# The Logic of Framing

The Framing Effect as a Non-Monotonic Decision Process of
Path Dependence

**MSc Thesis** *(Afstudeerscriptie)*

written by

**Michiel den Haan**
(born 31 January, 1991 in Aalsmeer, The Netherlands)

under the supervision of **Prof. Dr. Ing. Robert van Rooij**, and
submitted to the Board of Examiners in partial fulfillment of the
requirements for the degree of

**MSc in Logic**

at the *Universiteit van Amsterdam.*

**Date of the public defence:**
*26 June, 2015*

**Members of the Thesis Committee:**
Dr. Maria Aloni (chair)
Prof. Dr. Michiel van Lambalgen
Prof. Dr. Ing. Robert van Rooij (supervisor)
Prof. Dr. Frank Veltman

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

To hatch a crow, a black rainbow
Bent in emptiness
                    over emptiness
But flying

- Ted Hughes (1930 - 1998)

# Abstract

People's perception can be influenced by framing a message in different ways. For example, decision makers are often favourably disposed towards a medical treatment with a 60% success rate, but less often to one with a 40% failure rate. This suggests that human judgement or decision making is not merely concerned with the content of the message at stake, but also with the way in which this content is presented or 'framed'. This observation is at odds with standard theories of decision making, which prescribe that human judgement should be invariant under different descriptions of the same fact(s). In this thesis, I explore what framing is, how it makes itself manifest and how the underlying process of human decision making functions. I argue that this decision process is best to be characterised as a non-monotonic process of path dependence, and I present a formal model to make this idea more precise. This model has several advantages over existing models. It is able to provide an elaborate account of framing that is descriptively accurate, more uniform and has a larger scope than existing accounts. Furthermore, it allows for a thoughtful approach to the relationship between the framing effect and human rationality.

# Contents

# Introduction

Would you prefer a medical treatment with a 60% success rate or one with a 40% failure rate? And would you vote for an economic policy plan that yields 95% employment or one that yields 5% unemployment? Do you like your glass half full or half empty? These question may seem silly, but they take centre stage in research on 'framing', a topic studied extensively in, for example, cognitive psychology, political science, communication studies and behavioural economics. As it turns out, in all the cases above people are more often favourably disposed towards the former option than to the latter, despite the fact that both options seem to describe one and the same thing.

Framing is ubiquitous in our everyday lives. For example, we encounter press releases of big companies about 'restructurings' or 'reorganisations' rather than redundancies or job losses. We see advertisements of soda companies stressing the percentage of 'pure fruit juice' in their products rather than the percentage of sugar and additives. And we hear politicians talk about the success rate of their new job placement policy, rather than the failure rate.

This suggests that in making decisions, human beings do not merely take the *content* of the choice options at stake into account (e.g., the effectiveness of a medical treatment), but also the way in which these options are presented or 'framed' (e.g., in terms of success or failure). If this manner of presentation affects the judgement of the decision maker, one speaks of 'the framing effect'.

In this thesis, I will study what the framing effect is, how it comes about, how the underlying process of decision making works and what the consequences of framing are for human rationality. The results of this investigation will be used to construct a formal model that is able to provide an elaborate account of the framing effect.

### Framing and Decision Making

The framing effect will be studied in this thesis from a decision theoretic purview. The point of focus is how we perceive different frames, how differences in presentation influence decision making and how they can induce us to act.

Perhaps surprisingly, relatively little has been written about these issues. Despite the fact that framing is a fruitful research area today, it has long been neglected in theories about human decision making. One reason for this is that framing, in the words of Robert Entman, is a “fractured paradigm” (Entman 1993, p.51). That is, framing is studied by various disciplines, from different points of view, each focusing on different sides of the phenomenon, but the results of these endeavours are rarely connected.

For instance, in cognitive psychology various settings are studied in which the framing effect is observed, as well as factors that can enhance or attenuate the effect. Typically, these studies focus on individual test subjects. In political science and communication studies, on the other hand, the persuasiveness and effectiveness of framing is investigated, as well as it relation to, for example, source reliability and reputation effects. Here, studies typically focus on groups and the consequences for public opinion or the way in which public debates are perceived.

Furthermore, the framing effect has, for instance, been studied from linguistic perspectives, focusing on the semantic properties of different formulations. It has been studied from evolutionary perspectives, focusing on possible advantages of responding differently to positive and negative information. And it has been studied from contextual perspectives, focusing on regularities between certain beliefs or events and the words people choose to describe these events.

What is seldom done, however, is to provide an overarching analysis of the framing effect by combining the results from these various fields, studies and experiments. For a decision-theoretic study, such analysis is crucial. A systematic investigation into what framing is, the circumstances in which it appears or disappears and the various factors that are at work is needed in order to obtain a proper understanding of how framing influences our perception, decision making and behaviour.

Two decades ago, Entman articulated this problem as follows: “Despite its omnipresence across the social sciences and humanities, nowhere is there a general statement of framing theory that shows exactly how frames [...] make themselves manifest [...] or how framing influences thinking” (ibid.). Today, these issues still stand in need of clarification, and will therefore be taken as the point of departure of this thesis.

### Decision Theory and Rationality

A second reason why the framing effect has received little attention from decision theory has to do with the historical development of this research area.

For a long time, decision making has been studied from the purview of ‘perfectly rational agents’. According to this view, a decision maker is a fully rational agent that "is assumed to have knowledge of the relevant aspects of his environment [...] He is assumed also to have a well-organized and stable system of preferences, and a skill in computation that enables him to calculate, for the alternative courses of action that are available to him, which of these will permit him to reach the highest attainable point on his preference scale" (Simon 1955, p.99).

At the core of this conception of rationality lies the principle of description invariance, or, as logicians call it, the principle of extensionality. Kenneth Arrow characterises this principle as a "fundamental element of rationality, so elementary that we hardly notice it" (Arrow 1982, p.6). The principle states that preferences should be unaffected by the way a problem or choice is described. What counts is *what* is described, not *how* it is described.

From the perspective of a logician, this principle makes perfect sense. The logical implications of some state of affairs are taken to depend on this state of affairs itself, not on the way it is presented. For instance, whether I use the formulation 'the glass is half full' or 'the glass is half empty' to describe some state of affairs (the amount of liquid in a glass), the truth value of this state of affairs and the consequences it has depends on the logical properties and implications of the actual amount of liquid in the glass, not on its being described as half full or half empty. Logicians tend to say that "two formulas which have the same truth-value under any truth-assignments [are] substitutable *salva veritate* in a sentence that contains one of these formulas" (Bourgeois-Gironde and Giraud 2009, p.386).

As a result, standard decision theory, assuming that decision makers are perfectly rational, is unable to accommodate the framing effect. Given that the principle of extensionality holds, the effect is predicted not to occur. That the effect does occur in practice has long been dismissed as an insignificant aberration of the norms of rationality, not of interest for decision theory (ibid.). This was mainly due to the normative, rather than descriptive, character of theories of human decision making.

It was not until the 1980s, with the pioneering work of Amos Tversky and Daniel Kahneman, that the framing effect and its consequences for decision making were systematically investigated for the first time. I will discuss their work in detail in chapter 1. It is interesting to note that Tversky and Kahneman did not reject standard decision theory. In fact, they argue that the principle of description invariance, even though "descriptively invalid", is "normatively essential" (Tversky and Kahneman 1986, S251). That is, the principle should be valid in any a theory of choice that claims normative status (ibid., S253).

Hence, rather than rejecting standard decision theory, they argue that "the normative and the descriptive analyses of choice should be viewed as separate enterprises" (Tversky and Kahneman 1986, S275). If one wants to know what decisions humans *should* make, standard decision theory is the way to go. As argued above, in this setting the occurrence of the framing effect is precluded. Tversky and Kahneman themselves provide a descriptive account of human decision making, i.e., what decisions humans *do* make. The model they present is able to accommodate the framing effect, but this automatically means that for it to occur, some rationality principles have to be violated. This idea that the framing effect is irrational 'tout court' is still widespread today (e.g., O'Keefe 2007, Marcus 2008).

In this thesis, I will take a different course. The notion of rationality outlined above is a highly idealised and unrealistic one. If one adopts a more moderate notion of rationality, tailored to the human mind and the context of decision making, this conclusion no longer follows. Rather than dismissing the framing effect as 'irrational', I will argue that much more is to be said about framing. In fact, in some cases, violating the principle of description invariance can be perfectly reasonable.

### Outline

This thesis is structured as follows. In chapter 1, I will investigate how frames make themselves manifest, in order to obtain a thorough understanding of what the framing effect is and to fathom what factors are at work in its occurrence. An overview of the framing effect and its various appearances will be provided, and Tversky and Kahneman's Prospect Theory, a formal model that has dominated the framing literature for more than two decades, will be presented and criticised. I will argue that framing is a much more diverse phenomenon than generally recognised, and that there are various factors at work that cause different frames to trigger different associations. Furthermore, I will argue that despite the diversity of the framing effect, a similar decision process can be said to underlie all instances of framing.

In chapter 2, I will probe into this decision process that underlies the framing effect. Some alternatives to Prospect Theory will be discussed, and I will argue that they all fail to provide a robust account of the framing effect. Furthermore, some empirical results will be presented that will prove to be problematic for all existing accounts of framing. Based on this and other results, I will provide an extensive characterisation of the underlying decision process of framing. Furthermore, the consequences for the alleged irrationality of the framing effect will be explored.

In the final chapter, the insights gained so far will be used to formulate four desiderata for a formal model of the framing effect. I will develop a dynamic model that accommodates these four desiderata, and which provides a robust account of framing in its various forms.

# 1 The Diversity and Uniformity of Framing

In this chapter, the framing effect will be studied in its various shapes. I will assess the psychological and cognitive science literature on framing to investigate how it comes about and how it is related to our cognitive capacities. I will study a landmark framework typically associated with the framing effect, Prospect Theory, and will argue that it has some serious shortcomings. Furthermore, I will attack some common characterisations of the framing effect and argue that focus should shift to the information conveyed by different frames.

## 1.1 Framing in Thought and Communication

In order to obtain a thorough understanding of the framing effect, some words have to be said about preferences first. A preference can be taken to be a ranking order on a set of events or objects. For instance, one can prefer reading a book over watching a movie, an immediate ban on environmental pollution over a piecemeal reduction, or working less and receiving a lower pay over working more and receiving a higher pay. Particularly interesting for decision theorists is what properties this preference ordering has and should have. Well known ordering conditions include consistency, asymmetry, connectedness and transitivity (Resnik 2008, pp.22-24).

In practice, however, many, if not all of these conditions are violated under certain circumstances. The framing effect is often seen as such a circumstance that may lead to the violation of these conditions, especially consistency. It turns out that framing can induce people to prefer $a$ over $b$ when one specific formulation is chosen, whereas it can induce people to prefer $b$ over $a$ when another formulation is used.

To see how this framing effect works, I will give a characterisation of preferences and how they arise. This issue has been studied extensively, especially by psychologists and economists. Commonly, a preference is taken to be an attitude towards an object or event. This attitude is the outcome of our evaluative beliefs (positive and negative) about that object or event (T. Nelson and Oxley 1999, p.1040). One widely accepted way of representing this idea is the so-called ‘expectancy value model’, in which evaluative beliefs are represented as ‘considerations' or 'dimensions' of an issue that together make up an attitude.[1] For example, one's attitude towards a policy plan might consist of one's evaluative beliefs about the impact on the economy and the environment, the estimated costs of the plan, the perceived reliability of the politician that proposed it, etc. The weighted sum of all these beliefs or dimensions determines one's overall attitude towards an object or event.

The notion of a frame can be characterised in terms of such attitudes and dimensions. In the literature, framing is used in two different ways. First of all, there is so-called 'equivalence' or 'valence' framing, of which some examples have been considered in the introduction.[2] For now, suffice it to say that this involves casting the same information differently, such as speaking about 95% employment or 5% unemployment. Here we are concerned with one dimension of a decision problem, presented in different ways. As it turns out, phrasing a dimension in positive terms often leads to a more favourable evaluation than phrasing it in negative terms, despite the fact that the dimension under consideration is one and the same. Valence framing is typically studied in psychology, economics and cognitive science .

A second way in which framing is used is in the sense of so-called 'value', 'issue' or 'topic' framing (Druckman 2011, pp.281-283). This type of framing is concerned with the relative importance that is attached to *different* dimensions of a problem. For example, we may consider both arguments about individual freedom and public safety in making a decision. Or we may attach decisive value to the economic dimensions of a problem, thereby adopting an 'economic' frame of thought. Contrary to valence framing, we now have several distinct evaluative beliefs that are treated differently: one dimension is taken to be more important than another. Topic framing is typically studied in political science, communication studies and public relations.

So far, I have focused on framing in thought and the way in which attitudes are formed from relevant considerations. However, what determines which considerations we take into account and what weight we attach to them? Presumably, various factors such as prior experiences, cultural background, social environment and ongoing world events play a role in this (ibid., pp.283-284). One factor, however, is particularly significant for the present purposes of this thesis: communication with others, be it relatives, public authorities or the media, as this is arguably the most

[1] See for instance Feather (1982).

[2] In psychology, the term 'valence' refers to the "intrinsic attractiveness or aversiveness" of events, objects, situations or descriptions (Frijda 1986, p.207). A sentence can be said have 'positive valence' or 'negative valence', depending on whether its appeal is 'attractive' or 'aversive'.

important source of information we have for evaluating our preferences.

Not surprisingly, framing is not merely confined to thought, but plays a role in communication as well. Speakers can engage in 'valence framing' by stressing the positive side of an issue (e.g., employment) rather than the negative side. Or they can engage in topic framing, by stressing one dimension (e.g., the environmental impact) rather than another.

Therefore, frames can be said to lead a 'double life'. As political scientists Kinder and Sanders (1996) put it: "frames are interpretative structures embedded in [...] discourse. [...] At the same time, frames also live inside the mind; they are cognitive structures that help individual[s] make sense of the issues that animate [...] life" (ibid., p.164).

A framing effect, in its broadest sense, can be said to arise when a frame in communication affects an individual's frame in thought (Druckman 2011, p.282). At the end of this chapter, it will become clear what exactly this means (section 1.4.2). For now, I will confine my analysis to (equi)valence framing. However, one central claim of this thesis is that topic framing can be analysed in the same way as valence framing, and that one can question whether it is feasible to draw a clear distinction between valence framing and topic framing.[3]

## 1.2 Prospect Theory

In 1981, Amos Tversky and Daniel Kahneman published their now-famous article *The Framing of Decisions and the Psychology of Choice* in which they describe the results of their 'Asian disease' experiment. They presented a scenario to a group of students from Stanford University and the University of British Columbia in which the outbreak of an unusual Asian disease is imminent. The disease is expected to kill 600 people when no action is taken. Luckily, the U.S. government can cushion the impact of the disease by adopting one of the following two programmes (Tversky and Kahneman 1981, p.453):

> **Programme A**: If programme A is adopted, 200 people will be saved.
> **Programme B**: If programme B is adopted, there is 1/3 probability that 600 people will be saved and 2/3 probability that no people will be saved.

When asked which programme they prefer, 72% of the respondents prefer programme A over programme B, whereas 28% prefer programme B over programme

[3] See section 1.4.2.

A. Despite the fact the both programmes yield the same expected value ($1 \times 200 = \frac{1}{3} \times 600$), the majority of people prefers the certain rescue of 200 lives over the uncertain rescue of 600 lives. This is in line with the widely observed fact that human behaviour is risk averse.[4]

A second group of respondents is presented the same scenario but now the programmes are as follows:

> **Programme C**: If programme C is adopted, 400 people will die.
> **Programme D**: If programme B is adopted, there is $\frac{1}{3}$ probability that nobody will die and $\frac{2}{3}$ probability that 600 people will die.

Surprisingly, when asked which programme they prefer, 78% of the participants of this second group prefers programme D over programme C, even though C is 'effectively' identical to A and D to B (Tversky and Kahneman 1981, p.453). The only difference, it seems, is the way in which the programmes are presented or formulated. This experiment, in which different formulations of the same event or fact result in different preferences or actions, is the classical example of valence framing effects.

What exactly is going on? How can one account for the differences in preferences? The Asian disease experiment has two important features. First of all, uncertainty plays a big role. Both programme B and D have uncertain outcomes and thereby stand in contrast to the certain consequences of A and C. Secondly, the experiment can be interpreted in terms of gains and losses relative to some 'given' outcome. Programme B can be interpreted as putting at stake the sure gain of programme A (200 survivors), whereas programme D can be interpreted as a chance to reverse the sure loss of programme C (400 deaths). The choice between A and B therefore takes place in a 'gain frame', whereas the choice between C and D takes place in a 'loss frame'.

Tversky and Kahneman use these features in what they call Prospect Theory (PT) to explain the framing effect (Kahneman and Tversky 1979). PT is a theory to account for individual decision making under risk and was developed as a reaction to standard expected utility theory, which dominated decision theory for a long time. One problem with this expected utility theory is that it cannot account for risk-averse behaviour. For example, it turns out that people prefer to have \$2.400,- with certainty, rather than a 33% chance of winning \$2.500,-, a 66% chance of winning \$2.400,- and a 1% chance of winning nothing, even though this latter option has

[4] Cf. Kahneman and Tversky (1984).

a higher expected utility ($2.409,- vs. $2.400,-) (Kahneman and Tversky 1979, p.265).

According to Prospect Theory, we make use of various ‘heuristics and biases’ in decision making. These are cognitive aids that “reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations” (Tversky and Kahneman 1974, p.1124). A downside of this, however, is that some information may be lost or misrepresented in this process, thereby leading to “severe and systematic errors” (ibid.).

By incorporating these heuristics and biases in the model, Prospect Theory is able to accommodate the observed behaviour such as in the example above. Tversky and Kahneman distinguish two phases in the choice process: “an early phase of editing and a subsequent phase of evaluation” (Kahneman and Tversky 1979, p.274). In the editing phase, the different choice options are organised, reformulated and simplified. Various factors are at work here. Most importantly, according to Tversky and Kahneman, we tend to rephrase outcomes in terms of gains and losses relative to some reference point, rather than as final states of wealth. Apart from that, there are many other changes we make to the choices. For example, we tend to simplify them, such that a 49% chance to win $101,- is recorded as a 50% chance to win $100,-. After this editing has taken place, the decision maker chooses the prospect of the highest value, but because the perceived or edited prospects may differ from the original or actual ones, ‘anomalies’ in choice may occur (ibid., p.275).

The tendency for risk-averse behaviour may be explained in terms of reference points. Tversky and Kahneman argue that, in evaluating the different choices we have, we make use of a so-called value function that has two arguments: a reference point and the magnitude of the change (ibid., p.277). This value function, at least for many of our considerations, is likely to be concave above the reference point. That is, the difference between a $100,- gain and a $200,- gain appears to us to be larger than the difference between a $1.100,- gain and a $1.200,- gain, just as the difference between a temperature change of 3° C and 6° C is easier to notice than a temperature change of 33° C and 36° C (ibid., p.278).

The converse holds for losses. Below the reference point, the value function is likely to be convex. This means that when it comes to losses, there is a tendency for risk-seeking behaviour. For instance, whereas people tend to prefer a sure gain of $240,- over a 25% chance to gain $1000,- and a 75% chance to gain nothing, they prefer a 75% chance to loose $1000,- over a sure loss of $750,- (Tversky and

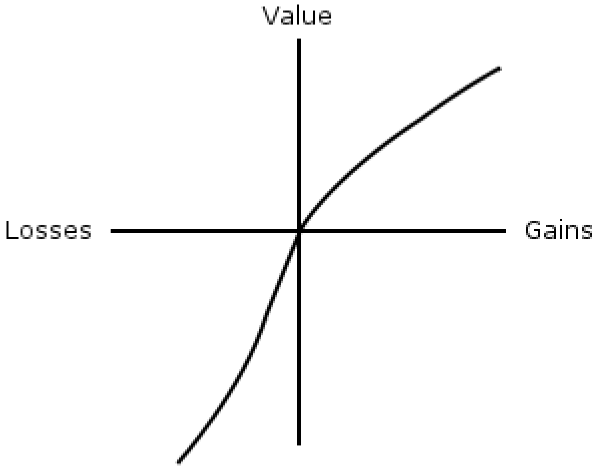

Figure 1: An example of an S-shaped value function.

Kahneman 1986, S255). As a result, the value function is commonly S-shaped (see figure 1).

In both gain and loss cases, the marginal value is expected to decrease with the magnitude of the gain or loss.[5] However, studies have shown that for most test subjects, the convex part is considerably steeper than the concave part. That is, losses loom larger than gains (e.g., Halter and Dean 1971, Barnes and Reinmuth 1976).

A famous example of this S-shaped value function is the so-called ‘favourite-longshot bias’. It turns out that in horse racing events, bets on long shots are most popular on the last race of the day (cf. Sobel and Travis Raines 2003).[6] A popular explanation for this phenomenon is that bettors do not consider individual bets in isolation, but rather take into account their gains or losses of the day. Suppose that someone has spent a day at the race track and has already lost $200,-. At the beginning of the day, the bettor may perceive a bet on a long shot as, say, a very small chance of winning $200,- and a big chance of losing $10,-. Given that her current return is $0,-, the potential loss is assessed as the difference between $0,-

[5] Note that deviations may occur due to special circumstances on preferences. Prospect Theory does leave room for this. For example, it seems natural to assume that a person’s aversion to losses increases sharply if the losses become so big that the person would be forced to sell their house. See Kahneman and Tversky (1979, pp.278-279).

[6] In this context, a long shot is a horse that has a very small chance of winning and, as a result, carries long odds.

and -$10,-. At the end of the day, however, the bettor may take her present loss of $200,- as her point of reference. As a result, the long shot bet may now be perceived as a very small chance to break even and a big chance of losing $210,-. The potential loss is now assessed as the difference between -$200,- and -$210,-. Because the value function is likely to be convex for losses, the potential loss of the bet at the beginning of the day looms larger than the loss of the same bet at the end.

Apart from value functions, Tvsersky and Kahneman also introduce weighting functions. That is, a subject attaches a certain decision weight to each perceived outcome of a choice. This decision weight does not only take into account the perceived likelihood of a specific outcome, but also the impact of an outcome on the desirability of prospects (Kahneman and Tversky 1979, p.280). As a result, decision weights behave differently from probabilities, and all weights together need not add up to 1. Nevertheless, the corresponding weighting function is expect to be positively related to the probability of an outcome. That is, we tend to attach greater weight to more probable outcomes. However, there is room for deviations. For example, Tversky and Kahneman suppose that very low probabilities are generally over-weighted (ibid., p.281). Due to this and other factors, the weighting function is expected to be non-linear (see figure 2).

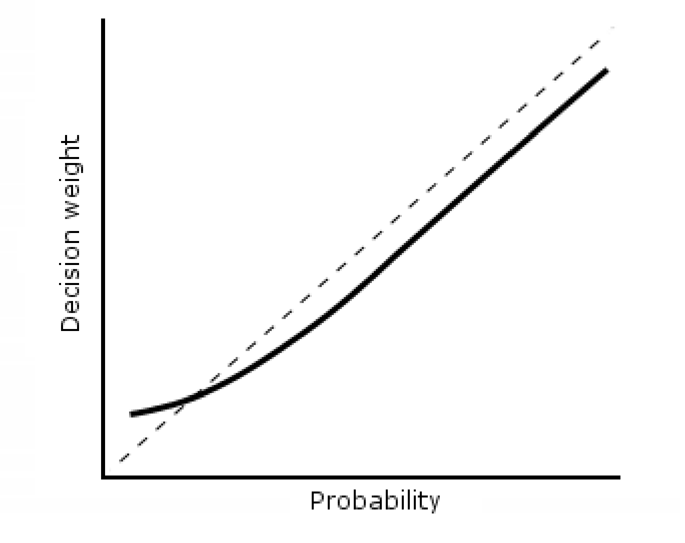

Figure 2: An example of a weighting function.

