics". *International Studies in Philosophy*, **16**(3), 35–57.

Raffman, Diana. 2013. *Unruly words: a study of vague language*. Oxford University Press.

Ransom, Keith, Voorspoels, Wouter, Perfors, Amy, & Dani, Navarro. 2017. A cognitive analysis of deception without lying. *Pages 992–997 of: Proceedings of the 39th Annual Conference of the Cognitive Science Society.* Cognitive Science Society.

Récanati, François. 2004. *Literal meaning*. Cambridge University Press.

Research Council, National. 2003. *The polygraph and lie detection*.

Revault d'Allonnes, Adrien, & Lesot, Marie-Jeanne. 2014. Formalising information scoring in a multivalued logic framework. *Pages 314–324 of:* Laurent, Anne, Strauss, Olivier,

Bouchon-Meunier, Bernadette, & Yager, Ronald R (eds), *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer.

Revault d'Allonnes, Adrien, & Lesot, Marie-Jeanne. 2015. Dynamics of trust building: models of information cross-checking in a multivalued logic framework. *Pages 1–8 of: International Conference on Fuzzy Systems*. IEEE.

Revault d'Allonnes, Adrien, Akdag, Herman, & Poirel, Olivier. 2007. Trust-moderated information-likelihood. A multi-valued logics approach. *Pages 1–6 of: Computability in Europe-Computation and Logic in the Real World*.

Rodenhäuser, Benjamin. 2014. *A matter of trust: Dynamic attitudes in epistemic logic*. Universiteit van Amsterdam [Host].

Rogova, Galina L., & Nimier, Vincent. 2004. Reliability in information fusion: literature survey. *Pages 1158–1165 of: Proceedings of the seventh international conference on information fusion*, vol. 2.

Rosch, Eleanor H. 1973. Natural Categories. *Cognitive psychology*, **4**(3), 328–350.

Russell, Bertrand. 1923. Vagueness. *The Australasian Journal of Psychology and Philosophy*, **1**(2), 84–92.

Rutschmann, Ronja, & Wiegmann, Alex. 2017. No need for an intention to deceive? Challenging the traditional definition of lying. *Philosophical Psychology*, **30**(4), 438–457.

Ryle, Gilbert. 2009. *The concept of mind*. Routledge.

Sakama, Chiaki. 2015. A Formal Account of Deception. *Pages 34–41 of: Proceedings of the AAAI Fall 2015 Symposium on Deceptive and Counter-Deceptive Machines*.

Sakama, Chiaki, & Caminada, Martin. 2010. The Many Faces of Deception. *Proceedings of the Thirty Years of Nonmonotonic Reasoning*.

Sakama, Chiaki, Caminada, Martin, & Herzig, Andreas. 2010. A Logical Account of Lying. *Pages 286–299 of: European Workshop on Logics in Artificial Intelligence*. Springer.

Sakama, Chiaki, Caminada, Martin, & Herzig, Andreas. 2014. A Formal Account of Dishonesty. *Logic Journal of the IGPL*, **23**(2), 259–294.

Samet, Michael G. 1975. Quantitative Interpretation of Two Qualitative Scales Used to Rate Militarv Intelligence. *Human Factors*, **17(2)**, 192–202.

Sapir, Edward. 1944. Grading, a study in semantics. *Philosophy of science*, **11(2)**, 93–116.

Sassoon, Galit W. 2012. A Typology of Multidimensional Adjectives. *Journal of Semantics*, **30**, 335–380.

Saul, Jennifer M. 2012. *Lying, misleading, and what is said: An exploration in philosophy of language and in ethics*. Oxford University Press.

Scarantino, Andrea, & Piccinini, Gualtiero. 2010. Information without truth. *Metaphilosophy*, **41**(3), 313–330.

Schaffer, Jonathan. 2008. Knowledge in the image of assertion. *Philosophical issues*, **18**(1), 1–19.

Scheppele, Kim L. 1988. *Legal secrets: Equality and efficiency in the common law*. University of Chicago Press.

Schum, David A. 1987. *Evidence and inference for the intelligence analyst*. Vol. 1. University Press of America.

Schweitzer, Nicholas. 1978. Bayesian analysis: estimating the probability of Middle East conflict. *Quantitative Approaches to Political Intelligence: The CIA Experience*, 23–24.

Scott, Gini G. 2006. The Truth About Lying. *NE: ASJA Press*.

Scriven, Michael. 1951. Paradoxical announcements. *Mind*, **60**(239), 403–407.

Seridi, Hamid, & Akdag, Herman. 2001. Approximate Reasoning for Processing Uncertainty. *JACIII*, **5**(2), 110–118.

Seymour, Travis L., & Kerlin, Jess R. 2008. Successful detection of verbal and visual concealed knowledge using an RT-based paradigm. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, **22**(4), 475–490.

