THE LONDON SCHOOL OF ECONOMICS
AND POLITICAL SCIENCE

Framing It Right? Normative representational standards for decision theory

by

Hadrien Mamou

A thesis submitted to the

Department of Philosophy, Logic and Scientific Method

of

The London School of Economics and Political Science

for the degree of

Doctor of Philosophy

London, October 15, 2024

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 56,416 words.

Abstract

Decision problems are non-exhaustive and selective representations that raise a number of philosophical issues. Framing effects are a case in point: Different ways of framing or representing a particular decision situation may lead to different choices, in violation of the Extensionality principle. Extensionality stipulates that different descriptions of the same outcome or option should lead to identical evaluations and choices. This dissertation examines what constitutes a correct representation of a decision situation for bounded agents. Taking a Moderate Humean perspective on rationality, I argue that the literature does not offer satisfactory standards of representation and fails to address a series of selection problems such as the ones raised by framing effects. Instead, I offer an instrumental account whereby agents represent decision situations in a way to achieve the ends they care about.

I examine two interpretations of irrational framing effects — as Extensionality violation and as unstable evaluation. I assess both principles and reject these two interpretations. Instead, I introduce the notion of evaluatively equivalent decision situations, and argue that illegitimate framing effects are inconsistent choices across evaluatively equivalent decision situations. The phenomenon illustrates one of the problems of selection that arise from assessing the rationality of decision-theoretic representations. I offer an alternative instrumental account of representation based on differences that the agent cares about, acknowledging that decision problems are partial, constructed, and value-driven representations.

Acknowledgements

I wish to warmly thank Richard Bradley for his unconditional support, for being a source of intellectual inspiration, and most importantly, for patiently coping with my unorthodox writing skills. Without him this dissertation would not exist.

I am grateful to The Department of Philosophy, Logic, and Scientific Method at the London School of Economics for being such a welcoming and understanding place for PhD students.

Special thanks to Julien Hauseux for making this document look decent, to Goreti Faria Da Costa for consistently reminding us that penguins do drink sea water, to Campbell Brown, Itay Rozen-Nissan, Johanna Thoma, as well as my fellow PhD students and friends.

Finally, I wish to thank my parents, my sister Roxane, Ben, Rafaella and Taliah for being there.

To Simon Mamou and Serge Le Diraison

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Representational standards for decision theory

Contents

1.1 Introduction

How should we correctly represent the decision problems we face? This dissertation is about this question and closely related philosophical issues of representation for decision theory.

Empirical decision theory attempts to explain and predict human agents' decisions. While economists traditionally do so by making hypotheses about rational choice and agency, psychologists examine the descriptive relevance and psychological plausibility of these hypotheses. Either way, these claims presuppose — sometimes implicitly — normative standards of rationality governing how one ought to make choices in order to be deemed rational. Indeed, economic models of decision-making presuppose a certain conception of rational choice. Analogously, psychological models may evidence biases of rationality that constitute systematic deviations from a pre-established normative perspective. The question of the legitimacy of these normative

standards is usually addressed by philosophers by examining our intuitions about the rationality of our choices in response to *decision problems*.

In informal terms, a problem may be a resistance or an obstacle to a project, to an obligation, or to the compliance with a norm. However, *decision problems* may also arise as a result of an opportunity rather than an obstacle: decision-makers typically face a decision problem when they are conflicted, in which case deliberation may help them settle for a certain course of action to overcome the conflict.

Decision problems come in various kinds. They may involve the evaluation of possible means to a given end; they may be imposed on us, such as when a referendum requires citizens to share their views on a political issue. They may be explicitly defined, as in the lab, where people are asked to choose between alternatives they did not construct. In real life, our environment may require us to make unanticipated choices: a friend may offer to go for a walk. All these are cases where you typically try to answer the question: "should I perform act A?", giving rise to a problem of decision.