By looking at the relation between the perceived value of a prospect and the weight that is attached to it, Prospect Theory is able to account for various observed attitudes towards risk. For instance, in the '\$2.400,- example' presented above, PT predicts that the difference in perceived value of winning \$2.400,- or \$2.500,- is presumably small, whereas the 1% of winning nothing is presumably over-weighted. If the extent of these effects is large enough, the 'edited' prospect of the \$2.500,- / \$2.400,- / \$0,- gamble may receive a lower evaluation than the prospect of receiving \$2.400,- with certainty.

For two decades, Prospect Theory has been dominant in the analysis of framing effects. In what way can PT account for these effects? Recall that Tversky and Kahneman argue that our use of various heuristics and biases in decision making can lead to systematic errors or anomalies. The framing effect is an example of such an anomaly.

According to Tversky and Kahneman, the most important feature of framing is that it emphasises one reference point rather than another (Tversky and Kahneman 1981, p.456). As the horse track example shows, this can result in differences in behaviour. That is, one can assess a single bet individually or from the point of view of one's current return, and this influences the way in which the potential gains and losses of the bet are evaluated. Something similar may be going on in the Asian disease experiment. Programme A and B are phrased in terms of gains, and hence it can be argued that the point of reference (the 'default' option) is a situation in which all people die. Programme C and D, on the other hand, are cast in terms of losses, and hence the point of reference is a situation in which no one dies. As a result, programme A and B are assessed on the concave gain part of the value function, whereas C and D are assessed on the convex loss part of the value function. This can explain why the programmes are evaluated differently, despite their (apparent) equivalence.

Related to this explanation of framing in terms of reference points is an observed pattern in human behaviour pertaining to certainty. When it comes to gains, it turns out that people tend to have a strong preference for certain outcomes rather than merely probable ones. When it comes to losses, the opposite is the case. For example, people tend to value a reduction of the probability of a harm from 1% to zero (much) higher than a reduction of the same harm from 2% to 1%, despite the fact that both reductions are similar in magnitude (ibid.). PT can accommodate this tendency by means of the weighting function. Since weights differ from probabilities, the 'certainty effect' will result in a weighting function that rises sharply when the probability of a choice option reaches 1.

This tendency is exploited in framing. PT predicts that the certainty of programme A is considered an asset, because it is presented as a certain gain, while the certainty in programme D is considered a disadvantage, because it is presented as a certain loss.

Many, if not all actions or events can be framed in either conditional or unconditional form. Tversky and Kahneman provide the following two examples:

> "[A]n insurance policy that covers fire but not flood could be evaluated either as full protection against the specific risk of fire or as a reduction in the overall probability of property loss. [...] [A study has shown] that a hypothetical vaccine which reduces the probability of contracting a disease from .20 to .10 is less attractive if it is described as effective in half the cases than if it is presented as fully effective against one of two (exclusive and equiprobable) virus strains that produce identical symptoms." (Tversky and Kahneman 1981, p.456)

By casting an action or event in a different way, one is able to narrow down or broaden the context in which this action or event is embedded. In this way, one is able to present a seeming probability as a certainty and vice versa, and thereby to exploit the difference in weight we attach to these outcomes. Hence, as said above, our use of 'heuristics and biases' may lead to systematic errors.

## 1.3 The Diversity of Framing

In a nutshell, Tversky and Kahneman explain the framing effect as follows. PT accommodates various asymmetries in the way in which we perceive and evaluate (probabilistic) events. By casting an event or action in a different way, one is able to exploit these asymmetries by shifting the reference point of the decision maker. This may lead to different evaluations of one and the same event, and thus result in a difference in preferences. In the past decades, the Asian disease experiment has been replicated, adapted and criticised. And as it turns out, the framing effect is a much more diverse phenomenon than the experiment suggests. As a result, Prospect Theory leaves many issues unanswered.

One important feature of the Asian disease experiment *and* the explanation offered by PT is that risk plays an important role. Both groups of respondents have to choose between a riskless prospect and a risky, all-or-nothing prospect. In a meta-analysis Levin, Schneider and Gaeth (1998) argue that this complicates the interpretation of the experiment's results. The presence of risk in the choice options makes it more difficult "to extract what influence the frame, as opposed to the risk, is having on information processing" (ibid., p.157). They distinguish three types of valence framing that remain undifferentiated in Tversky and Kahneman (1981). In a comparable meta-study, Kühberger (1998) makes a similar distinction.

### 1.3.1 Risky Choice Framing and Attribute Framing

First of all, there is 'risky choice framing', of which the Asian disease experiment is the classic example. It involves a choice between a riskless option and a risky option. Several experiments of this type of framing show a so-called 'choice reversal'. That is, the majority of subjects who are presented a positively framed problem prefer the option with a certain outcome rather than the risky one, while the majority of subjects who are presented the negatively framed problem prefer the risky option rather than the certain one.

As seen above, PT incorporates the widely observed tendency for risk-averse behaviour in gain situations and risk-seeking behaviour in loss situations. In this way, the framing effect can be accommodated. However, the framing effect has also been observed in settings in which no risk (or other potentially distorting factors) are involved. This type of framing has been named 'attribute framing'. Here, the different choice options are identical, except that one single attribute is subject to framing manipulation.

A few notable examples of attribute framing include an experiment by Davis and Bobko (1986) in which people had to value an employability development programme that "has placed 39.9% of all participants in either part-time or full-time jobs" while another group had to value a programme that "has failed to place 60.1% of all participants in either part-time or full-time jobs" (ibid., pp.130-133). These programmes are identical, but one attribute (the programme's result) has been either cast in terms of success rate or failure rate. The experiment shows that the former (success rate) is valued considerably higher than the latter (failure rate). Another experiment, by Wilson, Kaplan and Schneiderman (1987), involves a pregnant woman who carried haemophilia. 42% of the subjects feel that the baby should be aborted when confronted with the thesis that the baby had a 50% chance of being infected with haemophilia as well, whereas only 26% feel that the baby should be aborted when they are told that there was a 50% chance that the baby was 'clean' (ibid., p.55). Yet another experiment, by Levin and Gaeth (1988), shows that people evaluate a hamburger that is presented as '75% lean ground beef' more favourably than one that is presented as '25% fat ground beef' (Levin and Gaeth 1988, pp.375-376).

Nearly all studies on this type of framing show a 'valence-consistent shift', i.e., a tendency to evaluate positively framed attributes more favourably than negatively framed ones (Levin, Schneider and Gaeth 1998, p.160). How can this effect be explained using PT? This is not straightforward. There are two key differences between attribute framing and risky choice framing. First of all, in the former the decision maker does not face a set of choices, but is presented either with a positive or a negative frame of one and the same event. Secondly, in attribute framing no risk is involved. The decision maker is dealing here with one single, certain piece of information (e.g., the outcome of an employability programme) and evaluates this information differently depending on the frame that is used. As a result, "a direct prospect theory explanation of attribute framing results is not feasible because the theory is designed to address changes in preference for options varying in riskiness when each of a set of options is framed; it is not designed to address subtleties in evaluations of individual objects or events" (ibid., p.166).

One may argue that, just like PT, attribute framing leads to different reference points. However, this creates another problem. PT is able to explain how preferences are formed and evaluated *within* a certain frame, but not *across* frames. That is, PT can explain that in a gain frame we tend to prefer programme A over programme B, whereas in a loss frame we prefer D over C. Yet it does not say anything about the relative value we attach to programme A in the gain frame, as compared to its counterpart C in the loss frame. And precisely this is what is needed if one explains attribute framing in terms of different reference points. A 39.9% success rate may (supposedly) be evaluated from a reference point in which no success was expected. We thus end up with a value function with 0% as the origin. A 60.1% failure rate, on the other hand, (supposedly) gives rise to a different value function, with 100% as the origin. In the first case, we are thus operating in the gain area, and in the second in the loss area, but of different value functions. How can these two value functions be compared?

Tversky and Kahneman do not provide an answer to this question. In fact, they do not consider attribute framing at all within their model. One possible answer would be to point at the differences in slope of gain and loss frames. According to PT, the loss part is likely to be steeper than the gain part, but Tversky and Kahneman do not use this in explaining framing effect. One can argue that precisely this can explain why the same information processed as a loss results in less favourable evaluation than when processed as a gain.

It is questionable, however, whether this is really what is going on in attribute framing. Furthermore, one can question whether the supposed difference in reference points between gain and loss frame is warranted, and what this reference point should be (is our reference point in a gain frame really a 0% success rate?).

A simpler explanation, offered by Levin, Schneider and Gaeth (1998), is that positively labelled information evokes favourable associations in our brain, whereas negatively labelled information evokes unfavourable associations. Which one is relatively stronger does not matter. Rather, this effect on our associative memory affects the way in which a certain evaluation dimension is perceived, and thereby changes our subjective scale values. This is the cause of valence-consistent shifts (ibid., p.164).[7] I will come back to this idea in chapter 2.

### 1.3.2 Goal Framing

The final type of framing is 'goal framing'. Goal framing is taken to influence implicit goals adopted by an individual by focusing on the issue's potential to provide a benefit or to prevent a loss (ibid., p.167). One example is about a campaign to promote breast cancer prevention by encouraging women to engage in 'breast self-examination' (BSE). It turns out that women are more inclined to do so when they are told that 'research has shown that women who do not do BSE have a decreased chance of finding a tumour in the early, more treatable stages of the disease' rather than 'research has shown that women who do BSE have an increased chance of finding a tumour in the early, more treatable stages of the disease' (Meyerowitz and Chaiken 1987, p.504).

The most important difference with attribute framing and risky choice framing is that in this case, both positive and negative frames promote the same act. Contrary to risky choice framing and attribute framing, it is not the case that the positive frame focuses on something desirable and the negative frame on something undesirable. We do not have to decide whether we like a 75% lean ground beef' or a '25% fat' hamburger. Rather, BSE is presented as a good thing in both frames. Nevertheless, "a pamphlet stressing the negative aspects of not doing BSE [has] a greater persuasive impact than a pamphlet stressing the positive aspects of doing BSE" (ibid., p.507).

Other examples of goal framing include the preference of avoiding a 'credit

[7] See also Levin, Johnson et al. (1986).

card surcharge' rather than forgoing a 'cash discount' in the case of price differences between cash and credit card purchases (Ganzach and Karsahi 1995); the greater willingness to use mouth wash when presented with photos of 'bad' mouths rather than 'good' mouths (Homer and Yoon 1992); and the greater intentions to eat breakfast more often when presented with a booklet stressing the negative impact of not eating breakfast as opposed to a booklet stressing the positive impact of eating breakfast (Tykocinski, Higgins and Chaiken 1994). In all these cases, the loss frame has a greater impact than the gain frame. That is, people are less willing to accept losses than to forgo gains. Hence, goal framing "influences how persuasive the message will be" (Levin, Schneider and Gaeth 1998, p.174).

Goal framing is a type of framing that has received relatively little attention. Researchers typically use Prospect Theory to explain its effect, by translating the frames into implicit risks that subjects are trying to seek or avoid (ibid., p.176). For example, Meyerowitz and Chaiken (1987) assume that women perceive performing breast self-examination as riskier behaviour than not performing breast self-examination. They argue that BSE, even though beneficial in the long run, involves the short-run risk of detecting breast cancer. They write: "Deciding to perform BSE requires that [women] risk aversive consequences in the present (e.g., finding a lump, experiencing anxiety) in hopes of enhancing future outcomes (e.g., living a longer life)" (ibid., p.501). These longer term considerations are likely to be less influential in determining behaviour, because they are temporally remote. As a result, BSE is perceived as'risky' behaviour.

In the positive frame, the reference point (presumably) adopted by women is that they are healthy. Hence, they face the choice between the status quo (no BSE) or a 'gamble' (BSE) that may result in long run health benefits but at the expense of potentially finding a lump. As argued above, this short-run loss looms larger than the long-term gain. Using PT, one can explain that most women go for the risk-averse option (no BSE). A negative frame, on the other hand, may cause a shift from a 'positive' reference point (healthy) to a 'doubting' reference point: without examination, we cannot be sure that no lumps are present. Hence, we are now in the 'loss domain' and, according to PT, risk-seeking behaviour is to be expected here (performing BSE) (ibid.).

Even though some studies show that the fear of finding a lump is indeed an important reason for women not to perform BSE,[8] a similar case can be made to argue that not performing BSE is the most risky option. Furthermore, it can be

[8] Cf. Mahoney (1977) and Turnbull (1978).

questioned whether all instances of goal framing can be translated into terms of implicit risks. For instance, can one say that a cash discount is taken to involve more risk than a credit card surcharge? Or that eating breakfast more regularly is riskier than not eating breakfast more regularly? This seems rather strange. As a result, this 'translating' of goal frame experiments into implicit risky choice ones in order for them to fit within Prospect Theory remains a cumbersome, contested and, at times, contrived practice. Therefore, it remains to be seen whether PT can provide a uniform account of goal framing.

Rather than using this translation practice, one can also again refer to the slope of the value function. As seen above, this slope is (presumably) steeper for losses than for gains. In the case of attribute framing, I have argued that this might open a way for explaining why positive frames are rated more favourably than negative ones. It may possibly also be used to explain why negative frames are more persuasive than positive ones. In both cases, however, much work has to be done to adapt PT such that it is able to give a robust account of various instances of these types of framing, and to be able to make comparisons across frames, be it in terms of favourability or persuasiveness.

## 1.4 The Uniformity of Framing

The above showed that the framing effect is a much more diverse phenomenon than the Asian disease experiment suggests, and that Prospect Theory has not been designed to accommodate this diversity. There seem to be three different effects of positive frames and negative frames on the decision-making process. In cases in which some options are risky, positive frames lead to risk-averse behaviour, whereas negative frames lead to risk-seeking behaviour. In cases in which the positive and negative sides of an option are stressed, the positive frame leads to more favourable evaluations than the negative. And in cases in which the positive and negative consequences of an act are stressed, the negative frame is more persuasive or provides a stronger incentive to act than the positive one.

Does this mean that one can speak of three qualitatively distinct phenomena here, as Levin and colleagues argue? Not necessarily. Even though it is unmistakeably the case that there are differences between the different observed framing effects, there are striking similarities as well. I will argue that the framing effect can be attributed to different *information* conveyed by the different frames. Furthermore, this 'informational' process can even be said to underlie topic framing as well.[9]

So far, valence framing has been qualified in terms of ‘equivalence’. Tversky and Kahneman call the Asian disease programmes A and C (and B and D) ‘effectively identical’, Levin and colleagues characterise valence framing as casting ‘the same critical information’ in a different light and Kühberger talks about framing as giving different descriptions of ‘logically equivalent choice situations’.[10] But in what sense are the different choice options really equivalent? This is a difficult question to answer.

### 1.4.1 Equivalence, Semantics and Information

The first aspect that makes it difficult to establish the equivalence of two statements is that the meaning of a sentence and the information it conveys do not solely depend on what is explicitly stated. One example, argued for by Sher and McKenzie (2006), is that the conversational behaviour of speakers exhibits several regularities of which the interlocutors are (implicitly) aware. This means that there is a link between the words and formulations chosen by the speaker and their background knowledge about the situation.

One such regularity is that it is often the case that, given two logically equivalent statements $A$ and $B$, speakers are more likely to utter $A$ when some background condition $C$ obtains than when it does not (ibid., p.469). In this way, the speaker may ‘leak’ certain information by choosing a specific formulation. This information may subsequently be ‘absorbed’ by the hearer, thereby inducing some specific behaviour.[11]

One straightforward example is passive-form sentences and active-form sentences. Despite the fact that they seem to describe the same fact, they convey “different information about the relative prominence of the logical subject and the logical object of the sentences (e.g., in “The man was kissed by the woman”, the man is intended and interpreted to be more prominent than in “The woman kissed the man”)” (ibid., p.470).[12] Hence, when the speaker chooses one formulation over the other, the emphasis put on the subjects in the sentence is different, and this may tell the hearer something about the speaker’s intentions, preferences or knowledge.

[9] Recall that topic framing is the type of framing in which one stresses a specific dimension of a decision problem and leaves out others (e.g., stress economic consequences rather than environmental consequences), instead of stressing one and the same dimension in either positive or negative terms.

[10] Cf. Tversky and Kahneman (1981, p.453), Levin, Schneider and Gaeth (1998, p.150), Kühberger (1998, p.23).

[11] Cf. Corner and Hahn (2010).

[12] Cf. Johnson-Laird (1968).

Another example is that speakers are more likely to use positive phrases when a situation has increased or improved relative to some reference point, or when the situation turns out to be better than expected. The opposite holds for negative phrases. For instance, people tend to say that the glass is 'half full' when it was empty before, and 'half empty' when it was full before (McKenzie and J. Nelson 2003, p.598). Thus, the specific formulation chosen by the speaker may reveal information about a positive or negative trend or development. Similarly, stressing that 'the employability programme has placed 39.1% of the participants' may not only be a simple statement about the result of the programme, but possibly also tells us something about the speaker's expectations or assessment of the situation.

The consequence of this is that two seemingly equivalent statements $A$ and $B$ that (when taken at face value) seem to license the same inferences, may be informationally non-equivalent nonetheless. The hearer can draw different inferences from the fact that the speaker uttered $A$ rather than $B$. The way in which a proposition is couched may convey information about relative prominence, causal agency, positive or negative trends etc. perceived by the speaker. As a result, by choosing one frame rather than another, the speaker may implicitly make a recommendation or reveal additional information to the listener. Sher and McKenzie write: "different perceptions (of relative prominence, causal agency, etc.) lead speakers to choose different sentence forms, and listeners are able to draw corresponding conclusions from the speaker's choice of sentence form." (Sher and McKenzie 2006, p.470).

Information leakage by implicit speaker regularities is not the only factor that makes it difficult to establish the equivalence of two statements. The semantic properties of specific formulations play a role as well. Geurts (2013) argues that there is a link between the meanings of words and expressions on the one hand, and our evaluative and justificatory practices on the other (ibid., p.3). Consider, for example, the following two sentences about a crashed airplane that was carrying 600 passengers:

(1) 200 people survived

(2) 400 people died

These sentences seem to describe exactly the same event, and therefore seem to be (descriptively) equivalent. Yet, as Geurts remarks, if an agent believes that 'it is good that 200 people survived', it seems contradictory to assume that this agent also believes that 'it is good that 400 people died'. This is puzzling: the qualification 'it is good' holds for (1), but does not hold for the (seemingly) equivalent (2).

Geurts explains this as follows. He claims that the assessment of the (typical use of the) word ‘good’ is influenced by the alternatives that are posed by a sentence. Consider, for instance, the following sentences:

(3) Fred kicked Barney

(4) *Fred* kicked Barney

(5) Fred kicked *Barney*

Sentence (3) can have different meanings, depending on which word in the sentence is emphasised. If one puts emphasis on ‘Fred’ (4), the sentence means something like ‘Fred was the one who kicked Barney’. The relevant alternatives associated with the sentence are, for example, ‘John kicked Barney’, ‘Peter kicked Barney’, ‘Henry kicked Barney’, etc. One can say that ‘it is good that (3)’ if one believes that the situation in which Fred kicked Barney is better than (many of) the situations in which someone else kicked Barney.

If one puts emphasis on ‘Barney’ (5), on the other hand, the alternatives associated with the sentence are, for example, ‘Fred kicked George’, ‘Fred kicked Scott’, ‘Fred kicked Bob’. In this case, one can say that ‘it is good that (3)’ if the situation in which Fred kicked Barney is better than (many of) the situations in which Fred kicked someone else. In general, Geurts argues that “the core meaning of ‘good’ is something like the following: ‘It’s good that $\phi$’ means that $\phi$ ranks sufficiently highly on the relevant qualitative scale which orders [the alternatives of $\phi$]” (Geurts 2013, p.10).[13]

A set of alternatives can be ordered quantitatively and qualitatively. For example, ‘Fred has $n + 1$ children’ can be taken to be quantitatively stronger than ‘Fred has $n$ children’. On the other hand, one can also rank a set of alternatives qualitatively, for example based on how ‘probable’ or ‘desirable’ each alternative is.

According to Geurts, part of the meaning of the word ‘good’ is that it complies with the ‘alignment assumption’, i.e., the assumption that quantitative and qualitative rankings coincide. For example, ‘Fred earns $\$n{+}1$,-’ is quantitatively ‘stronger’ than ‘Fred earns $\$n$,-’. In case of qualitative ‘goodness’, the alignment assumption prescribes that the former is therefore also taken to be qualitatively ‘better’ than the latter. Hence, if an agent believes that it is good that Fred earns $\$n$,-, the agent (probably) finds it (even) better if Fred earns $\$n + 1$,-.

[13] Note that $\phi$ is an alternative of itself as well.

In general, when it comes to 'good', the 'default' assumption is to assume that more (or stronger) is better. This explains why people are willing to apply 'it is good that' to (1), but not to (2): '$n + 1$ died' is quantitatively stronger than '$n$ people died', but the agent's qualitative ranking is likely to be exactly the opposite. Hence, quantitative and qualitative rankings do not 'align' and using the word 'good' would be infelicitous. For '$n$ people survived', on the other hand, quantitative and qualitative rankings do coincide.

Notice that it *is* felicitous to say 'it is good that less than $n$ people died'. But in this case, the alignment assumption is fulfilled. The sentence is likely to be ranked qualitatively higher than 'it is good that less than $n+1$ people died', but the former is also *quantitatively* stronger than the latter (e.g., 'less than 5' is stronger than 'less than 6').

As a result, if the agent accepts the alignment assumption for both (1) and (2), a contradiction follows. That is, it would follow that the agent believes that '$n + 1$ people survived' is better than '$n$ people survived' and that '$n + 1$ people died' is better than '$n$ people died'. This shows that despite the seeming 'descriptive equivalence' of (1) and (2), different semantic properties are at work and thus different qualifications apply to the different sentences. Furthermore, sentences do not only license inferences about the explicit facts they state, but also about "*counterfactual* states of affairs, i.e., about what might have been the case, and these [can] turn out to be inconsistent" (Geurts 2013, p.12).

This can also explain why a policy plan resulting in 90% employment is rated higher than a policy plan resulting in 10% unemployment. If one supposes that the agent prefers high employment (and thus low unemployment), she can hold the believe that 'it is positive that the plan results in 90% employment', and hence give a favourable rating. In the latter case, however, it is infelicitous to qualify the statement 'the plan results in 10% unemployment' as 'positive', and this may result in a lower rating. The same holds for Tversky and Kahneman's programme A and C: a semantic analysis provides her with reasons to believe that programme A will receive more favourable ratings than programme C.

It is important to note that Geurts does not simply want to reduce the framing problem to a semantic problem. His main point is that different descriptions give rise to different alternatives. This again influences our decisions, as they are not only based on explicitly stated facts, but also on the decision we could have made but did not (ibid., pp.13-14). In this way, semantic properties and decision making are related, but the one does not fully determine the other.