Shafer, Glenn. 1976. *A mathematical theory of evidence*. Vol. 42. Princeton University Press.

Shannon, Claude E., & Weaver, Warren. 1949. *The Mathematical Theory of Information*. University of Illinois Press.

Shaw, R. 1958. The paradox of the unexpected examination. *Mind*, **67**(267), 382–384.

Shibles, Warren. 1987. Lying: A critical analysis. *Revue Philosophique de la France Et de l'Etranger*, **177**(1).

Shiffrin, Seana V. 2014. *Speech matters: On lying, morality, and the law*. Princeton University Press.

Simpson, David. 1992. Lying, liars and language. *Philosophy and Phenomenological Research*, **52**(3), 623–639.

Smith, David L. 2007. *Why we lie: The evolutionary roots of deception and the unconscious mind*. Macmillan.

Smullyan, Raymond M. 1978. *What is the Name of this Book? The Riddle of Dracula and Other Logical Puzzles: Mysteries, Paradoxes, Gödel's Discovery*. Prentice-Hall.

Solt, Stephanie. 2015. Vagueness and imprecision: Empirical foundations. *Annu. Rev. Linguist.*, **1(1)**, 107–127.

Sorensen, Roy. 2007. Bald-Faced Lies! Lying without the Intent to Deceive. *Pacific Philosophical Quarterly*, **88**(2), 251–264.

Sorensen, Roy. 2012. Lying with conditionals. *The Philosophical Quarterly*, **62**(249), 820–832.

Sorensen, Roy A. 1988. *Blindspots*. Oxford: Clarendon Press.

Spector, Benjamin. 2013. Bare numerals and scalar implicatures. *Language and Linguistics Compass*, **7**(5), 273–294.

Spohn, Wolfgang. 1988. Ordinal Conditional Functions: A Dynamic Theory of Epistemic States. *Pages 105–134 of:* Harper, William L., & Skyrms, Brian (eds), *Causation in Decision, Belief Change, and Statistics: Proceedings of the Irvine Conference on Probability and Causation*. Dordrecht: Springer Netherlands.

STANAG-2511. 2003. Standardization Agreement, Intelligence Report. *North Atlantic Treaty Organization*, **2511**, 1–15.

Stokke, Andreas. 2013. Lying, deceiving, and misleading. *Philosophy Compass*, **8**(4), 348–359.

Suchotzki, Kristina, Verschuere, Bruno, van Bockstaele, Bram, Ben-Shakhar, Gershon, & Crombez, Geert. 2017. Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin*, **143**(4), 428–453.

Tiersma, Peter. 2004. Did Clinton Lie: Defining "Sexual Relations". *Chicago-Kent Law Review*, **79(3)**.

TTA. 2001. TRAITÉ TOUTES ARMES. *Renseignement*.

Turri, Angelo, & Turri, John. 2015. The truth about lying. *Cognition*, **138**, 161–168.

Turri, Angelo, & Turri, John. 2016. Lying, uptake, assertion, and intent. *International Review of Pragmatics*, **8**(2), 314–333.

Turri, John. 2010. Epistemic invariantism and speech act contextualism. *Philosophical Review*, **119**(1), 77–95.

Veltman, Frank. 1996. Defaults in update semantics. *Journal of philosophical logic*, **25**(3), 221–261.

Verheyen, Steven, Dewil, Sabrina, & Egré, Paul. 2017. *Subjective meaning in gradable adjectives: The case of* tall *and* heavy. Manuscript, IJN, under review.

Verschuere, Bruno, Suchotzki, Kristina, & Debey, Evelyne. 2014. Detecting deception through reaction times. *Detecting deception: current challenges and cognitive approaches*, 269–291.

Viebahn, Emanuel. 2017. Non-literal lies. *Erkenntnis*, **82**(6), 1367–1380.

Vincent, Jocelyne M., & Castelfranchi, Cristiano. 1981. On the art of deception: how to lie while saying the truth. *Possibilities and limitations of pragmatics*, 749–777.

Vrij, Aldert. 2000. *Detecting lies and deceit: The psychology of lying and implications for professional practice*. Wiley.

Vrij, Aldert, Fisher, Ronald, Mann, Samantha, & Leal, Sharon. 2006. Detecting deception by manipulating cognitive load. *Trends in cognitive sciences*, **10**(4), 141–142.

Waismann, Friedrich. 1945. Verifiability. *Proceedings of the Aristotelian Society, Supplementary Volumes*, **19**, 119–150.

Wakeling, Edward. 1992. Who is telling the truth? *In:* Wakeling, Edward (ed), *Lewis Carroll's Games and Puzzles*. Dover Publications.

Wand, Yair, & Wang, Richard Y. 1996. Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM*, **39**(11), 86–95.

Wang, Richard Y. 1998. A Product Perspective on Total Data Quality Management. *Communications of the ACM*, **41**(2), 58–65.