The upshot of both Sher and McKenzie's theory of speaker regularities and Geurt's focus on semantic properties is that the meaning of a sentence or the information it contains extends well beyond the facts it explicitly describes. In other words, logical equivalence does not imply informational equivalence. This idea can be traced back to Paul Grice's theory of implicature, i.e., the theory that our conversations and the exchange of information are characterised by various implicit rules and norms, which partly determine the meaning of our speech acts. By complying to or breaking such rules, one is able to 'speak between the lines'.[14]

This lays bare a second shortcoming of Prospect Theory. As seen above, PT explains the framing effect of the Asian disease experiment in terms of risks, choices and reference points. However, a simple semantic analysis predicts that programme A will receive more favourable ratings than programme C, and this can explain (a large part of) the framing effect. Such analysis does not use the notions of risks, choice and reference points. In fact, it does not even refer to programmes B and D and how the different programmes are related. This suggests that the key ingredients of PT's explanation of framing may not be as vital as Tversky and Kahneman believe. At the same time, it shows that two factors that are nearly entirely neglected by PT, the semantic properties of descriptions and their underlying conversational norms, may be much more important than Tversky and Kahneman have recognised.

### 1.4.2 Logical Equivalence and the Three Types of Framing

The above shows that one may have very good reasons to draw different conclusions from different descriptions. Even though the descriptions may allude to the same event, they can nonetheless convey very different information. This sheds a new light on the long held conviction that the framing effect induces 'irrational' behaviour. I will come back to this in the next chapter.

Does this mean that valence framing has wrongly been linked to the notion of equivalence? If one defines equivalence as *informational* equivalence, as Levin and colleagues do, the answer is clearly 'yes'. However, what about 'logical', 'objective' or 'descriptive' equivalence? Can it not be argued that two statements, when merely looking at what is explicitly stated, describe one and the same state of affairs?

One can ask whether the answer to this question really matters. What is the significance of knowing that a positive frame is equivalent to a negative frame seen from an isolated, 'objective' or 'factual' standpoint, if this standpoint is not the

[14] Cf. Grice (1975).

one adopted by the decision maker? In other words, what is the significance of logical equivalence if the agent assesses the decision problem through an embedded, subjective, context-dependent standpoint in which the frames are (informationally) non-equivalent?[15]

Furthermore, the alleged logical equivalence of different valence frames is not always straightforward. First of all, there is a difference between logical equivalence *per se* and logical equivalence in a specific context. For example, in the Asian disease experiment '200 lives saved' is logically equivalent to '400 lives lost', but only because the context specifies that there are 600 people in the domain under consideration. In some experiments, this context is not made explicit. This seems unproblematic in cases like '95% employment' vs. '5% unemployment', as it can be taken to be common knowledge that one entails the other (and vice versa). But does the same hold for lean meat vs. fat or success rate vs. failure rate? Is it common knowledge that a 75% lean meat hamburger consists of 25% fat? And does it immediately follow that a project team with a 40% failure rate has a 60% success rate? Or do some test subjects consider the possibility that some projects are neither (real) failures nor (real) successes? In both examples, the authors do not specify the context and thus assume that the equivalence is apparent. In my view, this is a premature conclusion.

Secondly, it can also be questioned whether many examples of goal framing are logically equivalent. The positive goal frame is of the form $p \rightarrow q$, whereas the negative goal frame is of the form $\neg p \rightarrow \neg q$. However, these statements are not logically equivalent. That is, generally it cannot be concluded from the goal framing examples that $p$ is a necessary and sufficient condition for $q$ (i.e., that $p \leftrightarrow q$ holds). As a result, the agent presented with $p \rightarrow q$ need not conclude that $\neg p \rightarrow \neg q$ holds as well (and vice versa).

For example, if one is told that 'performing breast self-examination (BSE) increases the chance of finding a tumour in the early stages of the disease', then BSE is presented as a sufficient condition for increasing this chance, but not a necessary one. This means that not performing BSE does not necessarily lead to a lower chance of finding a lump. Furthermore, if eating breakfast regularly leads to certain benefits (e.g., increased concentration, higher productivity), it does not follow that these benefits can only be attained by eating breakfast.

[15] In the next chapter, I will argue that this non-equivalence mainly has to do with the *accessibility* of the information conveyed by frames, not necessarily with the information itself. In this way, the logical equivalence of two frames can play a role in explaining certain empirical results of psychological experiments, such as the attenuation of the framing effect over time.

When looking more closely to logical equivalence in framing experiments, in many of them smaller or bigger differences between the frames can be found. One example in which the (descriptive) equivalence of frames is quite blatantly violated is the mouth wash advertising experiment briefly mentioned above (see p.22). The authors claim that "[c]areful attention was devoted to construct ad copies that were as equivalent as possible" (Homer and Yoon 1992, p.22). However, when looking at the ad copies, significant differences can be found. In the positive frame, subjects are told the following:

> "You can enjoy fresh breath if you practice good oral hygiene. Healthy gums, cavity protection, and a germ-free mouth are assurances of clean breath. So brush, floss, and visit your dentist regularly. Since many people don't clean their teeth regularly, the extra care of rinsing with mouthwash can be important for fresh breath and good oral hygiene." (ibid., p.33)

In the negative frame, on the other hand, subjects are presented with quite a different text:

> "Your mouth may be full of oral-germs that cause foul-smelling breath, plaque and gingivitis. And you don't want ginivitis (sic.). Gingivitis is a gum disease characterized by red, swollen gums. If left untreated, it can progress to periodontitis, which can result in tooth loss. It also causes bad breath. Three out of 4 adults have gingivitis." (ibid.)

Given the above, one can argue that neither goal framing nor some examples of attribute framing or risky choice framing are to be regarded as 'proper' instances of framing effects, because the logical equivalence is at stake. I agree with this, but claim that this is only a matter of definition. That is, 'fully' logical equivalent or not, the same underlying mechanism is at work in all types of framing presented here. This holds for topic framing as well. To put it bluntly, topic framing and valence framing are two ends of a continuum. This can be argued for as follows.

Recall that I have maintained that the notion of informational (non-)equivalence is much more relevant to decision making than the 'aloof' notion of logical equivalence. The informational equivalence of two frames is hard to establish and, due to semantic and conversational considerations, likely to fail. I therefore prefer to define the framing effect in terms of 'dimensions' of a decision problem.

In the first section of this chapter, it was argued that a decision problem can be represented as consisting of several dimensions. The decision whether to build a new road has an economic dimension, an environmental dimension etc. I have characterised both valence framing *and* topic framing as the situation in which a frame in communication affects an individual's frame in thought: a situation in which the highlighting of a particular dimension *rather than another*, or the highlighting of a dimension in a particular way *rather than another* (frame in communication) influences the dimensions and their respective weight that together make up our attitude towards an issue (frame in thought).

In other words, the framing effect occurs when partial information about a dimension or set of dimensions (due to the emphasis of the speaker) leads to a different decision by the decision maker than if this emphasis were different. This emphasis can either be emphasis on a specific dimension (e.g., environmental impact) or emphasis on a specific side of a dimension (i.e., positive or negative). Both cases lead to different information than if other dimensions or formulations are emphasised.

It can be remarked that, stated this way, framing is a very broad phenomenon. This is chiefly true because topic framing is included in the definition. Topic framing can be defined as the situation in which stressing one (part of a) dimension rather than another influences the agent's attitude towards a problem. Since speakers usually stress only some information and leave out other, the 'frame in communication' condition is often satisfied. Furthermore, since decision makers usually are not fully informed about the dimension(s) at stake, it is likely that the emphasis put on this specific information indeed influences the decision maker's attitude towards the problem ('frame in thought' condition). Precisely this effectiveness can explain why framing is so ubiquitous, for example in politics and advertising. I will briefly return to this issue later.

In this setting, valence framing can be characterised as a situation in which a particular way of presenting one and the same dimension (or a set of dimensions) influences the agent's attitude towards a problem. 'Pure' valence framing is thus linked to a notion of descriptive (or logical) equivalence: the 'same' argument or description of the 'same' results is presented differently.

However, I have argued that this notion of equivalence is not the most important one. Some instances of framing deal with completely different dimensions (topic framing), some are about 'objectively' the same dimensions presented differently (valence framing), and some are about the same dimension only up to a certain degree (e.g., goal framing). In all these cases, the different frames convey different information, and this informational effect of the speaker's emphasis on the decision maker's attitude is central to the framing effect.

Therefore, topic and valence framing can be said to be the two ends of a continuum of one and the same mechanism. This is why the framing effect can be said to have a certain degree of uniformity. In section 3.2.2, the similarities and differences between the two types of framing will be investigated in detail using the formal model presented there.

The advantage of using these definitions is that attention has shifted to the notion of information. I will use this as a key ingredient in the next chapter to model and explain the framing effect.

## 1.5 Conclusion

In this chapter, I have provided an overview of the framing effect. I have shown that framing is a diverse phenomenon that comes in various forms and has been studied in various fields. Tversky and Kahneman's Prospect Theory, a formal model that has dominated the framing literature for more than two decades, has been presented and criticised. I have argued that it neglects the diversity of the framing effect, and that it only provides a proper explanation for one specific type, i.e., risky choice framing. Furthermore, it focuses on the notions of risk, choice and reference points, whereas these may not be the most important notions underlying the framing effect. Instead, I have suggested that the notion of information carried by a frame is vital, and that this may be influenced by the semantic properties of the different frames and by implicit conversational conventions. PT neglects these semantic and conversational dynamics.

Furthermore, the characterisation of frames in terms of (logical) equivalence, put forward by various authors, has been criticised. There is an important difference between informational equivalence and logical equivalence, and both types of equivalences are difficult to establish (especially the former). For classificatory purposes, one can say that valence framing is about stressing descriptively equivalent frames or dimensions, whereas topic framing is about stressing distinct dimensions. However, there is a whole gamut of framing effects in which the descriptive equivalence only holds up to some degree. This means that these instances of framing at the same time also stress, again up to some degree, different (parts of) dimensions of a decision problem. Hence, they are somewhere in between 'pure' valence framing and ‘pure’ topic framing.

More important than these ‘definitional’ issues is that in all cases (topic, valence or ‘hybrid’) a similar interplay between frames in communication and frames in thought can be said to underlie the observed framing effects. In what is to follow, I will investigate some alternatives to Prospect Theory, and ultimately provide a new model for the framing effect. In this model, differences in information conveyed by different frames are central and, in this way, a uniform representation of both topic framing and valence framing can be given.

# 2 The Underlying Decision Process of Framing

In this chapter, two alternatives to Prospect Theory (PT) will be investigated. These alternatives will provide some useful insights into the underlying decision process of framing. However, I will argue that just like PT, both approaches face some serious problems and fail to fully capture the framing effect. Afterwards, an extensive account of the underlying decision process of framing will be presented, which takes into account everything considered so far and in which the notion of (partial) information plays a key role. Finally, the consequences of this account for the (ir)rationality of the framing effect will be explored.

## 2.1 A Decision Model without Extensionality

Until recently, Prospect Theory has remained the only formal model for framing. Bourgeois-Gironde and Giraud (2009) argue that this is so because, as explained in the introduction, framing effects "have long been discarded as irrational and because formal models of decision theory have for a long time only cared about rational behavior" (ibid., p.386). In the last 15 years, this trend has shifted, and new models for framing have been developed. Still, up until today, there are only a handful of formal models of framing effects around.

### 2.1.1 DM/E

One such model is developed by Bourgeois-Gironde and Giraud (ibid.). This model, which will be called a 'decision model without extensionality (DM/E), draws on the alleged informational non-equivalence of different frames that describe the same event. It can be regarded as a formal model for the theory of implicit information leakage or speaker regularities, as explained in the previous chapter. Bourgeois and Giraud make use of the model developed by Richard Jeffrey in his *The Logic of Decision* (1983), but adapt it in such a way that the principle of extensionality no longer holds. In their model, logically equivalent statements no longer (necessarily) influence decisions in the same manner. This allows them to accommodate informational non-equivalence. I will give a brief sketch of the model to provide a (rough) idea how it works.[16]

In the Jeffrey model, preferences are represented by a binary relation $\succeq$ on a set of propositions $\mathcal{A}$. These propositions are taken to be elements of a Boolean algebra. For two propositions $a$ and $b \ \in \mathcal{A}$, $a \succeq b$ expresses that the decision maker prefers $a$ to be true rather than $b$. Furthermore, preferences are taken to correspond to (a version of) expected utility theory, since the following version of Bolker's theorem holds:

**Theorem 2.1** *Bolker's Theorem*
*There exists a function $U : \mathcal{A} \to \mathbb{R}$ and a probability measure $P : \mathcal{A} \to \{0, 1\}$ such that for all $a, b \in \mathcal{A}$:*[17]

(i) $U(a) \geq U(b) \Leftrightarrow a \succeq b$
(ii) $U(a) = U(a \ \wedge \ b) \, P(b \mid a) \, + \, U(a \ \wedge \ \neg b) \, P(\neg b \mid a)$, *with* $P(b \mid a) := \dfrac{P(a \ \wedge \ b)}{P(a)}$

This theorem shows that a utility function $U$ can be specified that behaves similar to (a version of) expected utility (condition ii) and that corresponds with $\succeq$ (condition i). According to Jeffrey, we make ('ratify') decisions based on an evaluation of the expected utility of the different options we have. Jeffrey writes: "A ratifiable decision is a decision to perform an act of maximum estimated desirability relative to the probability matrix the agent thinks he would have if he finally decided to perform that act" (Jeffrey 1983, p.16). In other words, decision making is maximisation of expected utility.

It is important to note that no distinction is made between acts or events and the different descriptions of these. A sentence is identified with its equivalence class,

[16] A few other models for decision making and the framing effect have been developed in which this theory of implicit information leakage plays a prominent role as well. See Giraud (2004) and Ahn and Ergin (2010). Furthermore, Bourgeois-Gironde and Giraud (2009) provide a second model, based on Ghirardato and Marinacci (2001). The specific (technical) details of neither of these models play an important role in this thesis, but it is useful to give the reader a rough idea of how this theory of implicit information leakage can be incorporated in a formal model. I have chosen to provide a brief sketch of the Jeffrey model as it presupposes the least amount of background knowledge.

[17] To ensure the existence of $U$, it is assumed that $\succeq$ satisfies the following conditions:

(1) $\succeq$ is complete and transitive

(2) Let $a, b$ be 'disjoint' if $a \ \wedge \ b = \perp$. For all disjoint $a, b \ \in \mathcal{A}$, $a \succ b \Rightarrow a \succ (a \vee b) \succ b$ and $a \sim b \Rightarrow a \sim (a \vee b) \sim b$

(3) For all pairwise disjoint $a, b, c \in \mathcal{A}$, if $a \sim b \not\sim c$ and $(a \vee c) \sim (b \vee c)$ then for all $d$ disjoint from $a$ and $b$, $(a \vee d) \sim (b \vee d)$

(4) For any monotone sequence $(a_n)$ in $\mathcal{A}$ such that $a = \bigvee a_n$ or $a = \bigwedge a_n$, if $b \succ a \succ c$, then there is an $N \in \mathbb{N}$ suc that $b \succ a_n \succ c$ for all $n \geq N$ (Bourgeois-Gironde and Giraud 2009, pp.388-389).

consisting of logically equivalent sentences that each express one and the same proposition. This means that Jeffrey implicitly neglects "any morphological differences between logically equivalent sentences" (Bourgeois-Gironde and Giraud 2009, p.389). As a result, the principle of extensionality, stating that the agent is indifferent between sentences that belong to the same equivalence class, is (implicitly) taken to hold.

As explained above, Bourgeois and Giraud argue that this principle is violated in framing situations. They use Jeffrey's model as a starting point and define a 'decision problem with framing' as a tuple $\langle \mathbb{E}, \mathbb{P}, \Phi, \succeq \rangle$. $\mathbb{P}$ is a set of propositions describing an event, $\mathbb{E}$ is a set of events, and $\succeq$ is a binary relation over $\mathbb{P}$.[18]

$\Phi$ is the set of 'frames', represented as a set of homomorphisms from $\mathbb{E}$ to $\mathbb{P}$. This is to be interpreted as follows: suppose that for a proposition $p$, frame $\phi$ and event $e$ it is the case that $p = \phi(e)$. Then $e$ is called the 'reference' of proposition $p$ relative to frame $\phi$ (ibid., p.392). By uttering a proposition, one thus refers to a certain event, couched in a certain frame.

Bourgeois and Giraud introduce the notion of 'good news' to bypass extensionality. 'Good news' is a piece of information that helps the agent to make the right decision. It is information one would like to have before making a decision, but which is not always available. For example, when choosing between two investment opportunities, we would like to know which opportunity has the biggest chance of being profitable. Similarly, when deciding which economic policy to adopt, we would like to know which policy will lead to the highest economic growth or the lowest unemployment. Good news is defined as follows:

**Definition 2.2** *Good News*[19]
*An event $k$ is a 'good news' for an event $e$ if:*

(i) $\phi(e \wedge k) \sim \phi'(e \wedge k)$, *for all* $\phi, \phi' \in \Phi$
(ii) $\phi(e \wedge \neg k) \sim \phi'(e \wedge \neg k)$, *for all* $\phi, \phi' \in \Phi$
(iii) $\phi(e \wedge k) \succ \phi(e \wedge \neg k)$, *for all* $\phi \in \Phi$

The first two conditions say that good news is equally beneficial to all frames when it occurs (or does not occur). The last condition is the most interesting one. It says that a situation in which $k$ is true is preferred over a situation in which $k$ is not. Hence, knowing that $k$ is true in a certain situation makes that situation more attractive.

[18] To be precise: $\succeq$ is a binary relation over $\mathbb{P} \backslash \{\bot\}$.

[19] From Bourgeois-Gironde and Giraud (2009, p.392).

Given that $\succeq$ satisfies the same conditions as in the Jeffrey model (see footnote 17), the following theorem now holds:

**Theorem 2.3** *Informational Non-Equivalence*[20]
*For each frame $\phi \in \Phi$, there is a probability measure $P_\phi$ on the set of events $\mathbb{E}$ such that for all frames $\phi, \phi'$, for all events $e$ and for all 'good newses' $k$ it holds that:*
$\phi(e) \succeq \phi'(e) \Leftrightarrow P_\phi(k \mid e) \geq P_{\phi'}(k \mid e)$

This theorem says that if one prefers one description of an event over another description of this same event, then the (perceived) probability that some good news $k$ happens given this first description is higher than the (perceived) probability that some good news $k$ happens given the second description. The converse holds as well. In this way, Bourgeois and Giraud are able to express that logically equivalent statements (describing one and the same event) may be informationally non-equivalent nonetheless, thereby leading to different utility values.

As a result, the principle of extensionality no longer holds. It is no longer the case that for all events $e, e'$ and frames $\phi, \phi', \psi, \psi'$:

$$\phi(e) \succeq \phi'(e') \Leftrightarrow \psi(e) \succeq \psi'(e') \qquad \textbf{(Extensionality)}$$

For example, it may be the case that $\phi(e) \succeq \phi'(e')$ because the perceived probability of $k$ (given $e$) is high in frame $\phi$. This results in a higher perceived utility of $e$. However, it does not automatically hold that $\psi(e) \succeq \psi'(e')$, as it may very well be that the perceived probability of $k$ given $e$ is low in frame $\psi$, thereby lowering the expected utility of $e$.

### 2.1.2 Assessment

This opens the door for the incorporation of the theory of implicit speaker regularities into Jeffrey's decision model. One can specify a mechanism that ensures that certain ways of speaking or describing events increase the perceived probability of good news, and use the model of Bourgeois and Giraud to explain that this influences the hearer's preferences. Bourgeois and Giraud themselves do not specify such mechanism. However, assuming that such a mechanism exists, one can explain attribute framing by assuming that the (perceived) chance of good news is higher in positive frames.

[20] See Bourgeois-Gironde and Giraud (2009, p.393).

For example, one can argue that a glass with 90% pure fruit juice seems more desirable than a glass with 10% added sugar if one assumes that positive aspects are stressed more often in cases that a product is healthy or tasty. In this way, the perceived probability of good news (i.e., that the juice is indeed healthy or tasty) is higher in the positive frame. Or one can argue that a policy plan resulting in 95% employment is more attractive than one resulting in 5% unemployment, if employment figures are more often used in situations of high economic growth.

For the other types of framing, however, this explanation is less straightforward. Take, for instance, the Asian disease experiment. One can argue that a medical programme that is effective is more often described in positive terms (e.g., amount of lives saved) rather than negative terms. However, in this case both programmes presented to the decision maker are phrased in the same terms. That is, the agent either has to choose between programme A and B (couched in terms of lives saved), or between C and D (couched in terms of lives lost). Hence, one would expect that the 'good news' effect is the same for the two choices.

So why do subjects prefer A over B, but D over C? Additional assumptions are required to explain the framing effect here, such as that good news has a larger impact in cases of certainty. However, whether this really is the case remains to be seen. Furthermore, this assumption is in conflict with the underlying notion of expected utility in the Bourgeois and Giraud model.

When looking at goal framing, such as frames promoting breast self-examination (BSE), they run into trouble as well. As seen above, we are more likely to perform an action when the negative rather than the positive consequences are stressed. How can this be explained with the model presented above? Again, it seems that additional assumptions are needed. For instance, it may be so that speakers tend to use negative frames more often in cases that doing nothing has serious consequences, thereby implicitly urging the hearer to act. However, whether this is true is again uncertain, and psychological research is needed to support this claim.

In sum, it can be concluded that a quick evaluation shows that the model presented above works fine when it comes to the 'cleanest' type of valence framing, i.e., attribute framing. For other types, in which potentially distorting factors such as risk play a role, however, the model cannot be straightforwardly applied.

Just like Prospect Theory, the reason for these troubles seems to be a matter of focus. While PT mainly focuses on the notions of risk, choice and reference points, thereby neglecting causes of informational non-equivalence such as semantic considerations and implicit speaker regularities, the model of Bourgeois and Giraud chiefly revolves around the notion of good news. This allows them to incorporate the implicit information conveyed by speaker regularities, but it leaves little room for other (less direct) factors that may play a role in the framing effect.

Rather, it seems that by using a specific frame, i.e., by casting events in a specific light, a variety of associations are triggered. Only some of these associations are due to implicit speaker regularities linked to specific formulations. Others are due to the descriptive content of the formulation, or to the semantic properties of the words used, and yet others due to the context in which these words are used. These associations can be regarded as the 'information' (in the broadest sense of the word) we are able to extract from what is presented to us.

For instance, as Geurts (2013) suggests, the use of specific words may cause specific other words (such as 'good' or 'bad') to become applicable and this may prime certain scales used for evaluation. As a result, positively framed events are likely to be evaluated along positive scales, whereas negatively framed events are likely to be evaluated along negative scales.[21] Another example, pertaining to context-dependent associations, is that the risk involved in programme B of the Asian disease experiment may elicit the association of 'putting at stake' the certain rescue of 200 lives of programme A, while the risky element of programme D may 'offer a chance' to revert the certain loss of 400 lives of programme C. Depending on the frame chosen, the choices one faces are put in a different context, and this subsequently influences how they are evaluated.

Hence, this interplay between formulations, associations and evaluative standards goes much further than just the process of good news conveyed by speaker regularities. This is only one aspect, just as the effects of risks, choices and reference points is. Various factors give rise to various associations, and thereby influence the decision process in various ways.

## 2.2 Framing as Path Dependence

In order to give a proper account of the framing effect, a model is needed that is able to accommodate this 'associative' process outlined above. In a recent paper, Gold and List (2004) have proposed a second alternative to PT. This approach represents the framing effect through the notion of 'path dependence'. Even though the model has several shortcomings, it makes use of the concept of decision paths that will turn out to be very fruitful later on.

[21] Note that a scale can be seen as an ordering among alternatives. Different frames induces different alternatives, and hence different scales.

### 2.2.1 Path Dependence

The notion of path dependence is often used in decision theory, the social sciences and economics to refer to the fact that future decisions can be (and often are) restrained by decisions made in the past. One example from economics is that, according to the received view, the (actual) inflation rate is partially determined by the expectations of investors about what future inflation will be. These expectations are again based on past experience. In this way, past inflation rates have a lasting influence on future inflation rates (Yared 1999).

Another example from political science is the following.[22] Suppose a government, consisting of three members, has to decide whether they will implement a new education project, a new health care project and a new defence project. Assume that the implementation of all three projects together is only allowed when taxes are increased, in order to avoid a budget deficit. This increase in taxes is unanimously rejected by all government members, and as a result only two out of three projects can be implemented. Furthermore, suppose that the opinions of the government members are as follows:

| | Education | Health Care | Defence | Increase Taxes |
|---|---|---|---|---|
| Member 1 | Accept | Accept | Reject | Reject |
| Member 2 | Accept | Reject | Accept | Reject |
| Member 3 | Reject | Accept | Accept | Reject |

Note that the opinions of all three members are consistent. Now consider the following situation:

In January, the proposal to increase taxes is considered and rejected unanimously by the government members. In February, the education proposal is considered and accepted, as two out of three members are in favour of the plan. In March, the health care proposal is considered and accepted, for the same reasons. Finally, in April the defence proposal is considered. Despite the fact that a majority of the government members is in favour of the proposal, acceptance is in conflict with the government's earlier commitments. As a result, the government must reject the defence proposal.

It is easy to see that if the proposals were considered in a different order, the government would have accepted different projects. Hence, one can say that the outcome is 'path dependent'.

[22] Example from List (2004).

### 2.2.2 LPD

Gold and List (2004) suggest that in situations in which the framing effect occurs, the notion of path dependence may play a role as well. They argue that the human decision-making process is a *sequential* process, and that information processed early on puts constraints on the processing of information later. They present a Logic of Path Dependence (LPD) to model this idea. This framework will serve as a starting point for the model I will present in chapter 3.

The process Gold and List have in mind is the following. When facing a decision problem, the decision maker has to make up her mind about a so-called 'target proposition', such as a preference expression of the form '$x$ is preferred to $y$' ($PREF(x, y)$). To do so, the decision maker makes use of a set of 'background propositions', containing both information about the decision problem and preference rules relevant to the target proposition. These background propositions make up the 'context' of the decision problem (ibid., p.259).

Gold and List use the classical framework of predicate logic to represent the target proposition and its context. Let $X$ be the set containing the target proposition and the context propositions. For each proposition $\phi \in X$, the decision maker has an 'initial disposition', a preliminary opinion about the truth of a proposition. This initial disposition is a counterfactual notion: "[the decision maker's] *initial disposition* on $\phi$ is the judgment (acceptance/non-acceptance) she *would* make on $\phi$ *if* she *were* to consider $\phi$ in isolation, with no reference to other propositions" (ibid.). This results in the following 'acceptance function':

**Definition 2.4** *Acceptance Function*
*An acceptance function is a function $\delta : X \to \{1, 0\}$, where $X$ is a set of propositions. For each $\phi \in X$, $\delta(\phi)$ represents the acceptance (1) or non-acceptance (0) of $\phi$.*

Furthermore, Gold and List define a so-called 'decision path'. This is "the order in which the agent considers the propositions in a sequential decision process" (ibid., p.260):

**Definition 2.5** *Decision Path*[23]
*A decision path is a function $\Omega : \{1, 2, \ldots, k\} \to X$, where $k$ is the number of propositions in $X$.*

The decision process on the target proposition can be characterised as follows. At each step $i$ along the decision path, the agent considers a proposition $\phi_i \in X$. If the

[23] From List (2004).

initial disposition of $\phi_i$ is consistent with the set of previously accepted propositions $\Phi$, $\phi_i$ is added to $\Phi$. If not, there is a conflict between $\phi_i$ and $\Phi$, which has to be resolved. Gold and List assume that the decision maker uses a modus ponens or a 'priority-to-the-past' rule to resolve tis conflict, in which previously accepted propositions and their logical consequences can overrule the initial disposition on $\phi_i$. This results in the following process:

**Definition 2.6** *Modus Ponens Decision Process*[24]
*Given a decision path $\Omega$ of length $n$, consider $\phi_1 := \Omega(1)$ in step 1, $\phi_2 := \Omega(2)$ in step 2..., $\phi_n := \Omega(n)$ in step $n$. Let $\Phi_t$ be the set of all propositions accepted up to and including t. $\Phi_t$ is defined by induction as follows:*

- *For t $= 0$, $\Phi_0 = \emptyset$*
- *For t $= k$, $\phi_k$ is considered. There are three cases:*
  - *(1) If $\Phi_{k-1}$ entails $\phi_k$, then $\Phi_k = \Phi_{k-1} \cup \{\phi_k\}$*
  - *(2) If $\Phi_{k-1}$ entails $\neg\phi_k$, then $\Phi_k = \Phi_{k-1} \cup \{\neg\phi_k\}$*
  - *(3) If $\Phi_{k-1}$ does not entail $\phi_k$ or $\neg\phi_k$, then*:

$$\begin{array}{ll} \Phi_k = \Phi_{k-1} \cup \{\phi_k\} & \textit{if } \delta(\phi_k) = 1 \\ \Phi_k = \Phi_{k-1} & \textit{if } \delta(\phi_k) = 0 \end{array}$$

The decision process ends either if the decision maker has made a decision on the target proposition, or if the final proposition of the decision path has been considered.

The Asian disease problem can now be represented as follows. Suppose that an agent has to decide on the target proposition $PREF(x, y)$ (i.e., which programme she prefers over the other). Furthermore, suppose the agent has the initial disposition to accept the following factual and normative propositions:[25]

(i) $SAVE(p_{a/c})$: Programme A / C saves some lives with certainty

(ii) $\neg SAVE(p_{b/d})$: Programme B / D involves the risk that no one will be saved

(iii) $DEATH(p_{a/c})$: Programme A / C entails the certain death of some people

(iv) $\neg DEATH(p_{b/d})$: Programme B / D offers the chance that no one will die

[24] Adapted from Gold and List (2004, p.261).

[25] Adapted from Gold and List (ibid., p.267). $p_{a/c}$ is to be read as $p_a$, referring to programme A, if the agent faces the choice between A and B; $p_{a/c}$ is to be read as $p_c$, referring to programme C if the agent faces the choice between C and D. The same holds for $p_{b/d}$ (mutatis mutandis).

(v) $SAVE(x) \ \wedge \ \neg SAVE(y) \rightarrow PREF(x, y)$: It is not worth taking the risk that no one will be saved

(vi) $DEATH(x) \wedge \neg DEATH(y) \rightarrow PREF(y, x)$: It is unacceptable that some people will die with certainty

It can be argued that the difference in presentation of the choice between A / B and the choice between C / D induces two different decision paths. The presentation of the former makes the propositions in terms of lives saved more focal ($i$, $ii$ and $v$), whereas the presentation of the latter makes the propositions in terms of deaths more focal ($iii$, $iv$ and $vi$). An agent presented with the first problem is therefore most likely to follow the path $\Omega$, with $\Omega(1) = SAVE(p_a)$, $\Omega(2) = \neg SAVE(p_b)$, $\Omega(3) = SAVE(x) \ \wedge \ \neg SAVE(y) \rightarrow PREF(x, y)$, $\Omega(4) = \ldots$, whereas an agent presented with the second problem is most likely to follow the path $\Psi$, with $\Psi(1) = DEATH(p_c)$, $\Psi(2) = \neg DEATH(p_d)$, $\Psi(3) = DEATH(x) \ \wedge \ \neg DEATH(y) \rightarrow PREF(y, x)$, $\Psi(4) = \ldots$

If one assumes that the set of accepted propositions $\Phi$ of both agents is deductively closed,[26] the first agent will reach a decision on the target proposition at $t = 3$ and conclude $PREF(p_a, p_b)$ (i.e., prefer programme A over B). The second agent will also reach a decision on the target proposition at $t = 3$, but will conclude $PREF(p_d, p_c)$ (i.e., prefer programme D over programme C). This is the case because for the first agent, $\Phi_3 = \{SAVE(p_a), \neg SAVE(p_b), SAVE(x) \ \wedge \ \neg SAVE(y) \rightarrow PREF(x, y)\}$, which entails $PREF(p_a, p_b)$, whereas for the second agent, $\Phi_3 = \{DEATH(p_c), \neg DEATH(p_d), DEATH(x) \ \wedge \ \neg DEATH(y) \rightarrow PREF(y, x)\}$, which entails $PREF(p_d, p_c)$. As seen, this is indeed what the majority of agents decides when facing the Asian disease decision problem.

With this model in mind, the link between framing and path dependence can be explored. One can say that a decision process is ‘path dependent’ with respect to a target proposition $\phi$ if there exist “at least two decision paths with mutually inconsistent outcomes on $\phi$” (Gold and List 2004, p.264). That is, there exist two paths $\Omega_1$ and $\Omega_2$ such that under one $\phi$ is accepted and under the other $\neg\phi$ is accepted. The framing effect can then be said to occur when two different presentations of a decision problem lead to a path-dependent decision process.

[26] That is, if the agent accepts all propositions in $\Phi$ and it is the case that $\Phi \models \psi$, then the agent also accepts $\psi$.

### 2.2.3 Assessment

In what situations can a path dependent decision process arise? Gold and List distinguish four conditions that, arguably, are satisfied by the initial dispositions of any agent that can be called rational. They argue that path dependence only arises if at least some of these conditions are violated. The four rationality conditions on $\delta$ are as follows (Gold and List 2004, p.262):

> **Completeness:** For all $\phi \in X$, $\delta(\phi) = 1$ or $\delta(\neg\phi) = 1$.
>
> **Weak Consistency:** For all $\phi \in X$, it is not the case that both $\delta(\phi) = 1$ and $\delta(\neg\phi) = 1$.
>
> **Strong Consistency:** The set $\{\phi \in X : \delta(\phi) = 1\}$ is logically consistent.
>
> **Deductive Closure:** Given a set of propositions $\Psi$, if $\Psi \models \phi$ and $d(\psi) = 1$ for all $\psi \in \Psi$, then $\delta(\phi) = 1$.

The completeness condition says that for all $\phi$ under consideration, the agent must have the initial disposition to either accept $\phi$ or accept its negation. The weak consistency condition says that the agent never has the initial disposition to accept both $\phi$ and its negation. The strong consistency condition says that all propositions the agent is willing to accept can be simultaneously true without leading to inconsistencies. And the deductive closure condition says that the agent has the initial disposition to accept those statements that logically follow from the ones already accepted.

A proposition $\phi$ can be said to be path dependent if the agent's dispositions are 'implicitly inconsistent'. This means that there are two sets of propositions, both accepted by the agent, such that the one entails $\phi$ and the other $\neg\phi$. This implicit inconsistency is a necessary and sufficient condition for path dependence. It is easy to see that for it to arise, the strong consistency condition has to be violated. That is, only if the accepted propositions cannot all be true simultaneously, is it possible that two opposing conclusions can be drawn from the set of accepted propositions.

Furthermore, if the agent's dispositions are complete and weakly consistent, then it follows that path dependence requires deductive closure to be violated as well (ibid., p.263). That is, if the agent does not (explicitly) accept both $\phi$ and $\neg\phi$ (weak consistency) while one path leads to $\phi$ and another to $\neg\phi$, it follows that one can deduce both conclusions from her set of accepted propositions. Since only one of these conclusions is accepted, the set of accepted propositions of an agent entails conclusions the agent herself does not accept. A consequence of this is that, for path dependence to arise, at least two rationality conditions have to be violated.

It is important to note that for the framing effect to occur, a path dependent decision process alone is not enough. Path dependence is a *logical* requirement for the framing effect, but there is an *empirical* requirement as well. That is, there have to exist "two ways of presenting the decision problem to the agent that, empirically, lead the agent to use these two decision paths" (Gold and List 2004, p.265). Hence, apart from the right logical conditions, there must also be an empirically feasible way of 'triggering' both paths for the framing effect to occur in practice. Gold and List do not specify what these empirical conditions are that induce the agent to use a specific path. I will (briefly) come back to this later.

As seen above, Gold and List's model succeeds in explaining the Asian disease experiment. But how about attribute framing and goal framing? Attribute framing seems to be unproblematic as well. One can argue that a glass of 90% pure fruit juice triggers the normative statements that 'pure fruit juice is healthy' and that 'one should choose a healthy option'. Hence, the agent is likely to end up appreciating the juice. A glass of juice with 10% added sugar, on the other hand, is likely to trigger the normative rules that 'sugar is unhealthy' and that 'one should avoid unhealthy food', thereby inducing the agent to dislike the juice.

The case of goal framing is less obvious. Here, the decision problem only provides us with a conditional fact, for instance, *if* one performs BSE, *then* one has a bigger chance of finding a lump at an early stage. So how does this induce us to perform BSE? And how can one explain that a negative frame is more persuasive? Gold and List do not provide an answer to these questions.

## 2.3 The Framing Effect as a Partial-Information Decision

So far, three formal models of the framing effect have been studied: Prospect Theory (PT), Bourgeois and Giraud's Decision Model without Extensionality (DM/E), and Gold and List's Logic of Path Dependence (LPD). I have argued that all these models, in their current form, can only partially account for or accommodate the various types of valence framing.

In the next chapter, I will endeavour to develop an improved model for the framing effect. However, before doing so, a careful representation of the underlying decision process of the framing effect must be given. The results of some experiments that have not been discussed so far will be helpful here. I will argue that these experiments provide some useful insights on the relation between our cognitive capacities and the framing effect.

### 2.3.1 The Elaboration Effect

The first interesting result is that the framing effect is observed to attenuate or even disappear when people put more effort in making their decision. Takemura (1994) shows that there is a significant link between elaboration time and the outcome of a decision process. He replicated the Asian disease experiment, but divided the testing subjects into different groups. These different groups were give different amounts of time to make their decision. While one group of subjects was asked to think about their choice for 10 seconds, the other group was asked to take 3 minutes before making a decision.

The results of the first group were more or less similar to the results Tversky and Kahneman observed: a significant amount of subjects preferred the riskless option in the positive frame and risky option in the negative frame. In the second group, however, no significant difference between the two frames was found.

To be precise, in group 1, 29 respondents picked the riskless option in the positive frame, and 12 picked the risky option. In the negative frame, these numbers were 13 and 28 respectively. Hence, a clear choice reversal can be observed. In group 2, however, 24 respondents picked the riskless option in the positive frame, and 17 picked the risky option. In the negative frame, these numbers were virtually the same, 23 vs. 18 (ibid., p.37).

What this shows is that the differences between positive and negative frames for the second group are much smaller. Furthermore, it is no longer the case that the positive frame is tied to risk-averse behaviour and the negative frame to risk-seeking behaviour. I will call this observed attenuation of the framing effect when elaboration is high the ‘elaboration effect’.

A similar pattern is observed when subjects are asked to provide a written justification for their decision: when subjects are asked to give a justification for their choice, the ratio between riskless and risky options was 21 vs. 24 in the positive frame, and 28 vs. 17 in the negative frame. This difference was below the significance threshold. For subjects that did not had to provide a justification, on the other hand, these numbers were 36 vs. 9 and 14 vs. 31 respectively, a clearly significant difference between positive and negative frames (ibid., p.36). Similar results are reported by Miller and Fagley (1991) and Almashat et al. (2008).

Another interesting result is that the framing effect tends to be less pronounced when people have to decide on personal issues or have to relate a decision to themselves. When asked to rate the likelihood that students cheat, subjects tend to give higher ratings when presented with the statement “65% of the students had cheated during their college career" than with the statement "35% of the students had never cheated". However, when asked to rate the likelihood that the subjects themselves would cheat, the outcomes were the same regardless of the frame the subject was presented with (Levin, Schnittjer and Thee 1988, p.521). This observation that the framing effect attenuates when subjects are asked to estimate their own performance has been confirmed by various other studies as well (see, e.g., Sniezek, Paese and Switzer 1990, Schneider 1995).

Furthermore, experiments show that the framing effect is less likely to occur when subjects are presented with rather unusual or extraordinary decision problems. For example, Beach et al. (1996) report no framing effect when subjects are asked to evaluate a toaster that is missing many important parts, whereas the framing effect did occur for toasters that are only missing some unimportant parts (ibid., pp.79-80). Similarly, Levin, Johnson et al. (1986) show that the framing effect is much weaker for gambles with extreme probabilities of winning or losing, than for gambles with more moderate probabilities.

For the frameworks discussed so far, these results are problematic or at least give rise to some difficult questions. Take, for example, the elaboration effect. For LPD, this poses a problem. It seems that neither the available decision time nor the requirement of justification changes anything about the decision path one follows. That is, a positive frame is still more likely to make positive facts and rules more focal, and hence put them more towards the beginning of the decision path. Thus, LPD would predict that subjects would follow the exact same decision process and so reach the same conclusion, regardless the degree of elaboration.

The elaboration effect is troublesome for PT and DM/E as well. Some additional assumptions are required to accommodate this effect. For example, it might be the case that the perceived good news fades over time. Or that the agent adopts a different reference point when elaboration is higher. However, it is unclear whether this is the case. Furthermore, such assumptions require justification, which is currently lacking.

Another aspect of the results of the framing experiments neglected so far is that testing subjects do not unanimously make the same choices. For example, in the Asian disease experiment, even though 72% of the subjects prefers programme A, a small part chooses programme B (28%). Similarly, even though a majority of the subjects does endorse a '39.9% successful placement' programme but does not support a '60.1% unsuccessful placement' one, in both cases there are (considerable) amounts of people who decide differently.

All three models that have been discussed fail to give a clear account of this lack of unanimity. From the perspective of LPD, one can point out that some people have different normative rules, and that therefore the factual propositions triggered by the presented frame lead to different conclusions. However, when looking at the (presumed) normative rules that are at work in the Asian disease experiment (such as 'one should not let people die with certainty'), these seem so uncontroversial (and natural) that it is highly unlikely that 28% of the subjects do not share them.

One can argue that PT and DM/E can account for the disagreement among subjects by saying that different people attach different (expected) utility values to different choices. However, this barely explains anything at all, for why and how these differences arise remains opaque. In what sense are the people that make different choices subject to the framing effect? Do they fail to perceive the 'good news' implicitly conveyed by a frame, and therefore attach lower expected utility to this frame? Or is their preference so strong that the framing effect is not enough to make them change their minds? Both models do not answer these questions.

### 2.3.2 The Underlying Decision Process

I think that, in all the cases presented above something else is going on. It seems that decision makers generally do not process all the available information before reaching a conclusion. Rather, they draw a 'tentative' conclusion based on the information that is close at hand. When explicitly asked to elaborate, when relating decisions to themselves, or when shaken up by the encounter of unusual situations, decision makers no longer rashly accept the first conclusion that springs to mind. Rather, they are inclined to take into account less readily available information. This may induce them to overrule their first conclusion and make a different decision.

One experiment that can be taken to support this claim shows that when agents are given full information at once, no framing effect is observed. For example, when the descriptions of the Asian disease programmes stress both the amounts of lives saved and lost, and hence both positive and negative information and associations are at hand, the results are more or less the same as group 2 in Takemura's elaboration experiments. Furthermore, there is no difference between frames in which the number of lives saved is put first and frames in which the amount of deaths is put first (Kühberger 1995, p.234).

A similar result is obtained by Sniderman and Theriault (2004) in the context of topic framing. They show that when test subjects are presented with two competing frames, such as a frame stressing the impact of a new law on individual freedom and a frame stressing the impact of this law on public safety, the framing effects mutually cancel out (Sniderman and Theriault 2004, pp.153-156).

Putting together the insights gained so far, I claim that the underlying decision process of the framing effect can be described as follows. As argued in section 2.1.2, the use of specific words or formulations in a specific context gives rise to various associations. Some of these associations are triggered by the 'objective' descriptive content of the proposition, others are grounded in semantic properties of the words used, are the result of patterns or regularities one is implicitly aware of and which are tied to specific formulations, or arise due to the context in which specific words are used.

I have called these associations the 'information' (in the broadest sense of the word) contained by a frame. What happens in framing is that some of these associations become more focal than others. That is, the use of a specific formulation triggers some associations right away. Other, less apparent information that can also be extracted from the decision problem requires more cognitive effort to be considered. Since the 'immediate' associations are often sufficient to reach a tentative conclusion, our decision process commonly ends before all the available information has been considered.

Some subjects, however, are not satisfied with this conclusion and can decide to take more information into account. This information is less focal and therefore requires more cognitive effort to process. Several factors can induce an agent to indeed make this effort. In this section, a few of them have been discussed: when one is asked to take more time (e.g., by providing a justification), when the problem gets personal, or when the mind is 'shaken up' by an unusual or extreme situation.

However, it stands to reason that other, less direct factors play a role as well, such as the subject's intelligence, mood, vigilance etc. This can also explain the lack of unanimity. While most subjects stop the decision process as soon as they have reached a first conclusion on the target proposition, some others proceed and take more information into account.

It is important to note that this process of considering information is the same for both topic framing and valence framing. Successful topic framing causes information about a certain *aspect* of a problem to be processed first, and thereby often leads to a conclusion largely based on information pertaining to this aspect (e.g., economic benefits). Valence framing causes *positive or negative* information related to a problem to be considered first, thereby often leading to a conclusion based on this positive or negative information. In both cases, the same underlying process of giving priority to the primed information is at work.

Hence, the notion of 'partial information' can be said to be at the heart of the framing effect. A frame greatly influences the order in which the information contained by a decision problem is processed (i.e., which associations are triggered immediately and which require more cognitive effort), and thereby influences the information used to reach a first (and often final) decision.

In a way, this idea of partial information is reminiscent of a central claim put forward by Daniel Kahneman in his best-selling book *Thinking, Fast and Slow* (2011). Kahneman argues here that human beings have two systems or modes for decision making. There is a 'hasty' mode, in which we use various heuristics and biases to quickly decide on an issue. As a consequence, some information may be lost or tarnished in this process. Secondly, there is a 'slow' mode, in which we take more time to carefully consider all available information. Prospect Theory is designed to model this first mode, while the second one much more resembles classical logic.

In the next chapter, I will shown that it need not be the case that human beings indeed use two separate systems for decision making. I will present one and the same model that can accommodate both partial and full information processing.

In sum, one can say that the underlying decision process of the framing effect has three major tenets; (1) the central role played by a variety of associations or information extracted from a decision problem (not merely confined to 'risk', 'reference points' or 'good news'); (2) the varying degrees of accessibility of this information, chiefly dependent on the frame that is used and the cognitive effort it takes to access (implicit) information; and (3) the resulting consequence that we often only take partial information into account when making decisions, as this is often enough to reach a (tentative) conclusion.

It is not hard to see that Gold and List's notion of path dependence fits well with these tenets (especially 2 and 3). Therefore, it will play a key role in the model presented in chapter 3. However, despite the fruitfulness of this notion, I have also argued that the model provided by Gold and List themselves is too static and narrow to provide a full account of all facets of the framing effect. As it turned out, the model faces problems in representing goal framing, and it cannot provide a proper explanation of the results discussed in the previous section, such as the elaboration effect. The model lacks a flexible mechanism for looking beyond our first, preliminary, conclusion and for considering less focal information.

Apart from this, there is another problematic aspect of LPD that has not been discussed so far: the alleged violation of the standards for rationality that Gold and List take to be necessary for the framing effect to occur. I will turn to that now.