Wark, David L. 1964. The Definition of Some Estimative Expressions. *Studies in Intelligence*, **8**(4), 67–80.

Wason, Peter C. 1966. Reasoning. *New Horizons in Psychology*, **1**, 135–151.

Weber, Elke U., & Hilton, Denis J. 1990. Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance*, **16**(4), 781.

Weiner, Matt. 2007. Norms of assertion. *Philosophy Compass*, **2**(2), 187–195.

Wiegmann, Alex, & Willemsen, Pascale. 2017. How the truth can make a great lie: An empirical investigation of lying by falsely implicating. *Pages 3516–3621 of: Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Wiegmann, Alex, Samland, Jana, & Waldmann, Michael R. 2016. Lying despite telling the truth. *Cognition*, **150**, 37–42.

Wiegmann, Alex, Rutschmann, Ronja, & Willemsen, Pascale. 2017. Empirically Investigating the Concept of Lying. *Journal of Indian Council of Philosophical Research*, **34**(3), 591–609.

Williams, Bernard. 1985. Ethics and the Limits of Philosophy. *Philosophy*, 156–173.

Williams, Bernard. 2002. *Truth and Truthfulness: An Essay in Genealogy*. Princeton: New Jersey: Princeton University Press.

Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford University Press.

Wittgenstein, Ludwig. 1953. *Philosophical investigations*. Oxford: Blackwell.

Wright, Crispin. 1995. The epistemic conception of vagueness. *The Southern journal of philosophy*, **33**(S1), 133–160.

Zadeh, Lotfi A. 1978. Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems*, 3–28.

Zlotnick, Jack. 1972. Bayes' theorem for intelligence analysis. *Studies in Intelligence*, **16**(2), 43–52.

Résumé

Cette thèse vise à mieux *définir* ainsi qu'à mieux *évaluer* les stratégies de tromperie et de manipulation de l'information. Des ressources conceptuelles, formelles et expérimentales sont combinées en vue d'analyser des cas *standards* de tromperie, tels que le *mensonge*, mais aussi *non-standards*, tels que les *inférences trompeuses* et *l'omission stratégique*.

Les aspects définitionnels sont traités en premier. J'analyse la définition traditionnelle du *mensonge* en présentant des résultats empiriques en faveur de cette définition classique (dite 'définition subjective'), contre certains arguments visant à défendre une 'définition objective' par l'ajout d'une condition de fausseté. J'examine ensuite une énigme logique issue de R. Smullyan, et qui porte sur un cas limite de tromperie basé sur une *règle d'inférence par défaut* pour tromper un agent *par omission*.

Je traite ensuite des aspects évaluatifs. Je pars du cadre existant pour l'évaluation du renseignement et propose une typologie des messages fondée sur les *dimensions descriptives* de *vérité* (pour leur contenu) et *d'honnêteté* (pour leur source). Je présente ensuite une procédure numérique pour l'évaluation des messages basée sur les *dimensions évaluatives* de *crédibilité* (pour la vérité) et de *fiabilité* (pour l'honnêteté). Des modèles numériques de plausibilité servent à capturer la crédibilité *a priori* des messages puis des règles numériques sont proposées pour *actualiser* ces degrés selon la fiabilité de la source.

Mots-Clés

Mensonge, Tromperie, Omission, Qualité de l'information, Désinformation, Mésinformation, Évaluation de l'information, Vague, Logique épistémique, Traitement du renseignement.

Abstract

This thesis aims at improving the *definition* and *evaluation* of deceptive strategies that can manipulate information. Using conceptual, formal and experimental resources, I analyze *three deceptive strategies*, some of which are *standard cases* of deception, in particular *lies*, and others *non-standard cases* of deception, in particular *misleading inferences* and *strategic omissions*.

Firstly, I consider *definitional aspects*. I deal with the definition of *lying*, and present new empirical data supporting the traditional account of the notion (called the 'subjective definition'), contradicting recent claims in favour of a falsity clause (leading to an 'objective definition'). Next, I analyze non-standard cases of deception through the categories of *misleading defaults* and *omissions of information*. I use qualitative belief revision to examine a puzzle due to R. Smullyan about the possibility of triggering a *default inference* to deceive an addressee *by omission*.

Secondly, I consider *evaluative aspects*. I take the perspective of military intelligence data processing to offer a typology of informational messages based on the *descriptive dimensions* of *truth* (for message contents) and *honesty* (for message sources). I also propose a numerical procedure to evaluate these messages based on the *evaluative dimensions* of *credibility* (for truth) and *reliability* (for honesty). Quantitative plausibility models are used to capture degrees of *prior* credibility of messages, and dynamic rules are defined to *update* these degrees depending on the reliability of the source.

Keywords

Lying, Deception, Omission, Information Quality, Disinformation, Misinformation, Information Evaluation, Vagueness, Epistemic Logic, Intelligence Data Processing.