## 2.4 The Rationality of Framing

As seen in the introduction, the framing effect has typically been interpreted as (yet another) sign of human irrational behaviour. This is echoed in LPD, as it directly links the occurrence of the framing effect to the violation of some rationality principles (such as deductive closure or the consistency of our beliefs). Similarly, in PT the framing effect is labelled as a 'systematic error', a violation of the invariance principle, caused by our system of heuristics and biases.

In this section, I will argue that this conclusion is unsatisfying, and that one should look at the framing effect from the purview of the decision maker as a bounded rational agent.

### 2.4.1 Rationality and Information

In the introduction, I have focused on the notion of description invariance. Many authors take this to be a cornerstone of rationality. However, as argued in chapter 1, the descriptive equivalence of two formulations does not mean that these formulations are equivalent in all respects. Logically equivalent statements may very well contain (very) different information, and this information may have varying degrees of accessibility.

As a consequence, the violation of 'objective' description invariance in the framing effect does not necessarily mean that human behaviour is irrational. One may have very good reasons to violate this principle. We (implicitly) notice various patterns and regularities in the (linguistic) behaviour of others. This allows us to 'read between the lines' and extract useful information from the fact that a specific formulation is used rather than another. This information is not contained in the 'objective descriptive content' of the uttered statement, but may well be very relevant for the decision one has to make.

I claim, therefore, that the logical properties (equivalence) or the descriptive content of a statement under consideration is not of primary importance in the decision-making process. There is some (potentially) more useful information, related to the specific formulation that is used and the context in which it is uttered. Various non-logical factors that can provide us with such useful information have been indicated, such as semantic properties of the formulation used, regularities we implicitly observe, contextual factors etc.

This does not mean that logical equivalence plays no role at all in the decision process. Rather, this information becomes available to us (or is processed by us) later on along the decision path. For example, if a doctor stresses the mortality rate of a medicine rather than its success rate, the first associations that spring to mind are related to this negative term 'mortality' and to the fact that the doctor used this specific formulation rather than another. This often provides us with enough clues to reach a conclusion we find acceptable.[27]

However, if we are not satisfied, some more distinct observations and associations come into view, such as the logical equivalence of the doctor's statement with a statement in terms of survival rate. This latter statement again induces various (more positive) associations. The processing of this information can lead to a more 'balanced' decision, as more information is taken into account. This explains why the framing effect is observed to attenuate when the degree of elaboration of the decision maker increases.

This does not mean that our initial decision is 'irrational'. Rather, it is based on partial information, but this need not have any (negative) normative implications. I will argue that the rationality of a 'partial-information decision' depends on whether one has good reasons of being satisfied with it. What exactly qualifies as a 'good reason' for satisfaction in this case? In other words, under what circumstances is it justified to rely on the most focal (and often 'non-logical') information, rather than full information? There are several candidates.

One of them is the time the decision maker has available. Sometimes we simply cannot consider all the information directly and less directly conveyed by a decision problem, but rather have to make up our minds within a limited time.

From an evolutionary perspective, it seems reasonable to assume that this ability to quickly draw conclusions is vital. However, this automatically means that we have to give priority to some information, rather than others. Given our limited time, the information we are able to process need better be as relevant as possible.

[27] The question whether our cognitive processes of generating associations does indeed operate in such way will not be discussed further in this thesis. There are some theories and data that are in support of this idea. One such theory is that our associations are largely generated by means of their 'semantic overlap' with the statement we process (see for example Holyoak and Thagard 1989, p.301). Similarly, data collected by Russo, Medvec and Meloy (1996), among others, shows that there is a strong link between positive terms and positive associations. However, the question how exactly the process of associations functions is a live and widely debated one. It is well beyond the scope of this thesis to take position in this debate.

Examples of highly relevant information are clues that signal an increased chance of 'good news'. Since logical equivalence is, as have been just argued, less important than some other information, a lack of time may be a good reason to accept a partial-information conclusion.

A second candidate for a 'good reason' pertains to the payoff or importance of the decision problem. I have argued that some information is more accessible than others, and that this influences the cognitive effort it takes to both extract and process this information. In some cases, the problem at stake may simply be not important enough to put much effort in processing less readily available information.

A third candidate is the perceived reliability or authority of the source. As has been argued above extensively, we have several 'sensors' for detecting certain patterns in human behaviour, which subsequently trigger associations. In some cases, we have reasons to doubt whether these associations are warranted, but in (many) other cases, speakers do indeed (implicitly or explicitly) leak information by choosing specific words. Hence, depending on the reliability or authority of the source, it may be reasonable to assume that the source does indeed leak more information about a decision problem than the 'objective' descriptive content of her words suggests. In such cases, it may actually be better to stick to our first decision, as this decision is primarily based on this leaked ('non-logical') information.

There are probably more candidates to be found, but the general idea is clear: whether it is beneficial to put effort in processing as much information as possible depends heavily on the situation one is in and on what is at stake.

A question that springs to mind is what the consequences of this are for the subjects that participated in, for example, the Asian disease experiment. Were they right to be satisfied with their initial, partial-information decision? These questions reveal the problematic character of many of the framing experiments. Various factors that can potentially justify a subject's 'hasty' decision, are difficult to assess in experiments.

The artificial set-up of experiments makes it very difficult for subjects to assess whether it is justified to 'read between the lines' or not, whether they should put effort in making a careful decision or whether it suffices to follow their first intuition. Furthermore, the decisions made by the subject have no real consequences. Thus, incentives to consider more information than needed to reach a first conclusion are often distorted.

Fawcett et al. (2014) support this observation. They remark that many experiments of decision making are based on highly simplified contexts that ”bear little resemblance to the complex, heterogeneous world in which animals (including humans) have evolved” (Fawcett et al. 2014, p.153). As a result, many of the biases and errors that are observed in such studies may be due to the artificial situation presented to the decision maker, rather than the (mal)functioning of our cognitive system.

They argue, therefore, that one has to be careful in drawing conclusions about human reasoning based on some simple, unnatural decision tasks.[28] I agree with this conclusion. There is no clear ‘yes’ or ‘no’ to the question whether the framing effect is a failure of our rationality. In some cases, the cognitive process that underlies the framing effect may indeed lead us astray. However, in (many) others it may in fact lead to the best decision, given the circumstances we are in. A careful assessment of the context and set-up of the decision problem is needed to be able to say anything about these matters.

This last proviso is an important one for it means that, at least in my view, rationality should be assessed in the light of the circumstances or context in which the decision is made. As a consequence, there is more to rationality than merely ‘objective’ or ‘logical’ considerations.

### 2.4.2 Bounded Rationality

The upshot of the above is that, in my view, it is not of much use to measure rationality based on some unattainable ideal that does not reflect human cognitive capacities. Our preferences are formed based on the information and preference rules that are available to us at a certain time, constrained by several contextual factors such as the time, interests and cognitive capacities.

In other words, we are ‘bounded rational agents’. This concept of ‘bounded rationality’ was first introduced by Herbert Simon in his famous article *A Behavioral Model of Rational Choice* (1955). In this article, he attacks the notion of the ‘economic man’ as postulated by traditional economic theory. This economic man is a hyper-rational, nearly omniscient person that is able to calculate, in a well-organised, stable way, which course of action will yield the highest payoff on his preference scale, using all resources available to him (Simon 1955, p.99).

[28] A similar conclusion is reached by Stenning and van Lambalgen (2008). They combat the “tendency to push logic to the fringes [of cognition], in the wake of results allegedly establishing its irrelevance to human reasoning” (ibid., p.347). In these results, the specific context in which decisions are made is often neglected. If we do take this context into account, and subsequently adapt our logic to this, logic can be said to play “a much wider role in cognition than is customarily assumed” (ibid.).

Simon argues that this concept of the economic man is too much of an idealisation of actual, observed behaviour to be useful as a scientific postulate. He opposes this concept with his concept of man as a bounded rational agent (“a choosing organism of limited knowledge and ability”) (Simon 1955, p.114). This conception does take into account the influence of the limited capacities of man as well as contextual constraints on the decision-making process. Just as the maximum speed at which we can move constrains the set of available behaviour alternatives, so may “limits on computational capacity [be] important constraints entering into the definition of rational choice under particular circumstances” (ibid., p.100).

Simon argues that rational behaviour should not be defined as making the optimal decision based on all given information. Decision making is a delicate (and therefore costly) endeavour: it takes time and effort to collect and process information. Therefore, one should regard decision making as the process of picking a satisfactory solution given several limitations (e.g., pertaining to the availability of time, information, computational capacities etc.). I will return to this in the next chapter.

As seen above, framing influences or manipulates the order in which information is processed by us, and this may result in different choices or behaviour. Not necessarily because we are irrational, but rather because of the limited resources that are at our disposal. This forces us to draw conclusions based on partial information, either because only partial information is available, or because we only have the resources/incentives to process partial information.

Linked to this notion of bounded rationality is the view that our process of entailment, the information that is regarded to yield justifiable conclusions, is non-monotonic. That is, because we draw conclusions based on partial information, we can retract these conclusions when new information becomes available (or is processed).

In the process in which preferences are formed, this happens all the time. For example, we might prefer to commute to work by bike rather than by bus, until we come to know that it is raining outside. Similarly, we might prefer to save 200 lives with certainty rather than to save 600 with a $1/3$ chance, until we realise that this means that 400 people are sent to death.

A corollary of this is that we mostly draw conclusions (form preferences) based on ‘default rules’ or ‘rules of thumb’. These are rules that are not strict, universally applicable laws that always hold. Rather, many (if not all) of the rules we use to form a preference allow for exceptions, and hence these rules may or may not apply depending on the context. For example, one preference rule may be: 'Generally, I prefer to travel by bike rather than by bus' (e.g., because the former is healthier). However, this rule may not apply when it is raining. Hence, rainy days may be a systematic exception to the rule. Furthermore, there may be accidental exceptions as well. For example, one may decide to travel by bus today because one did not sleep very well last night.

As a result, one may draw the conclusion that one prefers option $x$ based on some preference rule, only to change this preference to $y$ as soon as one learns that this preference rule does not apply (or is overruled by another).

This notion of 'default rules' sheds a new light on Gold and List's conclusion that the framing effect only occurs when our beliefs are inconsistent. They take our preference rules to be strict, universal laws. Under some circumstances, these laws license contradictory results. If one looks at the preference rules we use from the perspective of a bounded rational agent, this conclusion no longer holds. Since we are dealing with limited resources and information, we draw conclusions based on the assumption that things we do not know of play no role. In other words, our preference rules are not strict, universal laws, but rather defeasible tools that help us draw conclusions based on the information available.

Different decision paths may yield opposing conclusions, but these conclusions are *tentative*. They are partial information decisions, based on the available resources. Due to the non-monotonic principle of entailment, it may very well be that when more information is considered, only one (or neither) of these conclusions follows. As a result, our total belief set need not be inconsistent at all.

In the next chapter, I will show that by making use of a non-monotonic decision process based on default rules, path dependence (and hence the framing effect) may arise, despite the fact that the decision maker's beliefs are perfectly consistent. I believe that this fits better with the above considerations about rationality than Gold and List's conclusion that some rationality principles *must* be violated for the framing effect to occur.

## 2.5 Conclusion

In chapter 1, the non-equivalence of information conveyed by different frames has been identified as a key notion for the framing effect. In this chapter, I have looked at two recent models of framing that partly incorporate this idea, a Decision Model without Extensionality (DM/E) and a Logic of Path Dependence (LPD). I have argued that both models, as well as Prospect Theory (PT), fail to fully accommodate the various types of framing identified in chapter 1. Furthermore, all three models struggle with some empirical results presented in this chapter, such as the 'elaboration effect'.

I have argued that the notion of 'information' should be conceived in a very broad sense: a specific formulation gives rise to various associations, facts and rules. Some of these are due to the 'objective' descriptive content of this formulation, others to the semantic properties, implicitly registered patterns or contextual factors related to this formulation.

This information is processed in a sequential way by the decision maker, and the frame in which an event is described greatly influences the accessibility of the information: some associations are more vivid than others, depending on the frame that is used. Since we are able to draw conclusions based on very little information, we often make a decision based on the most accessible information, thereby neglecting less focal information.

This leads to the framing effect: as framing influences the accessibility of information, different formulations cause different pieces of information to become focal. This focal information is processed first, and hence can lead to different decisions. I have argued that the decision process underlying the framing effect is the same for all types of framing, including topic framing.

Finally, I have argued that human beings are bounded rational agents, and that the framing effect need not be irrational nor be caused by inconsistent beliefs. Rather, a 'rational' decision depends on the information, time and cognitive capacities that are available, and on the added value of processing more information given the extra effort it takes.

In the next chapter, I will develop a dynamic model of the framing effect that is able to do provide an account of framing in its various forms, and that does justice to the insights gained so far.

# 3 The Logic of Framing

In the previous chapters, I have investigated what the framing effect is and how the underlying decision process functions. From the insights gained there, the following desiderata for a model of framing that is more comprehensive than the ones discussed can be formulated:

- First of all, such a model has to reflect that, in making decisions, we tend to take a wide array of information or associations into account, coming from various sources.
- Secondly, the model has to represent the underlying decision process of the framing effect outlined in the previous chapter. As a corollary, this means that the model has to be able to accommodate both topic framing and valence framing, in all its forms (risky choice, attribute and goal framing), since the underlying process is, in my view, (roughly) the same.
- Thirdly, the model has to be able to account for the lack of unanimity among subjects, and for the fact that the framing effect is less pronounced when the level of elaboration of the decision maker is high.
- Finally, the model has to do justice to our human nature as bounded rational agents. This means that it should leave room for revision of one's beliefs in the light of new information, and for preferences rules that allow for exceptions. Furthermore, it should avoid the conclusion that the framing effect is irrational 'come what may', but rather should leave room for more balanced considerations.

In this chapter, I will present a model that is able to meet all these requirements. I start out with a default logic presented by Raymond Reiter, and subsequently add a mechanism for resolving conflicts between preference rules. Afterwards, Gold and List's notion of decision paths will be adapted and incorporated into the framework. Finally, the model will be tested, some examples will be provided and its ability to meet the four desiderata outlined above will be assessed.

## 3.1 A New Model for the Framing Effect

In the previous chapter, it was argued that our preferences rules are likely to allow for exceptions and that the notion of entailment in human reasoning is non-monotonic. The endeavours to develop a new model for the framing effect therefore start out with a non-monotonic logic that can express default rules.

### 3.1.1 Reiter's Default Logic

One of the first logical frameworks that was able to meet these requirements is called 'Default Logic' and was presented in 1980 by Raymond Reiter. Even though there are many other non-monotonic logics around today, most of them make use of propositional logic rather than predicate logic, as Reiter does. Since framing is about stressing different sides (or properties) of the same object or event, it is convenient to make use of the greater expressivity of predicate logic. Therefore, Default Logic is a good starting point for the purposes of this thesis.[29]

A default theory $\Delta$ is a pair $(D, W)$. $W$ is a theory containing first-order sentences and can be regarded as a set of facts that are currently known (a 'background theory'). $D$ is a set of default rules. These rules have the following form:

$$\frac{\mathrm{A}(x) : M\mathrm{B}_i(x), \ldots, M\mathrm{B}_n(x)}{\mathrm{C}(x)} \qquad \textbf{(Default Rule)}$$

$A$ denotes the 'prerequisite' of the default rule $d$ and $C$ the 'consequent'. $B_i \ldots B_n$ are the 'justifications' of the rule. I will refer to these three parts as $pre(d)$, $cons(d)$ and $just(d)$ respectively. $M$ is to be read as 'it is consistent to assume'. For instance, $BIRD(x)$ : $\mathrm{M}FLY(x)$ / $FLY(x)$ means: "if $x$ is a bird and it is consistent to assume that $x$ has the property 'fly', then it follows that $x$ flies". These default rules will sometimes be abbreviated and written as '$A(x) \rightsquigarrow C(x)$'.

A default without a prerequisite is denoted as $\top$ : $\mathrm{M}B_i(x) \ldots \mathrm{M}B_n(x)/C(x)$ (or simply $\rightsquigarrow C(x)$) and is called a 'prerequisite-free default rule'. These rules will turn out to be particularly useful in resolving conflicts between preference rules, and in the analysis of goal framing.

Default Logic allows us to infer $FLY(tweety)$ from $BIRD(tweety)$ and the default rule $BIRD(x) \rightsquigarrow FLY(x)$, as long as we have no evidence to the contrary.

[29] There are a few other predicate default logics around, most notably Circumscription Logic by McCarthy (1980). These frameworks are generally more profound than Reiter's model, but for the purpose of this thesis (and for the sake of simplicity) his Default Logic will suffice. Note, however, that nothing in what is to follow relies on this choice, and the model presented here will work equally well with a different underlying (predicate default) logical structure.

That is, if none of the justifications is inconsistent with our current beliefs, the 'default conclusion' follows. This inference is ensured by the so-called 'closed-world assumption':

$$\frac{: M\neg\text{PRED}_i(x_1, \ldots, x_n)}{\neg\text{PRED}_i(x_1, \ldots, x_n)} \qquad \textbf{(Closed-World Assumption)}$$

This assumption says that one supposes for all $n$-ary predicates $PRED_i(x_1, \ldots, x_n)$ that $\neg PRED_i(x_1, \ldots, x_n)$ holds whenever it is consistent to do so (Reiter 1980, p.84). This means that, in the reasoning process, we are only concerned with the set of facts and (default) rules that are at our disposal, and hence are allowed to draw (tentative) conclusions based on what is known (rather than being restrained by what is unknown).

As a result, the logic is not monotonic. For example, let $A = \{BIRD(x) \rightsquigarrow FLY(x), BIRD(tweety)\}$, then $A \models FLY(tweety)$. That is, given that all we know is that birds typically fly and that tweety is a bird, we are justified to conclude that tweety can (presumably) fly. However, suppose that we learn some new facts, such that our belief set expands to $B = A \cup \{PENGUIN(tweety), \forall x \; PENGUIN(x) \rightarrow \neg FLY(x)\}$. Then despite the fact that $A \subseteq B$, it is not the case that $B \models FLY(tweety)$, for it is no longer consistent to assume that the default can be applied. In words, if we subsequently learn that tweety is not only a bird but a penguin as well and that penguins do not fly, we are forced to retract our earlier conclusion.

Under the assumption that we are bounded rational agents, it seems straightforward that we have an incomplete theory ($W$) of the world. Default rules can be regarded as ways to 'extend' this incomplete theory by using the resources we have available. That is, they allow us to fill gaps in our theory of the world by using observed regularities (things that are 'typically' the case). In technical terms, the set of default rules ($D$) induces an 'extension' ($E$) of $W$ (ibid., p.87).

In what is to follow, it is assumed that the default theory is 'closed'. This means that the formulas in $W$ and $D$ do not contain free variables.[30] An extension $E$ is defined as follows:

[30] This is mainly for simplicity's sake. A generalisation for arbitrary theories can be bound in Reiter (1980, pp.115-129).

**Definition 3.1** *Classical Extension*[31]
*Let $\Delta = (D, W)$ be a closed default theory. $E$ is an extension of $\Delta$ iff it is the smallest set satisfying the following three properties:*

*(1)* $W \subseteq E$
*(2)* $E = \mathrm{Th}(E)$ *(i.e., the deductive closure of E)*
*(3) For any default* d *(of the form* $pre(d)$ : $just(d)$ / $cons(d)$*), if* $pre(d) \in E$ *and* $\neg just(d) \cap W = \emptyset$ *then* $cons(d) \in E$*.*

In words, an extension $E$ of our theory $W$ is a deductively closed set of sentences, starting out with $W$ and in which subsequently all default rules that can be applied are applied (i.e., as much gaps of $W$ as possible have been filled).

Note that this process of extending a given $W$ may be nondeterministic. As $W$ may contain multiple default rules, "[d]ifferent applications of the defaults [can] yield different extensions and hence different sets of beliefs about the world" (Reiter 1980, p.87). For example, suppose you meet a 21 year old student ($W = \{ADULT(a), STUDENT(a)\}$). Furthermore, you believe that adults are typically employed ($ADULT(x) \rightsquigarrow EMP(x)$) and that students are typically not employed ($STUDENT(x) \rightsquigarrow \neg EMP(x)$). Our knowledge base lacks any information about whether the student is employed. This can be overcome by extending our $W$ using defaults. However, if we use the first default, we can conclude that the student is (presumably) employed, and, as a result, are no longer able to apply the second default, as this yields the opposite conclusion. The converse holds if we use the second default first. This means that there are multiple (complete) extensions of our belief set $W$.

### 3.1.2 Prioritised Default Logic

How can one choose between these different, mutually exclusive, extensions? In other words, which one is the 'best', given the situation? In the example above, one can question, based on the information available, whether one conclusion really is better than another. That is, as long as one does not receive more information that may indicate that one rule is more relevant than another, it may be better not to draw any conclusion at all.[32]

[31] Adapted from (Reiter 1980).

[32] One such piece of information that may shed a new light on the relevance of the two competing default rules is that students are typically adults ($STUDENT(x) \rightsquigarrow ADULT(x)$). This allows one to regard students as exceptions to the rule $ADULT(x) \rightsquigarrow EMP(x)$. See Bastiaanse and Veltman (forthcoming) for a principled method, drawing on Circumscription Logic, for prioritising certain rules over others based on their semantic properties.

However, in the context of preferences, which is the main concern here, the situation is different. We often do have a preference for one action rather than another, despite the fact that we have arguments for both. Hence, it is reasonable to assume that we attach different weights to our preference rules. This weight determines which preference rule is stronger.

For example, suppose that someone wants to be in good shape. This may induce this person to prefer to travel by bike rather than by car. However, suppose that she also does not want to be sweaty all day. This may induce her to prefer to travel by car rather than by bike. Even though she has arguments for both sides, she may find the latter more important than the former.

In this case, both arguments are *semantically* unrelated to each other (except for the fact that they yield opposite conclusions). Neither preference is defeated by the other, nor conveys any information about being a special case or atypical circumstance that may affect the other rule. They are simply two arguments for opposing conclusions.

This happens a lot in the context of preference rules. Since the arguments are unrelated, we are not able to give a systematic procedure to determine which rule should prevail based on the semantic properties of the rules/facts known. Nevertheless, it is often the case that an agent does find one preference rule more important than another. To represent this in the model, a superiority relation will be introduced to express priorities among preference rules.

Following Brewka and Eiter (2000), a priority relation $<$ is added to the model that provides an ordering on the default rules . $<$ is a strict partial order, and $d < d'$ indicates that default $d$ is preferred over default $d'$. A prioritised default theory, then, is a triple $\Delta = (D, W, <)$. Since $<$ is a partial order, the preference ordering among certain defaults can be left unspecified (ibid., p.30). For the sake of simplicity, however, in what is to follow $<$ is taken to be a total ordering, and $\Delta = (D, W, <)$ is called a 'fully prioritised' default theory.[33]

In the case of two conflicting extensions of some default theory $\Delta$, these priorities are used to ensure that the (intuitively) correct extension comes up as the preferred one. Brewka and Eiter formulate two principles that a system for preference handling must satisfy in order to guarantee the correct handling of priorities:

[33] Note that an arbitrary prioritised default theory can be reduced to a fully prioritised one by defining $E$ as a prioritised extension of $(D, W, <)$ iff $E$ is a prioritised extension of some fully prioritised default theory $(D, W, <')$ such that $< \subseteq <'$, see Delgrande, Schaub and Tompits (2000, p.380).

**Principle 1:**
If one has two extensions of a theory $\Delta$, $E_1$ and $E_2$, generated by the defaults $d_1$ and $d_2$ respectively, and it is the case that $d_1 < d_2$, then it cannot be that $E_2$ is a preferred extension.

**Principle 2:**
If $E$ is a preferred extension of $\Delta$ and $d$ is a default such that $pre(d) \notin E$, then $E$ is a preferred extension of $\Delta' = (D \cup \{d\}, W, <')$ if $<'$ and $<$ have the same priority ranking among the defaults in $D$. That is, if one adds a default to the theory that is not applicable in a preferred extension, then this default cannot cause this extension to become non-preferred.

The intuition behind this second principle is that "whether to believe a formula $[cons(d)]$ or not should depend on the priorities of the defaults contributing to the derivation of $[cons(d)]$, not on the priorities of defaults which become applicable when $[cons(d)]$ is believed [...] In other words, a belief set is not blamed for not applying rules which are not applicable." (Brewka and Eiter 2000, p.31).

With these two principles in mind, a selection procedure for choosing between competing extensions based on the priorities of the default rules can be specified. A default rule $d$ will be called 'active' in a set of formulas $W$ if it is applicable in $W$ but has not been applied so far. That is, $d$ is active if the prerequisite of $d$, $pre(d)$, is in $W$, if $\neg just(d) \cap W = \emptyset$ and if $cons(d) \notin W$. With this in mind, a selection operator $C(\Delta)$ is defined:

**Definition 3.2** *Selection operator* $C$[34]
*Let* $\Delta = (D, W, <)$ *be a closed, fully prioritised default theory.* $C(\Delta)$ *is defined as follows.* $C(\Delta) = \bigcup_{i \leq 0} E^i$*, where* $E^0 = Th(W)$ *(the deductive closure of W), and for every ordinal* $i > 0$*:*

$$E^i = \begin{cases} \bigcup_{j<i} E^j & \textit{if no default from D is active in } \bigcup_{j<i} E^j \\ Th(\bigcup_{j<i} E^j \cup \{cons(d)\}) & \textit{otherwise, where } d \in D \textit{ is the minimal default w.r.t} < \textit{active in } \bigcup_{j<i} E^j \end{cases}$$

This operator $C$ provides us with an algorithm for selecting a prioritised extension $E_i$ of $\Delta$.

[34] Adapted from Brewka and Eiter (2000) and Delgrande, Schaub and Tompits (2000).

This selection process goes as follows.[35] One starts at $E^0$ with $Th(W)$ and subsequently considers all default rules $d \in D$, starting with the one with the highest priority (e.g., $d_1$). Hence, in $E^1$, one checks if $d_1$ is active (i.e., whether it can be applied but has not been so far). If so, $cons(d_1)$ is added to $Th(W)$ and one moves on to $E^2$. If not, one checks the other default rules, starting with one with the second highest priority, then the third highest priority, then the fourth highest priority etc. until one has found a default rule that is active. When such active default rule is found, its consequent is added to $Th(W)$, and one moves to $E^2$. At $E^2$, one does the same, again considering the highest active default first, and adding its consequence to $\bigcup_{j<2} E^j$. This process is repeated until no more defaults can be applied.

Now what may happen is that if one adds some $cons(d_i)$ to $E^n$, a more prioritised default $d_h$ becomes applicable in $E^{n+1}$. The operator $C$ ensures that in this next round $E^{n+1}$, $cons(d_h)$ is indeed added to the extension. Hence, the defaults with the highest priorities are applied as soon as this is possible.

Unfortunately, this operator $C$ alone is not sufficient to guarantee that the correct extension comes up as preferred (i.e., that the preferred extension complies with the two principles stated above). Consider the following example:

**Example 3.3** *Violation of Principle 1*
*Let* $\Delta = (W, D, <)$ *where* $W = \emptyset$ *and* $D$ *consists of the following rules, with* $d_1 < d_2 < d_3$*:*

*(*$d_1$*)* $PRED_1(x) : \mathrm{M}PRED_2(x)\ /\ PRED_2(x)$
*(*$d_2$*)* $\top : \mathrm{M}\neg PRED_2(x)\ /\ \neg PRED_2(x)$
*(*$d_3$*)* $\top : \mathrm{M}PRED_1(x)\ /\ PRED_1(x)$

This theory has two classical extensions, $E_1 = Th(\{PRED_1(x), PRED_2(x)\})$ and $E_2 = Th(\{PRED_1(x), \neg PRED_2(x)\})$. That is, if one does not take priorities into account, one can either apply $d_2$ first, or $d_3$, as they are both prerequisite free (in both cases, $pre(d) = \top$). In the former case, $d_1$ can no longer be applied, in the latter case it can.

It is easy to see that $C(\Delta) = E_2$. One starts out with $E^0 = \emptyset$, and subsequently applies $d_2$, as this is the default rule with the highest priority that is active. Hence,

[35] Note that $E^i$ is used to refer to the stages of the selection operator $C$, whereas $E_i$ is used to refer to an extension of a default theory $\Delta$. As seen above, in a classical setting (i.e., without taking priorities into account), it is possible that multiple extensions ($E_i$) of a theory $\Delta$ exist. The operator $C$ is designed to pick the most preferred one by looking at priorities. Hence, the final stage of the operator $C$, $E^{final}$, is equal to one of the classical extensions $E_i$.

$E^1 = Th(\{\neg PRED_2(x)\})$. In $E^2$, $d_3$ is the highest active default rule, and so $PRED_1(x)$ is added to the set. Rule $d_1$ cannot be applied, as it is defeated by $\neg PRED_2(x)$, thus one ends up with $Th(\{PRED_1(x), \neg PRED_2(x)\})$.

This violates principle 1. Given that $d_1 < d_2$, one should prefer the extension in which $d_1$ is applied (namely, $E_1$). To overcome this problem, the theory $\Delta$ has to be 'preprocessed' before applying $C$. A prioritised extension $E$ of $\Delta$ is defined as follows:

**Definition 3.4** *Prioritised Extension*[36]
*Let $\Delta = (D, W, <)$ be a closed, fully prioritised default theory. $E$ is a prioritised extension of $\Delta$ iff the following is satisfied:*

*(*a) *$E$ is a classical extension of $\Delta$ (see Definition 3.1)*

*(*b) *$E$ is a prioritised extension of a 'preprocessed' $\Delta_E = (D_E, W, <_E)$. This prioritised extension is obtained by applying $C$ to $\Delta_E$ (i.e., $E = C(\Delta_E)$), where $\Delta_E$ is constructed from $\Delta$ as follows:*

*(*i) *For every default $d \in D$ such that pre$(d) \notin E$, it holds that $d \notin D_E$*

*(*ii) *For every remaining default, pre$(d)$ is replaced by $\top$ such that one obtains $d^\top$, the prerequisite-free version of $d$*

*(*iii) *For any $\delta_1, \delta_2 \in D_E$, $\delta_1 <_E \delta_2$ iff $d_1 < d_2$ where $d_i = \max_<\{d \in D \mid d^\top = \delta_i\}$.*

Hence, one applies $C$ to some preprocessing $\Delta_E$ of $\Delta$ and if the resulting $E$ is equal to a classical extension $E$ of $\Delta$, one has found the preferred extension. To see how this works, it is best go provide some examples.

**Example 3.5** *Preferred Extension*
*Let $\Delta = (W, D, <)$ be the same as in example 3.3.*

As explained before, this theory has two classical extensions, $E_1 = Th(\{PRED_1(x), PRED_2(x)\})$ and $E_2 = Th(\{PRED_1(x), \neg PRED_2(x)\})$. Obviously, our preferred extension must be one of these (condition (a)). These two extensions are used to generate $\Delta_{E_1}$ and $\Delta_{E_2}$, respectively, and subsequently apply $C$.

First consider $E_1$. Given that $d_2$ and $d_3$ are prerequisite free, and given that the prerequisite of $d_1$, $PRED_1(x)$, is in $E_1$, condition (i) is satisfied vacuously. Condition (ii) makes that $D_{E_1}$ consists of the following rules:

[36] Adapted from Brewka and Eiter (2000) and Delgrande, Schaub and Tompits (2000).

($d_1'$) $\top : \mathrm{M}PRED_2(x) \; / \; PRED_2(x)$

($d_2'$) $\top : \mathrm{M}\neg PRED_2(x) \; / \; \neg PRED_2(x)$

($d_3'$) $\top : \mathrm{M}PRED_1(x) \; / \; PRED_1(x)$

Condition (iii) ensures that the ordering $<_E$ remains the same (except that now $d_1$ is replaced by $d_1^{\top}$). If one applies $C$ to $\Delta_{E_1}$, one starts out with $E_0 = Th(W) = \emptyset$, adds $cons(d_1')$ in $E_1$, and adds $cons(d_3')$ in $E_2$. Hence, one ends up with $C(\Delta E_1) = Th(\{PRED_1(x),\; PRED_2(x)\}) = E_1$. As a result, $E_1$ satisfies both requirements (a) and (b) of definition 2.3 and thus is a preferred extension of $\Delta$.

Notice that $E_2$ is not a preferred extension. $D_{E_2}$ consists of the same $d_1', d_2', d_3'$ as $D_{E_1}$. This means that $C(\Delta E_2) = C(\Delta E_1)$, and, as seen above, this is not equal to $E_2$. Hence, $E_2$ violates requirement (b).

By following the above procedure, one will always end up with a unique preferred extension that complies with both Principle 1 and 2, given that a preferred extension exists.[37] This latter, however, need not always be the case. Consider the following example:

**Example 3.6** *No preferred extension*[38]
*Let* $\Delta = (W, D, <)$ *where* $W = \emptyset$ *and* $D$ *consists of the following rules, with* $d_1 < d_2 < d_3$*:*

*($d_1$):* $PRED_1(x) \; : \; \mathrm{M}\neg PRED_2(x) \, / \, \neg PRED_2(x)$
*($d_2$):* $\top \; : \; \mathrm{M}PRED_2(x) \; / \; PRED_2(x)$
*($d_3$):* $PRED_2(x) \; : \; \mathrm{M}PRED_1(x) \; / \; PRED_1(x)$

There is one classical extension $E = Th(\{PRED_1(x), PRED_2(x)\})$. However, $E$ is not preferred. This is because the preprocessed $D_E$ contains $\{d_1^{\top},\; d_2,\; d_3^{\top}\}$. Thus, if one applies $C$ to $\Delta_E$, the rule with the highest priority that is active is considered first, $d_1^{\top}$, and so one adds its consequent, $\neg PRED_2(x)$, to the extension. Afterwards, one considers the rule with the second highest priority that is active. This is $d_3$: since $\neg PRED_2(x)$ is added to the theory, it is no longer justified to apply $d_2$. Hence, $PRED_1(x)$ is added to the extension and one ends up with $C(\Delta_E) = Th(\{\neg PRED_2(x),\; PRED_1(x)\} \neq E$.

This is not necessarily a bad thing, as it only occurs when the agent's preferences together with her priorities are inconsistent (i.e., when the former are incompatible with the latter). Hence, if the agent fails to generate a prioritised extension, she will not draw any conclusions, but rather will be alarmed that something is wrong.

[37] For a proof, see Brewka and Eiter (2000, p.38).

[38] Adapted from Brewka and Eiter (ibid.).

### 3.1.3 Decision Paths

As a final step, the notion of a decision path is added to the framework. This notion is inspired by Gold and List, but a different characterisation of the decision process will be given.

**Definition 3.7** *Decision Path*
*Let $\Delta = (D, W, <)$ and let $X = D \cup W$. A decision path on $X$ is a bijective function $\Omega : \ (1, 2, \ldots, n) \rightarrow X$, where $n$ is the number of elements (facts and default rules) in $X$.*

The sequential processing of information can be represented with this decision path, just as seen in chapter 2. However, since an underlying non-monotonic logic with default rules is used here, the tentative character of the conclusions the agent becomes clear. That is, by moving on along the decision path, she processes pieces of information and draws conclusions on the fly. These conclusions are not unassailable truths, but are based on what is known to the agent at a certain point. In subsequent points on the decision path, new information becomes available to her, and this may lead to new insights.

For the sake of simplicity, the notion of 'initial disposition' of the agent is dispensed with. It is simply assumed that the agent is willing to accept all the associations/information she 'extracts' from a decision problem. If this results, at a certain point, in a contradictory belief set, the decision process ends and the agent has to reconsider her associations. I will come back to this later. First, I will define how our default theory $\Delta$ expands as one walks along the decision path:

**Definition 3.8** *Information Processing*
*Let $X$ be a set of formulas with $n$ elements, representing the information (in the broadest sense of the word) contained by a decision problem. $X$ includes both 'facts' and preference rules. Let $\phi_{target}$ be the goal of the decision process. Let $\Omega$ be a decision path on $X$ and let $\phi_j \in X := \Omega(j)$. Let $<_n$ be some (partial) priority ranking on the defaults in $X$.*

*The agent's theory $\Delta_j$ at step $j$ of the decision path is defined as follows:*

*(1) $\Delta_0 = (D_0, W_0, <_0)$, where $D_0 = W_0 = <_0 = \emptyset$*
*(2) For all $0 < j \leq n$, $\Delta_j = (D_j, W_j, <_j)$ it holds that:*
*If $\phi_j$ is a default, $W_j = W_{j-1}$, $D_j = D_{j-1} \cup \{\phi_j\}$*
*If $\phi_j$ is fact, $D_j = D_{j-1}$, $W_j = W_{j-1} \cup \{\phi_j\}$*
*In both cases, $<_j = <_n\restriction D_j \ (= \{(d, d') \mid d, d' \in D_j \ \& \ (d, d') \in <_n\})$*

At each step $j+1$ of the decision path, the agent expands her default theory ($\Delta_j$) by adding a new formula ($\phi_{j+1}$) to either her current set of facts ($W_j$) or default rules ($D_j$), depending on the nature of $\phi_{j+1}$. The specific course of the decision path determines which formula is added at step $j+1$. Furthermore, she updates her priority relation ($<_j$) to reflect her (hard-wired) priorities among preference rules.

Using this theory $\Delta_j$, the agent comes to a decision in the following way:

**Definition 3.9** *Decision Process*
*Let $\Delta_j$ be a theory at step j of the decision path of length $n$. Let $E_j$ be the preferred extension (if existent) of $\Delta_j$ (i.e., $E_j = C(\Delta_j)$. The decision process goes as follows:*

(I) *If $\phi_{target} \notin E_j$ then move on to $j+1$*

(II) *If $\phi_{target} \in E_j$ and $j = n$, then decide $\phi_{target}$*

(III) *If $\phi_{target} \in E_j$ and $j \neq n$, then decide $\phi_{target}$ if one is satisfied with this conclusion, otherwise move on to $j+1$*

In short, what happens is that at each point of the decision path, the agent adds a new piece of information to her theory $\Delta$. Subsequently, she extends this theory based on the information she has to see whether she can reach a (tentative) conclusion $\phi$ about the target proposition. If not, she moves on to process more information.

If she does reach a conclusion $\phi$, and furthermore has no more information to process, she simply decides $\phi$. However, if the agent reaches a conclusion but has not yet made it until the end of the decision path (i.e., when she has tentatively reached her goal but has not yet fully processed all information), she can either stop reasoning and conclude $\phi_{target}$ (despite its tentative status) or she can move on and take more information into account. Which option she takes, depends on whether she is 'satisfied' with this tentative decision or not.

As seen in the previous chapter, one can have good reasons to be satisfied with a tentative conclusion. That is, I have argued that under certain circumstances, accepting a tentative conclusion is the 'best' or 'rational' choice to make. However, not all reasons qualify as 'good'. Sometimes, one rashly accepts the first conclusion that springs to mind (out of laziness, disinterest, impatience etc.), without carefully considering or even being aware of the fact that it is based on partial information.

It is difficult to give a proper characterisation of this process of 'satisfaction', as many factors and considerations seem to be at work. I will not attempt to provide a formalisation or systematisation of this notion here, but instead refer to the literature on 'satisficing'. This notion, introduced by Herbert Simon (1956), refers to a decision-making process in which an agent is not in search of the ‘perfect’ solution, but rather takes it to be sufficient when some constraints are satisfied.

For example, suppose that an agent is looking for a needle to sew a pouch, and suppose that she is standing next to a haystack containing several needles. Assuming that the sharpest needle is the most efficient for the job and produces the best result, she can either search the haystack until she has found the sharpest needle in there, or she can stop searching as soon as she has found a needle sharp enough for her purposes. A ‘satisficer’ would go for the second option, either because it is too costly or time consuming to search for the ‘optimal’ solution, or because the agent simply feels no need to consider any alternatives once a solution has been found that is ‘good enough’ (Simon 1997, p.296). Several frameworks have been developed to represent this idea. Two classic examples are Cyert and March (1963) and Odhnoff (1965).[39]

What is most important about the representation of the decision process given above is that one is often able to draw conclusions based on only a few associations, but is also able to take more information into account. As a result, the decisions of agents are based on varying amounts of information.

The framing effect, then, can be said to occur when different frames lead to different conclusions when the same *amount* of information is taken into account. That is, given two agents with the same preferences and priorities, one presented with frame $a$, the other with frame $b$, one speaks of framing when the agents come to different conclusions at the same stage $t$ of the decision path.

By making use of the dynamic aspect of the model, the difference between valence framing and topic framing can now be expressed as follows. Valence framing, in its purest form, is characterised by the fact that the different frames induce different decision paths on the same underlying set of associations $X$. This means that under full information, the model would predict that different frames lead to the same decision (given that the priorities among preference rules are the same). This is in line with the ‘elaboration effect’ observed by Takemura (1994). I will come back to this in the next section.

[39] In the context of framing, a problem arises if one represents ‘satisfaction’ based on the interplay between the expected ‘payoff’ of a more informed conclusion and the additional ‘costs’ of processing the extra information needed for reaching this conclusion. One can say that an agent is satisfied with a conclusion $\phi_j$ at stage $j$ if the marginal costs of a more informed conclusion $\phi_k$ at stage $k$ exceed the marginal ‘revenue’ of this conclusion. However, it seems that an agent can only properly gauge the marginal cost and revenue of processing more information if she knows what this information (and the resulting conclusion) entails. Hence, she must (implicitly) take this information into account when determining whether she is satisfied or not with the current conclusion. But the main point of partial-information decisions is exactly that she does *not* take all information into account.

One speaks of topic framing, on the other hand, if different frames induce decision paths on different sets of associations. In other words, in topic framing the frames do not convey the same information. As argued in chapter 1, these two concepts of valence and topic framing form a continuum, rather than a dichotomy, and in many experiments of framing, the two frames convey 'more or less' the same information.

Finally, it is important to note that for framing to actually occur in practice, more is needed than just two frames that can lead to different decisions. The circumstances in which the frame is posed have to be such that the decision maker is actually induced to follow the different decision paths. For example, the source has to be credible enough for the decision maker to go along with her (rather than rejecting the frame and the information it conveys right away). This is what Gold and List have called the 'empirical condition'.[40]

## 3.2 The Various Types of Framing in the New Model

Now that the adapted model for the framing effect has been presented, it is time to put it to the test. First of all, it has to be checked whether the model succeeds in accommodating the various types of framing as identified in chapter 1: risky choice framing, attribute framing, goal framing and topic framing.

### 3.2.1 Risky Choice Framing

The results of the Asian disease experiment can be explained very much along the lines of Gold and List's Logic of Path Dependence. However, I now have the tools to explain the findings of Takemura (1994) as well, by showing how an increased degree of elaboration can affect the amount of information that is processed.

Recall the set of information conveyed by the decision problem of choosing between programme A and B, and between C and D, as presented in 2. For now, I assume that both problems are (logically) equivalent and give rise to the same information (even though the accessibility of the various pieces of information is different under the different frames). In both cases, $W$ consists of the following propositions:[41]

[40] For a better understanding of the circumstances in which this condition is satisfied (i.e., in which successful communication takes place), a game theoretic analysis can be useful. In the conclusion, I will do some suggestions for future research on this topic, drawing on the results of this thesis.

[41] Again, $p_{a/c}$ is to be read as $p_a$ for an agent presented with the programme A / programme B choice, and as $p_c$ for an agent presented with the programme C / programme D choice. For $p_{b/d}$, this is $p_b$ and $p_d$ respectively.

(i) $SAVE(p_{a/c})$: Programme A / C saves some lives with certainty

(ii) $\neg SAVE(p_{b/d})$: Programme B / D involves the risk that no one will be saved

(iii) $DEATH(p_{a/c})$: Programme A / C entails the certain death of some people

(iv) $\neg DEATH(p_{b/d})$: Programme B / D offers the chance that no one will die

Furthermore, the following (defeasible!) preferences rules are in $D$:

(v) $SAVE(x) \wedge \neg SAVE(y) \rightsquigarrow PREF(x, y)$: Generally, it is not worth taking the risk that no one will be saved

(vi) $DEATH(x) \wedge \neg DEATH(y) \rightsquigarrow PREF(y, x)$: Generally, it is unacceptable that some people will die with certainty

Again, the frame of programmes A and B, couched in terms of lives saved, is more likely to put propositions $(i)$, $(ii)$ and $(v)$ up front on the decision path, while the frame of programmes C and D is more likely to trigger propositions $(iii)$, $(iv)$ and $(vi)$ first.

For an agent presented with the former frame ('agent 1'), at $t = 3$ this results in a theory consisting of $W = \{SAVE(p_a),\ \neg SAVE(p_b)\}$ and $D = \{SAVE(x) \wedge \neg SAVE(y) \rightsquigarrow PREF(x, y)\}$. This theory has one (prioritised) extension, in which $PREF(p_a, p_b)$ is added to $W$. Hence, the agent tentatively concludes that she prefers programme A over B. For an agent presented with the latter frame ('agent 2'), $W' = \{DEATH(p_c),\ \neg DEATH(p_d)\}$ and $D' = \{DEATH(x) \wedge \neg DEATH(y) \rightsquigarrow PREF(y, x)\}$ at $t = 3$. This leads her to the conclusion (extension) that programme D is preferred over C.

This is perfectly in line with the results observed by Tversky and Kahneman. Now, in chapter 2, I have argued that some testing subjects are not satisfied with this tentative conclusion, and rather take more information into account. This number of unsatisfied subjects increases when subjects are encouraged to elaborate on their decision, or when they are presented with an unusual or personal decision problem. In all these cases, subjects are triggered to 'think twice' before making a decision. That is, they move on to $t = 4$.

At $t = 4$, the 'immediate' or 'directly accessible' associations have already been processed, and other, less focal considerations are taken into account. For example, agent 1 comes to realise that programme A not only saves lives with certainty, but also results in the certain death of other lives ($DEATH(p_a)$). Subsequently ($t = 5$), she processes the fact that programme B offers a chance that no one will die ($\neg DEATH(p_b)$). This triggers the preference rule that sending people to death is unacceptable ($DEATH(x) \wedge \neg DEATH(y) \rightsquigarrow PREF(y,x)$) at $t = 6$.

Now suppose that she takes this rule to be stronger than her earlier preference rule ($(vi) < (v)$). At $t = 6$, her default theory includes propositions $(i) - (vi)$. This theory results in the following two classical extensions: $E_1 = Th(\{(i),\ (ii),\ (iii),\ (iv),\ PREF(p_a, p_b)\})$ and $E_2 = Th(\{(i),\ (ii),\ (iii),\ (iv),\ PREF(p_b, p_a)\})$. Using the prioritisation procedure outlined above, in both cases one ends up with a preprocessed $D_E$ consisting of the rules:

(vii) $\rightsquigarrow\ PREF(p_a, p_b)$

(viii) $\rightsquigarrow\ PREF(p_b, p_a)$

That is, since the prerequisites of both $(v)$ and $(vi)$ are in $E_1$ as well as $E_2$, one replaces them with their prerequisite-free versions $(vii)$ and $(viii)$. If one now applies the operator $C$ to the resulting default theories $\Delta_{E_1}$ and $\Delta_{E_2}$, in both cases one ends up with the prioritised extension $E = Th(\{(i),\ (ii),\ (iii),\ (iv),\ PREF(p_b, p_a)\})$. Since this is equal to $E_2$, $E_2$ is the only extension that meets both conditions (a) and (b) for a prioritised extension (see section 3.1.2).

Hence, agent 1 now concludes $PREF(p_b, p_a)$. In other words, she prefers programme B over A, contrary to her earlier conclusion. For agent 2, the process will be similar, and in the end her priorities among the default rules will determine her final choice.

By adding this possibility to process more information than necessary for reaching a first conclusion, different subjects may draw different conclusions based on different *amounts* of information, even though the information itself may be (roughly) the same for all subjects. In this way, the model is able to explain both the observed lack of unanimity among subjects, and the attenuation of the framing effect due to increased elaboration, unusual and personal decision problems.[42] As seen earlier, in cases of full information, the priority the subject attaches to the preference rules that are at work will determine her final choice. It seems reasonable to assume, as Takemura (1994) observed, that in the case of the Asian disease experiment, the proportion of people choosing the certain outcome and the proportion choosing the uncertain outcome will be roughly the same under full information.

[42] Note that in this example, I use quite simple and straightforward associations and preference rules. In practice, however, different subjects can have different associations, and the preference rules used may be much more complex. Besides the amount of information processed, these differences and increased complexity may also be sources for different choices among subjects.

### 3.2.2 Attribute Framing and Topic Framing

Can the model account for the other types of framing (attribute, goal and topic framing) as well? Both attribute and topic framing seem to be unproblematic. In the former case, a positive frame about, for example, employment figures is likely to first trigger associations related to the positive effect of high employment on economic growth. A negative frame, in terms of unemployment, is more likely to first trigger negative associations, highlighting the negative impact of unemployment on the economy. As a result, the 'default' conclusion in the positive frame is more likely to be positive, whereas the converse holds for the negative frame.

Only after this directly accessible information has been processed do the agents that are not yet satisfied consider the fact that $x\%$ employment at the same time means $(100 - x)\%$ unemployment, and vice versa. This additional information may lead to a more balanced conclusion.

For topic framing, in which the different frames are not (taken to be) logically equivalent but rather stress different aspects or dimensions of a decision problem (such as environmental impact or economic benefits of building a new road), there is a similar underlying process. If the agent is willing to accept the information that is presented to her, the frame gives priority to the associations related to the dimension that is stressed. This may result in a tentative opinion of the issue, and the stressed dimension is likely to play a dominant role here. For example, if one is willing to accept an argument about the economic benefits of building a new road, this frame will trigger various associations in which (positive) economic considerations prevail.

If the agent is not satisfied with this conclusion, she may process more information related to the issue at stake. Here, the difference with valence framing, as outlined above, comes into play. In the case of topic framing, the information conveyed by other frames may not be accessible at all, no matter how much time one takes. That is, due to the logical equivalence of risky choice framing and attribute framing, the agent was, with the lapse of time, able to derive the negative associations from the positive frame and vice versa. Hence, the underlying information set $X$ of both frames was the same.[43] In topic framing, however, one cannot generally derive arguments related to the environmental impact of building a new road from arguments related to the economic benefits of this road. As a result, when all associations directly related to one frame are processed, the agent may consider

[43] Recall that I have argued that the occurrence of frames that are perfectly equivalent is quite rare, and that in many occasions the two frames are only equivalent up to some degree.

other arguments she happens to know about, but there is no guarantee that her final decision will be the same under all frames.

This highlights a second important difference between valence framing and topic framing (apart from the (non)-equivalence of the full information set $X$ of the different frames). In a valence framing context, a decision maker is prone to the framing effect when *both* the 'empirical condition' is met (i.e., when she accepts the frame presented to her and sets off to follow the accompanying decision path) *and* when she is satisfied with a *partial*-information decision (i.e., when she does not reach the end of the decision path). In a topic framing context, on the other hand, the first condition alone is already sufficient. This is because, in this case, the different frames convey different information and hence can lead to different decisions *even if* the decision maker takes all available information into account. This may be an important reason why topic framing is so ubiquitous, for instance in politics.

It is important to note that, despite these differences, in both risky choice, attribute and topic framing the underlying process is very much the same: the influence of a frame on the accessibility of certain information has a large initial effect on the decision process.

### 3.2.3 Goal Framing

There is one type of framing that has not been discussed so far, and of which none of the existing frameworks discussed could provide a proper representation: goal framing.

As has been explained in chapter 1, goal framing has received relatively little attention, and as a result, there are few explanations of the phenomenon around. The interesting thing about goal framing is that the negative frame, rather than the positive one, is more persuasive. That is, people are more likely to perform a certain act when the negative consequences of not performing this act are stressed.

This has generally been linked to a widely observed 'negativity bias' in our processing of information. That is, "negative events appear to mobilize physiological, affective, cognitive, and certain types of social resources to a greater degree than do positive or neutral events" (Taylor 1991, p.72).

There are various explanations for this bias. A popular one reverts to evolutionary considerations. Pratto and John (1991) argue, for example, that events with negative consequences tend to be of "greater time urgency" than events with positive consequences (ibid., p.380). That is, averting harm often requires one to make decisions as quickly as possible, while "positively valenced activities, such as feeding and procreation, are less pressing" (Pratto and John 1991, p.380). They argue that this has led to the development of a cognitive mechanism of 'automatic vigilance' that directs our "attentional capacity to undesirable stimuli", thereby inducing us to focus more on negative information than on positive information (ibid., p.390).

Other researchers do not so much focus on the greater priority we give to negative information, but rather on the greater weight we attach to it. It ties the greater persuasive impact of negative information to observed behavioural patterns such as the 'status quo bias' and 'loss aversion'. That is, people tend to avoid change, and in particular loss (Kahneman, Knetsch and Thaler 1991).

Even though this is one of the most robust and widely observed psychological phenomena, surprisingly little is known about its underlying causes. Most authors take it to be a byproduct of our cognitive system(s) (e.g., Rick 2011) or a flaw of the mind. Camerer (2005) suggests, for example, that loss aversion is in most cases "an exaggerated emotional reaction of fear, an adapted response to the prospect of genuine, damaging, survival-threatening loss" (ibid., p.133).

I would like to propose a different suggestion why negative information is more persuasive, not necessarily incompatible with the explanations just mentioned. My hypothesis is that the greater persuasive effect of negative information has to do with our preferences rules and the amount of information that is required to reach a conclusion. It seems reasonable to assume that we have various prerequisite-free defaults of the form: 'Generally, avoid bad consequence $X$' ($\rightsquigarrow \neg X$) or 'Generally, seek good consequence $Y$' ($\rightsquigarrow Y$). Negative information is in conflict with these rules, whereas positive information is not. To avoid such conflict, we are induced to act.

In the positive frame, an act for attaining a certain positive outcome is described. For example, the agent is told that if she performs breast self-examination (BSE), she has a higher chance of finding a tumour in the early stages of the disease. This information is perfectly compatible with her (presumed) preference rules to avoid sickness and, when ill, to get well as soon as possible.

This harmony between information and preference rules is important. Since performing BSE is a sufficient, but not necessary condition for good health, it does not (immediately) follow that not performing BSE is a bad thing. That is, the frame does not provide (direct) information that not performing BSE either leads to a harm or makes it impossible to attain a good consequence (e.g., good health). Hence, both performing BSE and doing nothing are compatible with our preference rules.

In the negative frame, on the other hand, the agent is presented with a situation in which a negative outcome obtains. For example, she is told that if she does not perform BSE, she has a lower chance of finding a tumour. This suggests that not performing BSE leads to a lower chance of being in good health, a consequence she (presumably) wants to avoid. Thus, there is a direct conflict with the agent's preference rules to avoid harmful consequences and seek good consequences. As a result, the negative frame contains more direct information to induce the agent to perform BSE.

To illustrate this, consider the following example about the breast cancer prevention campaign. Assume that the positive frame triggers the following associations:

(1) $BSE(x) \rightarrow TUMOUR(x)$: If you perform BSE, you have a higher chance of finding a tumour in the early stages of the disease

(2) $TUMOUR(x) \rightarrow HEALTHY(x)$: If you find a tumour in the early stages, you have a higher chance of getting well

The negative frame, on the other hand, triggers the following associations:

(3) $\neg BSE(x) \rightarrow \neg TUMOUR(x)$: If you do not perform BSE, you have a lower chance of finding a tumour in the early stages of the disease

(4) $\neg TUMOUR(x) \rightarrow \neg HEALTHY(x)$: If you do not find a tumour in the early stages of the disease, you have a lower chance of getting well

In both cases, I assume that the agents have the following (prerequisite-free) preference rule:

(5) $\rightsquigarrow HEALTHY(x)$: Generally, one wants (a high chance) to get well

Suppose the positive frame is presented to an agent ('agent 1'). It seems reasonable to assume that the frame first triggers association $(1)$, then $(2)$ and then $(5)$. At $t = 3$ of the decision path, she (per default) concludes $HEALTHY(x)$, and her (prioritised) extension $E_3$ is $Th(\{BSE(x) \rightarrow TUMOUR(x),\ TUMOUR(x) \rightarrow HEALTHY(x),\ HEALTHY(x)\})$. This is not enough to decide on the target proposition (i.e., to perform or not perform BSE). Hence, if she feels so inclined, she has to move on to $t = 4$ to process more information in order to be able to make a decision. Since it is questionable whether the positive and negative frame are logically equivalent, this additional information may or may not be identical to $(3)$ and $(4)$. For the purposes of this thesis, this does not really matter.

What is more interesting is the decision process of an agent presented with the negative frame ('agent 2'). It seems reasonable to assume that the negative frame first triggers $(3)$, then $(4)$ and then $(5)$. Hence, at $t = 3$, the agent tentatively concludes $HEALTHY(x)$ as well. Her (prioritised) extension $E_3$ then is $Th(\{\neg BSE(x) \rightarrow \neg TUMOUR(x),\ \neg TUMOUR(x) \rightarrow \neg HEALTHY(x),$ $HEALTHY(x)\})$. Contrary to the extension of agent 1, the extension of agent does induce her to conclude $BSE(x)$, by modus tollens.[44]

That is, her default preference $HEALTHY(x)$ can only be realised if it is not the case that $\neg TUMOUR(x)$ obtains. This again requires the agent to avoid that $\neg BSE(x)$ obtains. She can only do so by performing BSE.

Thus, the agent presented with the negative frame is already able to draw a conclusion (i.e., to perform BSE) at $t = 3$ of the decision path, whereas the agent presented with the positive frame requires more information (and effort) to reach the same conclusion. This can explain why the former is more persuasive. Provided that an agent presented with the positive frame *is* able to derive the information conveyed by the negative frame, it takes her two more steps to come to the same conclusion.

Whether this representation does justice to what is really going on in goal framing, is an open question. I am not aware of any studies relating goal framing to information processing that can vindicate or refute this hypothesis. One interesting question is whether subjects are able to perceive informational differences between positive and negative goal frames. This is an open issue. O'Keefe (2007) acknowledges the lack of logical equivalence in goal framing, but argues that "it is probably unwise to assume that the difference between these two conditionals is readily apparent to casual observers" (ibid., p.154). Corner and Hahn (2010) argue that this view is untenable and that the "differences between the arguments are more than stylistic" (ibid., p.160).

One possible set-up for an experiment to test whether there is an informational difference between goal frames is the following. Take four groups of subjects. The first group is presented with a positive conditional, the second with its negative

[44] Note that $(1) - (4)$ are represented as 'regular' implications in this example. If they are taken to be default rules as well, the conclusion $BSE(x)$ no longer follows. This is because 'defeasible' modus tollens (i.e., on default rules) is not a valid inference in Reiter's logic. To overcome this problem, one could use a more comprehensive underlying logic, such as (a variant of) Circumscription Logic. See Bastiaanse and Veltman (forthcoming, pp.12-13) for more about this.

counterpart, the third with a positive bi-conditional, and the fourth with its negative counterpart. If all groups are given enough decision time to process all available information, one would expect that groups 3 and 4 come (roughly) to the same conclusions as the frames are presumably descriptively equivalent. Whether groups 1 and 2 come to the same conclusion depends on whether the frames are perceived as informationally equivalent. Even though this experiment does not focus on 'casual observers', it could provide us with some hints about the relationship between information, preference rules and decision-making in goal framing.

One problem for the explanation of goal framing given above is that quite a few studies and experimental results question the use of modus tollens in human reasoning. As it turns out, people tend to use modus tollens considerably less often than modus ponens, and furthermore are much more inclined to doubt the validity of a modus tollens inference.[45]

This led Rips (1994), for example, to suggest that, contrary to modus ponens, our reasoning system does not contain a direct rule for modus tollens. As a result, "subjects would have to derive the conclusion of [a modus tollens argument] by means of an indirect proof" (ibid., p.178). This derivation takes time and cognitive effort to execute, which can explain why we tend to use modus tollens less often and in many cases do not (immediately) see its validity.

However, even though modus tollens may not be the best way of representing the decision process, the underlying idea about the relationship between goal framing and information processing is worth delving into. A nice feature of representing goal framing in this way, is that it can account for the fact that positive frames are *less* persuasive than negative frames. That is, positive frames do not lead to a different decision than negative frames (as is the case with the other types of framing), but people are less inclined to perform one and the same act (e.g., BSE). I attribute this to the larger amount of information required by the positive frame to yield a conclusion.

Other nice features include that the same procedure can be used for all examples of goal framing, and hence that one does not have to draw on a vague or contrived notion of implicit risk (as Prospect Theory does). As argued in chapter 1, this notion of risk may play a role in some cases of goal framing but certainly not in all. Furthermore, the explanation of goal framing fits the explanations provided for the other types of framing, as they all revolve around the notion of focal information

[45] See Evans, Newstead and Byrne (1993, p.46) for a discussion and an overview of some experimental results on the use of modus tollens in human reasoning.

triggered by a frame.

Thus, even though more work has to be done, the model provides an interesting starting point for representing the goal framing decision process.

## 3.3 Associations and Rationality

In the previous section, I have shown that the model is able to meet two out of four of the desiderata formulated at the start of this chapter: the ability to accommodate the various types of framing and the various experimental results discussed in the previous chapter, such as the elaboration effect and lack of unanimity. How about the other two desiderata, the ability to do justice to the wide array of information involved in decision making and our nature as bounded rational agents?

The first one is quite straightforward. At the heart of the model lies the idea that when we are presented with a decision problem, we embark on a decision path that leads us past the propositions of a set of associations $X$. The model itself does not put any restrictions on this set: it can include any piece of information that springs to mind in the decision-making process. In the examples I have provided, I have mainly used facts that follow immediately from the descriptive content of the decision problem (e.g., the amount of lives saved by a medical programme) and some preference rules that take these facts as their input (antecedent). However, as argued before, our set of associations may contain information pertaining to the semantic properties of the words used, implicitly observed behavioural regularities, contextual factors etc as well. Therefore, if required, the model allows for a more detailed and complex representation of the decision process and the information that is processed.

The final desideratum, i.e., the bounded rational nature of human cognition, is also deeply entrenched in the model. The non-monotonic notion of entailment, the use of defeasible preference rules and the sequential information process are all tailored to express the fact that decision making is a ‘hard-fought’ process, confined by the information, time and cognitive resources that are available to the agent. Furthermore, the possibility of terminating the decision process before the end of the decision path has been reached allows the model to express the fact that many decisions are not ‘optimal’ outcomes information-wise. Rather, they are the result of a circumstantial process in which an agent can be satisfied with a decision that is ‘good enough’ given the context she is in.

As a result, the conclusion that the framing effect can only arise when the beliefs of the decision maker are (implicitly) inconsistent or when some other rationality principles are violated, no longer holds. The framing effect may very well arise when we are dealing with fully consistent belief sets.

Consider, for example, the theory $\Delta = (D, W, <)$ at the final stage of the decision path of a decision maker presented with the A / B choice of the Asian disease problem: $W = \{SAVE(p_a),\ \neg SAVE(p_b),\ DEATH(p_a),\ \neg DEATH(p_b)\}$ and $D = \{SAVE(x)\ \wedge\ \neg SAVE(y) \rightsquigarrow PREF(x,y),\ DEATH(x)\ \wedge\ \neg DEATH(y) \rightsquigarrow\ PREF(y,x)\}$. Due to the use of default rules, whose application can be defeated when new information becomes available, the agent can consistently belief all formulas in $\Delta$. The presumed inconsistency, taken to be inevitable by many authors (including Gold and List) has been transformed into two mutually exclusive classical extensions the agent's theory $\Delta$ gives rise to (one containing $PREF(p_a, p_b)$, the other $PREF(p_b, p_a)$).

However, these mutually exclusive extensions do not have any repercussions for the rationality of the agent's belief set. Rather, it indicates that the agent cannot, given her current information state, make a choice between the two conflicting proposition, unless one assumes, as I do here, that she finds one preference rule more compelling than the other. If this preference is added by means of a priority relation $<$, one unique preferred extension remains, containing either $PREF(p_a, p_b)$ or $PREF(p_b, p_a)$.

As a result, none of the four rationality conditions outlined by Gold and List have to be violated: our belief set can be complete, weakly and strongly consistent, and deductively closed, while at the same time giving rise to the framing effect. This does not mean, however, that the agent's belief set is necessarily consistent either. The belief set can very well be (implicitly) inconsistent or the agent can violate principles such as completeness or deductive closure. What happens in such cases is that the decision process will be exactly the same as in the consistent case, except that now there may, eventually, come a point on the decision path in which the agent becomes aware of the inconsistency. It seems likely that the agent will then defer her judgement, and will rethink her beliefs.

Thus, the model presented here shows, just as has been argued before, that the issue of (ir)rationality is a much more complex one than is often assumed. Irrational behaviour is not necessarily related to the *occurrence* of phenomena such as the framing effect, but rather to the underlying rationale for this occurrence. That is, I have argued that one must look at the underlying reasons or process that led the decision maker to be satisfied with a certain partial-information conclusion.

As a consequence, the principle of description invariance or extensionality (with which this thesis started out) may not hold in partial-information decisions. But this does not need to have any (negative) normative implications. That is, because two descriptions that are fully equivalent content-wise may differ significantly with respect to the accessibility of this content (the order in which the pieces of information are triggered), there are circumstances in which the agent is normatively justified ('has good reasons') to make different decisions under equivalent frames. Hence, right decisions need not be invariant to differences in descriptions. If the agent decides to process all information conveyed by a proposition, on the other hand, the principle *is* predicted to hold for rational agents. In this case, the agent has judged that differences in accessibility are not significant enough to stop the decision process early, and hence only takes factors related to the content of the proposition into account, not factors related to the presentation of this content.

As a result, the principle of description invariance is not as "normatively essential" as Tversky and Kahneman take it to be (Tversky and Kahneman 1986, S251). This thesis shows that whether the principle holds or not is not (necessarily) related to whether an agent makes the decisions she *should* make. It can be violated in rational decisions, and it can hold in irrational decisions.

This may also have consequences for Tversky and Kahneman's conclusion that "the normative and the descriptive analyses of choice should be viewed as separate enterprises" (ibid., S275). The relationship between normative and descriptive approaches in studying human decision making is a vexed topic, and it is well beyond the scope of this thesis to assess these consequences in detail.[46] However, it is interesting to note that the model presented here does justice to various empirical results about actually observed decision-making behaviour, but at the same time allows for a normative assessment of this behaviour through the notion of 'satisfaction' and the reasons a decision maker has for being satisfied with her decision.

## 3.4 Conclusion

In this chapter, I have presented a new model for the framing effect. The model combines a non-monotonic logic with the notion of a decision path, and adds a mechanism for resolving conflicts among preference rules. In this way, the model is able to accommodate the four desiderata outlined at the beginning of the chapter, which were based on the insights about the framing effect that have been gained in

[46] See Elqayam and Evans (2011) and Oaksford and Chater (2007) for two (opposing) views on this issue.

this thesis.

By making use of a set of associations triggered by a decision problem, the first desideratum can be incorporated, a broad notion of 'information' conveyed by a frame, drawing on various sources. Secondly, by using decision paths and defeasible preference rules, the underlying decision process of framing, as put forward in chapter 2, can be represented. In this way, all types of framing distinguished before, risky choice, attribute, goal and topic framing, can be represented in a uniform way.

Thirdly, by using a non-monotonic logic, the model is able to represent the process of assessing new information. In this way, I showed how different agents can make different decisions based on different amounts of information. This allows the model to explain for the lack of unanimity among subjects and the situations in which the framing effect is observed to attenuate (the 'elaboration effect'). Finally, by giving the conclusions we draw a tentative status and by pointing out their contextual character (based on the information that is available) the model is able to avoid the conclusion that the framing effect is irrational 'come what may'. Rather, I argued, the (ir)rationality of the framing effect depends on the reasons we have for being satisfied with a partial-information conclusion.

# Conclusion

In this thesis, I have investigated what the framing effect is (chapter 1), how the underlying decision process works (chapter 2), and how it can be represented in a model (chapter 3). For a chronological summary of this thesis, the reader is referred to sections 1.5, 2.5 and 3.4. Here, I will discuss the most important results and their consequences, point out some shortcomings and do some suggestions for future research.

## Results

*Characterisation of Framing*

In this thesis, it showed that framing is a much more diverse phenomenon than is usually assumed. There are various types of framing, with various characteristics. As a result, the typical (sloppy) characterisation of framing in terms of (logical) equivalence does not suffice. In this thesis, I showed that the presumed equivalence of different frames is not so straightforward. Furthermore, even if two frames are fully equivalent content-wise, there can still be considerable differences with respect to the accessibility or distinctness of this content. Therefore, attention should shift from the descriptive content of a frame to a much more elaborate notion of information or association, and to the accessibility of the different pieces of information conveyed by a frame. This information can come from various sources, and the decision maker takes various (contextual) factors into account. In doing so, she extends well beyond the 'objective', 'pertinent' or 'aloof' attitude towards decision problems that is assumed in standard decision theory.

*Valence Framing and Topic Framing*

In the literature, valence framing and topic framing are often treated as two distinct phenomena. The former is mainly studied in psychology, whereas the latter is mainly studied in political science. In this thesis, I have used literature from both disciplines, as well as other disciplines such as economics and cognitive science, and showed that the two phenomena are more alike than has generally been recognised. They do not make up a dichotomy (equivalent / non-equivalent), but rather form a continuum. Both valence framing and topic framing can be said to be the result of a similar underlying decision process. By making use of the notions of informational (non)-equivalence and the ‘logical’ and ‘empirical’ condition of the framing effect, the similarities and differences between the two types of framing can be made precise.

*Existing Models of Framing*

Three different models of the framing effect have been discussed: Prospect Theory, the Decision Model without Extensionality and the Logic of Path Dependence. This thesis has shown that all three models face serious problems in providing a full account of the framing effect and in accommodating various experimental results, such as the ‘elaboration effect’. I have argued that this is mainly due to a wrong (or too narrow) focus (e.g., on reference points, risks or ‘good news’) and a too static or one-sided explanation of the underlying decision process.

*Underlying Decision Process*

This thesis has provided an extensive characterisation of the underlying decision process of framing. Different frames trigger (sometimes different) pieces of information, in a different order. Decision makers do not process these pieces of information all at once, but rather take the most focal ones into account first. We are used to deal with incomplete theories of the world, and therefore make use of various default rules that allow us to draw (tentative) conclusions based on little information (i.e., we ‘read between the lines’). As a result, the most focal associations are often enough for us to reach a first conclusion about the decision problem. In many cases, this first conclusion is ‘good enough’, and hence we do not have an incentive to take less focal pieces of information into account as well. Thus, by changing the order in which information is processed (framing), one is able to influence the decision process.

*Rationality*

Framing has long been regarded as the result of human irrationality (or the irrationality of the agents prone to it). Both Prospect Theory and the Logic of Path Dependence draw heavily on this assumption. In this thesis, I have argued that it is mistaken to answer the question whether framing is irrational with a clear ‘yes’ or a clear ‘no’. The fact that human beings do not live up to some highly idealised notion of rationality does not justify the conclusion that we are irrational. One should take into account the information available to the agent, the circumstances she is in and the reasons she has for being satisfied with a partial-information conclusion, before being able to draw insightful conclusions about the (ir)rationality of certain behaviour. An agent may have very good reasons to take only partial information into account. For example, since speakers tend to use specific formulations more often in specific situations, a speaker may unwittingly 'leak ' information by choosing one formulation rather than another. This can be a proper justification for the decision maker to attach more weight to this information. Furthermore, time constraints and the fact that processing less readily available information can be cognitively demanding can be good reasons as well for accepting a tentative conclusion, depending on the situation the decision maker is in.

This also means that the principle of description invariance or extensionality is not the 'cornerstone' of rationality it is often taken to be. One can have very good reasons for violating the principle, as well as very bad reasons for holding on to it.

*New Model of Framing*

Finally, this thesis presented a new model for the framing effect. This model incorporates the above results in order to do justice to the diversity of framing, the role played by different pieces of information with different degrees of accessibility, the underlying decision process that is put forward and the nature of human beings as bounded rational agents. In this way, the model is able to provide an elaborate account of framing, and is in line with various experimental results (such as the 'elaboration effect') the other models struggle with. Furthermore, the model is able to accommodate goal framing, a type of framing that has been studied considerably less than other types of framing. Only a few, schematic psychological explanations are around for this phenomenon, and none of the existing frameworks can deal with it in a satisfactory, uniform way. The model presented here handles all instances of goal framing in a similar vein, and suggests a link between negative information, preference rules and our ability to draw conclusions that is worth probing into.

## Consequences, Shortcomings and Suggestions

*Reason-Based Accounts vs. Utility-Based Accounts*

In the introduction, it showed that standard decision theory, based on expected utility, fails to account for the framing effect. Morphological differences between equivalent frames are deemed irrelevant by rational agents in making a decision, and as a result, the framing effect is predicted not to occur.

Prospect Theory and Bourgeois and Giraud's Decision Model without Extensionality adapt the way in which the expected utility of the different prospects is perceived such that the specific formulations used do become relevant. Both models, as well as standard decision theory, however, take decision making to be a matter of utility computation. That is, they assume that a frame is taken into account as a whole and that this yields a certain utility value. How this value comes about and what role the specific content of the frames play in this computational process is not specified.

On the other hand, Gold and List's Logic of Path Dependence, as well as the model presented in this thesis, provide a reason-based account for the framing effect. Frames are no longer viewed as 'monolithic' choice options, but rather as sets composed of various pieces of information that are taken into account by the agent. Decision making is no longer a matter of utility maximisation, but of drawing conclusions based on the information at hand.

These two approaches, utility-based (or value-based) and reason-based, have typically been applied to distinct domains. Shafir, Simonson and Tversky (1993), for example, say the following: "Reason-based analyses have been used primarily to explain non-experimental data, particularly unique historic, legal and political decisions. In contrast, value-based approaches have played a central role in experimental studies of preference and in standard economic analyses" (ibid., p.12).

This does not mean that both approaches are only useful for their own domains. It is clear that reason-based accounts can be used for explaining experimental studies of preference, such as the framing effect, as well. It would be interesting to investigate how exactly these two approaches are related. On the one hand, it can be argued that considerations and calculations of utility can be represented as reasons relevant in decision making, and thereby be incorporated in a reason-based account (ibid., p.35). Bourgeois and Giraud, however, show that the reverse is also possible. That is, through the notion of good news and the resulting higher expected utility, the way utility values come about "can be given an interpretation in terms of arguments", they write (Bourgeois-Gironde and Giraud 2009, p.394).

*Goal Framing*

In the literature, the greater persuasiveness of negative information and negative consequences has been described extensively, and this can explain why negative goal frames tend to induce people to act more than positive goal frames do. The model presented in this thesis is able to accommodate goal framing by arguing that negative information is in direct conflict with our preferences rules, whereas positive information is not. Hence, in the former case, action is required to resolve this conflict, whereas in the latter, no such action is needed.

For this to work, assumptions had to be made about the deductive closure of our belief set. More specifically, the agent draws the conclusion to perform an action by applying modus tollens. As has been explained, it remains to be seen whether this is really what is going on, and more research is needed to vindicate (or refute) this claim. I have done some suggestions for testing the relationship between the information conveyed by goal frames and the effect on decision making (see p.77).

*Associations and Satisfaction*

The model presented in this thesis uses two ingredients that are interesting to develop further. First of all, frames are taken to trigger various associations. However, the question how this process works, how these associations come about and are related, has been largely put aside. As has been indicated, there are a few theories of associative thought around, and it is worthwhile to find out how these can be integrated (formally) in the model.

Secondly, the model allows agents to be 'satisfied' with a partial-information conclusion. Here, a link has been made with the notion of 'satisficing', but again it would be interesting to see how this can be incorporated in the model.

*Persuasion, Ubiquity, Game Theory*

Finally, the results of this thesis give rise to two interesting follow-up questions. I have tried to answer the questions what the framing effect is and how the underlying decision process works. It would be interesting to investigate why framing is so ubiquitous in our everyday lives and how it can be used to persuade people.

This is linked to the 'empirical condition' of the framing effect. Recall that in this thesis, a distinction has been made between a logical condition and an empirical condition. The logical condition pertains to the order in which we process available information (decision path) and to our ability to draw (tentative) conclusions based on partial information. The empirical condition, on the other hand, pertains to the actual, empirical circumstances under which the decision maker is willing to accept a framed message and to follow the accompanying decision path.

One can use game theory to investigate under which conditions an agent takes a message (or frame) to be credible (i.e., in what situations successful communication occurs and what situations it does not). However, standard game theory predicts that such communication only occurs when the preferences of both the message sender and the decision maker (message receiver) are aligned.[47]

The rationale behind this is as follows. Given that many people are, in one way or another, familiar with the phenomenon of framing, one would expect that people pay attention to it in decision making. Hence, it would seem to be fairly difficult to 'abuse' framing for one's own ends. That is, assuming that decision makers are aware of the potential 'dangers' of framing, it seems hard to use framing to induce people to make different decisions than they otherwise would have made. Only when the interests of message sender and message receiver coincide, is the latter willing to accept the message of the former.[48]

Yet, this prediction of standard game theory is at odds with the observed ubiquity of framing. For some reason, framing does seem to be 'profitable' and effective. Hence, a similar problem as touched upon in the introduction presents itself. Just as standard decision theory fails to accommodate the existence of the framing effect, so does standard game theory fail to accommodate its persistence as a persuasive tool.

Standard decision theory and game theory share many underlying assumptions, for example with respect to perfect rationality, full information and computational capacities of the agents involved. Therefore, the insights gained in this thesis about framing and decision making may be helpful to account for the relation between framing and communication as well.[49]

For example, it may sometimes require a lot of effort to compute the optimal solution in a game. If one assumes that not all players are able or willing to take this effort, someone may frame a message hoping others do not see through it. A decision maker may assess the message based on the 'prima facie' information it contains (i.e., the information that is directly accessible), rather than its full information. In such situation, a message may seem credible to her, but may contain unfavourable information that could have changed her mind if it would have been more focal.

Furthermore, as seen in this thesis, people are often dealing with an incomplete theory of the world and use default rules to fill the gaps. This suggests another deviation from the standard game theoretic assumptions, namely that agents have complete information about the game that is played. It seems much more realistic to assume that the agents' information is incomplete. For example, in the case

[47] See, for example, Rabin (1990), Farrell (1993) and Stalnaker (2006).

[48] See also the literature on 'Dutch books', such as Vineberg (2011).

[49] See van Rooij and de Jaegher (2013) for a discussion of the assumptions underlying game theory, and the consequences of altering them.

of judging the credibility of a message, the receiver may not know the payoffs of the sender with certainty, and hence may not know whether their preferences are aligned.

These suggestions are all very informal and schematic. Altering the assumptions of standard game theory may lead to a better match between the predictions of game theory about successful communication and the observed practice in which communication and persuasion are ubiquitous. However, much more research is needed to test the feasibility of these suggestions and to develop a formal framework that can accommodate them.

# Bibliography

Ahn, D.S. and H. Ergin (2010). 'Framing Contingencies'. In: *Econometrica* 78 (2), pp.655–695.

Almashat, S. et al. (2008). 'Framing Effect Debiasing in Medical Decision Making'. In: *Patient Education and Counseling* 71 (1), pp.102–107.

Arrow, K.J. (1982). 'Risk Perception in Psychology and Economics'. In: *Economic Inquiry* 20 (1), pp.1–9.

Barnes, J.D. and J.E. Reinmuth (1976). 'Comparing Imputed and Actual Utility Functions in a Competitive Bidding Setting'. In: *Decision Sciences* 7 (4), pp.801–812.

Bastiaanse, H. and F. Veltman (forthcoming). 'Making the Right Exceptions'. Institute for Logic, Language and Computation, University of Amsterdam.

Beach, L.R. et al. (1996). 'Differential versus Unit Weighting of Violations, Framing, and the Role of Probability in Image Theory's Compatibility Test'. In: *Organizational Behavior and Human Decision Processes* 65 (2), pp.77–82.

Bourgeois-Gironde, S. and R. Giraud (2009). 'Framing Effects as Violations of Extensionality'. In: *Theory and Decision* 67 (4), pp.385–404.

Brewka, G. and T. Eiter (2000). 'Prioritizing Default Logic'. In: *Intellectics and Computational Logic. Papers in Honor of Wolfgang Bibel*. Ed. by S. Hölldobler. Vol. 19. Applied Logic Series. Dordrecht: Springer, pp.27–45.

Camerer, C. (2005). 'Three Cheers—Psychological, Theoretical, Empirical—for Loss Aversion'. In: *Journal of Marketing Research* 42 (2), pp.129–133.

Corner, A. and U. Hahn (2010). 'Message Framing, Normative Advocacy and Persuasive Success'. In: *Argumentation* 24 (2), pp.153–163.

Cyert, R.M. and J.G. March (1963). *A Behavioral Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall.

Davis, M.A. and P. Bobko (1986). 'Contextual Effects on Escalation Processes in Public Sector Decision Making'. In: *Organizational Behavior and Human Decision Processes* 37 (1), pp.121–138.

Delgrande, J.P., T. Schaub and H. Tompits (2000). ‘A Compilation of Brewka and Eiter’s Approach to Prioritization’. In: *Logics in Artificial Intelligence. European Workshop, JELIA 2000 Málaga, Spain, September 29 – October 2, 2000 Proceedings*. Ed. by M. Ojeda-Aciego et al. Vol. 1919. Lecture Notes in Computer Science. Berlin: Springer, pp.376–390.

Druckman, J.N. (2011). ‘What’s It All About? Framing in Political Science’. In: *Perspectives on Framing*. Ed. by G. Keren. New York, NY: Psychology Press, pp.279–301.

Elqayam, S. and J.S.B.T. Evans (2011). ‘Subtracting “Ought” from “Is”: Descriptivism versus Normativism in the Study of Human Thinking’. In: *Behavioral and Brain Sciences* 34 (5), pp.233–248.

Entman, R.M. (1993). ‘Framing: Toward Clarification of a Fractured Paradigm’. In: *Journal of Communication* 43 (4), pp.51–58.

Evans, J.S.B.T., S.E. Newstead and R.M.J. Byrne (1993). *Human Reasoning: The Psychology of Deduction*. Exeter: Lawrence Erlbaum Associates.

Farrell, J. (1993). ‘Meaning and Credibility in Cheap-Talk Games’. In: *Games and Economic Behavior* 5, pp.514–531.

Fawcett, T.W. et al. (2014). ‘The Evolution of Decision Rules in Complex Environments’. In: *Trends in Cognitive Sciences* 18 (3), pp.153–161.

Feather, N.T., ed. (1982). *Expectations and Actions: Expectancy-Value Models in Psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Frijda, N.H. (1986). *The Emotions*. Cambridge: Cambridge University Press.

Ganzach, Y. and N. Karsahi (1995). ‘Message Framing and Buying Behavior: A Field Experiment’. In: *Journal of Business Research* 32 (1), pp.11–17.

Geurts, B. (2013). ‘Alternatives in Framing and Decision Making’. In: *Mind and Language* 28 (1), pp.1–19.

Ghirardato, P. and M. Marinacci (2001). ‘Risk, Ambiguity, and the Separation of Utility and Beliefs’. In: *Mathematics of Operations Research* 26 (4), pp.864–890.

Giraud, R. (2004). *Framing under Risk: Endogenizing the Reference Point and Separating Cognition and Decision*. Ref. bla04090. Université Panthéon-Sorbonne.

Gold, N. and C. List (2004). ‘Framing as Path Dependence’. In: *Economics and Philosophy* 20 (2), pp.253–277.

Grice, H.P. (1975). ‘Logic and Conversation’. In: *Syntax and Semantics. Speech Acts*. Ed. by P. Cole and J.L. Morgan. Vol. 3. New York, NY: Academic Press, pp.41–58.

Halter, A.N. and G.W. Dean (1971). *Decisions under Uncertainty, with Research Applications*. Cincinnati, OH: South-Western Publishers.

Holyoak, K.J. and P. Thagard (1989). ‘Analogical Mapping by Constraint Satisfaction’. In: *Cognitive Science* 13, pp.295–355.

Homer, P.M. and S. Yoon (1992). ‘Message Framing and the Interrelationships among Ad-Based Feelings, Affect, and Cognition’. In: *Journal of Advertising* 21 (1), pp.19–33.

Hughes, T. (1970). ‘Two Legends’. In: *Crow: From the Life and Songs of the Crow*. London: Faber & Faber.

Jeffrey, R.C. (1983). *The Logic of Decision*. Chicago, IL: University of Chicago Press.

Johnson-Laird, P.N. (1968). ‘Shorter Articles and Notes the Interpretation of the Passive Voice’. In: *Quarterly Journal of Experimental Psychology* 20 (1), pp.69–73.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Kahneman, D., J.L. Knetsch and R.H. Thaler (1991). ‘Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias’. In: *Journal of Economic Perspectives* 5 (1), pp.193–206.

Kahneman, D. and A. Tversky (1979). ‘Prospect Theory: An Analysis of Decision under Risk’. In: *Econometrica* 47 (2), pp.263–291.

— (1984). ‘Choices, Values, and Frames’. In: *American Psychologist* 39 (4), pp.341–350.

Kinder, D.R. and L.M. Sanders (1996). *Divided by Color: Racial Politics and Democratic Ideals*. Chicago, IL: University of Chicago Press.

Kühberger, A. (1995). ‘The Framing of Decisions: A New Look at Old Problems’. In: *Organizational Behavior and Human Decision Processes* 62 (2), pp.230–240.

— (1998). ‘The Influence of Framing on Risky Decisions: A Meta-Analysis’. In: *Organizational Behavior and Human Decision Processes* 75 (1), pp.23–55.

Levin, I.P. and G.J. Gaeth (1988). ‘How Consumers are Affected by the Framing of Attribute Information Before and After Consuming the Product’. In: *Journal of Consumer Research* 15 (3), pp.374–378.

Levin, I.P., R.D. Johnson et al. (1986). ‘Framing effects in decisions with completely and incompletely described alternatives’. In: *Organizational Behavior and Human Decision Processes* 38 (1), pp.48–64.

Levin, I.P., S.L. Schneider and G.J. Gaeth (1998). ‘All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects’. In: *Organizational Behavior and Human Decision Processes* 76 (2), pp.149–188.

Levin, I.P., S.K. Schnittjer and S.L. Thee (1988). ‘Information Framing Effects in Social and Personal Decisions’. In: *Journal of Experimental Social Psychology* 24 (6), pp.520–529.

List, C. (2004). ‘A Model of Path-Dependence in Decisions over Multiple Propositions’. In: *American Political Science Review* 98 (3), pp.495–513.

Mahoney, L.J. (1977). ‘Early Diagnosis of Breast Cancer: The Breast Self-Examination Problem’. In: *Canadian Family Physician* 23, pp.91–93.

Marcus, G. (2008). *Kluge: The Haphazard Evolution of the Human Mind*. New York: Houghton Mifflin.

McCarthy, J. (1980). ‘Circumscription. A Form of Non-Monotonic Reasoning’. In: *Artificial Intelligence* 13 (1-2), pp.27–39.

McKenzie, C.R.M. and J.D. Nelson (2003). ‘What a Speaker’s Choice of Frame Reveals: Reference Points, Frame Selection, and Framing Effects’. In: *Psychonomic Bulletin & Review* 10 (3), pp.596–602.

Meyerowitz, B. and S. Chaiken (1987). ‘The Effect of Message Framing on Breast Self-Examination Attitudes, Intentions, and Behavior’. In: *Journal of Personality and Social Psychology* 52 (3), pp.500–510.

Miller, P.M. and N.S. Fagley (1991). 'The Effects of Framing, Problem Variations, and Providing Rationale on Choice'. In: *Personality and Social Psychology Bulletin* 17 (5), pp.517–522.

Nelson, T.E. and Z.M. Oxley (1999). 'Issue Framing Effects on Belief Importance and Opinion'. In: *The Journal of Politics* 61 (4), pp.1040–1067.

Oaksford, M. and N. Chater (2007). *Bayesian Rationality. The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Odhnoff, J. (1965). 'On the Techniques of Optimizing and Satisficing'. In: *The Swedish Journal of Economics* 67 (1), pp.24–39.

O'Keefe, D.J. (2007). 'Potential Conflicts between Normatively-Responsible Advocacy and Successful Social Influence: Evidence from Persuasion Effects Research'. In: *Argumentation* 21 (2), pp.151–163.

Pratto, F. and O.P. John (1991). 'Automatic Vigilance: The Attention-Grabbing Power of Negative Social Information'. In: *Journal of Personality and Social Psychology* 61 (3), pp.380–391.

Rabin, M. (1990). 'Communication between Rational Agents'. In: *Journal of Economic Theory* 51, pp.144–170.

Reiter, R. (1980). 'A Logic for Default Reasoning'. In: *Artificial Intelligence* 13, pp.81–132.

Resnik, M.D. (2008). *Choices. An Introduction to Decision Theory*. 8th ed. Minneapolis, MN: University of Minnesota Press.

Rick, S. (2011). 'Losses, Gains, and Brains: Neuroeconomics Can Help to Answer Open Questions about Loss Aversion'. In: *Journal of Consumer Psychology* 21, pp.453–464.

Rips, L.J. (1994). *The Psychology of Proof. Deductive Reasoning in Human Thinking*. Cambridge, MA: MIT Press.

Rooij, R. van and K. de Jaegher (2013). 'Argumentation with (Bounded) Rational Agents'. In: *Bayesian Argumentation. The Practical Side of Probability*. Ed. by F. Zenker. Dordrecht: Springer, pp.147–161.

Russo, J.E., V.H. Medvec and M.G. Meloy (1996). ‘The Distortion of Information during Decisions’. In: *Organizational Behavior and Human Decision Processes* 66 (1), pp.102–110.

Schneider, S.L. (1995). ‘Item Difficulty, Discrimination, and the Confidence-Frequency Effect in a Categorical Judgment Task’. In: *Organizational Behavior and Human Decision Processes* 61 (2), pp.148–167.

Shafir, E., I. Simonson and A. Tversky (1993). ‘Reason-Based Choice’. In: *Cognition* 49 (1), pp.11–36.

Sher, S. and C.R.M. McKenzie (2006). ‘Information Leakage From Logically Equivalent Frames’. In: *Cognition* 101 (3), pp.467–494.

Simon, H.A. (1955). ‘A Behavioral Model of Rational Choice’. In: *The Quarterly Journal of Economics* 69 (1), pp.99–118.

— (1956). ‘Rational Choice and the Structure of the Environment’. In: *Psychological Review* 63 (2), pp.129–138.

— (1997). *Models of Bounded Rationalilty. Empirically Grounded Economic Reason*. Vol. 3. Cambridge, MA: MIT Press.

Sniderman, P.M. and S.M. Theriault (2004). ‘The Structure of Political Argument and the Logic of Issue Framing’. In: *Studies in Public Opinion: Attitudes, Nonattitudes, Measurement Error, and Change*. Ed. by W.E. Saris and P.M. Sniderman. Princeton, NJ: Princeton University Press, pp.133–164.

Sniezek, J.A., P.W. Paese and F.S. Switzer (1990). ‘The Effect of Choosing on Confidence in Choice’. In: *Organizational Behavior and Human Decision Processes* 46 (2), pp.264–282.

Sobel, R.S. and S. Travis Raines (2003). ‘An Examination of the Empirical Derivatives of the Favourite-Longshot Bias in Racetrack Betting’. In: *Applied Economics* 35 (4), pp.371–385.

Stalnaker, R. (2006). ‘Saying and Meaning, Cheap Talk and Credibility’. In: *Game Theory and Pragmatics*. Ed. by A. Benz, G. Jäger and R. van Rooij. Basingstoke: Palgrave MacMillan, pp.83–100.

Stenning, K. and M. van Lambalgen (2008). *Human Reasoning and Cognitive Science*. Cambridge, MA: MIT Press.

Takemura, K. (1994). ‘Influence of Elaboration on the Framing of Decision’. In: *The Journal of Psychology* 128 (1), pp.33–39.

Taylor, S.E. (1991). ‘Asymmetrical Effects of Positive and Negative Events’. In: *Psychological Bulletin* 110 (1), pp.67–85.

Turnbull, E.M. (1978). ‘Effect of Basic Preventive Health Practices and Mass Media on the Practice of Breast Self-Examination’. In: *Nursing Research* 27 (2), pp.98–102.

Tversky, A. and D. Kahneman (1974). ‘Judgment under Uncertainty: Heuristics and Biases’. In: *Science* 185 (4157), pp.1124–1131.

— (1981). ‘The Framing of Decisions and the Psychology of Choice’. In: *Science* 211 (4481), pp.453–458.

— (1986). ‘Rational Choice and the Framing of Decisions’. In: *The Journal of Business* 59 (4), S251–S278.

Tykocinski, O., E.T. Higgins and S. Chaiken (1994). ‘Message Framing, Self-Discrepancies, and Yielding to Persuasive Messages: The Motivational Significance of Psychological Situations’. In: *Personality and Social Psychology Bulletin* 20 (1), pp.107–115.

Vineberg, S. (2011). *Dutch Book Arguments*. Version Summer 2011. The Stanford Encyclopedia of Philosophy. URL: `http://plato.stanford.edu/archives/sum2011/entries/dutch-book/`.

Wilson, D.K., R.M. Kaplan and L.J. Schneiderman (1987). ‘Framing of Decisions and Selection of Alternatives in Health Care’. In: *Social Behaviour* 2, pp.51–59.

Yared, F. (1999). ‘Path Dependence in Expected Inflation: Evidence from a New Term-Structure Model’. PhD thesis. Graduate School of Business, University of Chicago.