One should distinguish the conceptual stage of elaboration of a formal decision problem (typically a matrix of options, states and outcomes), and the stage of decision-making and choice, which is typically the only part of decision theory and rational choice theory under normative scrutiny. For now, I broadly define the representation of a decision problem as an umbrella term: it includes the *agential activity of elaborating and representing a problem of decision, from an initial decision situation, to a formal decision problem expressed in decision-theoretic terms*. Issues of representation may thus involve any of the two stages of representation (initial and final), as well as the processes and mechanisms required to conceptually make sense of the activity of representation itself. While the initial stage describes the object of representation (or domain), the final stage describes the formal output of the representational activity (the representation).

With these distinctions in mind, the question of the nature of the normative rationality standards governing representation arises. *What constitutes an adequate or correct representation of a decision problem from the perspective of practical rationality?*

Prima facie, one may be tempted to apply the same standards to both decision-making and to the representation of decision problems. After all, if the possible representations of a particular decision problem are not unique, a given specification can be interpreted as the result of a choice of representation. According to Resnik (1987), the fact that a decision situation can be specified in several ways raises a philosophical issue for decision theory. Broadly put, choosing the best decision problem specification (or matrix representations) amounts to choosing between decision tables, which in itself is a second-order decision, requiring another problem specification, and thus leading to an infinite regress of decision problems. In chapter 3, an alternative approach will be considered, which distinguishes the standards of rationality overarching decision theory and rational choice theory on the one hand, and those governing the elaboration and representation of a formal decision problem on the other. On that view, representation and decision making

are ultimately different activities that practical rationality cannot treat on a par. In any case, the general question this dissertation attempts to answer can be defined as follows:

What are the standards of rationality against which the representations of a decision situation ought to be assessed?

This dissertation does not try to answer the research question exhaustively although it will address a series of issues that the question raises. In the remainder of this introduction, I will clarify the perspective of the dissertation, its scope, and the desiderata I set for a convincing philosophical answer to the research question. I will then briefly sketch the key claims I intend to defend, and give an overview of the main body of this work.

1.2 Framing effects as the starting point of the discussion

Let's start with framing effects, a phenomenon that has been much discussed in the economics and philosophy literature. A paradigmatic illustration of framing effects is Kahneman and Tversky's Asian disease experiment.[1] A disease is expected to kill 600 people. Subjects are being asked to make choices between two pairs of policies:

- First offer:
  - Program A: "200 people will be saved"
  - Program B: "there is a $1/3$ probability that 600 people will be saved, and a $2/3$ probability that no people will be saved"
- Second offer:
  - Program C: "400 people will die"
  - Program D: "there is a $1/3$ probability that nobody will die, and a $2/3$ probability that 600 people will die"

In practice, people usually prefer A to B and D to C. Traditionally, these results are treated as inconsistent, as programs A and C, and B and D are respectively deemed identical, if we only consider the number of deaths and the probabilities. However, in order to claim that agents are inconsistent, we must suppose that the two descriptions represent the same (or equivalent) decision problems. In particular, can we say that agents are inconsistent if "not die" and "be saved" are not equivalent descriptions? Is it the case that the subject only cares about deaths and lives as consequences? In order for us to say whether an agent is inconsistent in her choices, we must be able to say when she faces the same decision problem. A framing effect only arises when this is the case.

A decision-maker is subject to a framing effect when her choice is sensitive to changes of frames or descriptions of the same decision problem. Framing effects are traditionally considered

[1] Tversky and Kahneman 1981.

irrational on the ground that they violate some version of the extensionality or Invariance principle: different descriptions of the same decision problem should lead to the same choice. The options are then valued differently under different descriptions to the extent that they are not perceived as identical or equivalent in the relevant way.

The decision-theoretic literature has interpreted the Asian disease case in various ways. In particular, when assessing the normative status of extensionality, two issues should be disentangled. The first one is raised by Schick (1991) and Bermúdez (2020). Both authors question the extensionality principle by arguing that different descriptions of the same situation may legitimately lead to different choices if these descriptions stem from different ways of "seeing" that situation. Frames as well as understandings are thought of on the model of Frege's distinction between the object, or referent, and its mode of presentation, or sense.

An alternative interpretation of the Asian disease experiment and framing effects has been given by Dietrich and List (2016). Instead of viewing framing effects as the effects of changing frames of the same situation, they are conceived from the experimenter's point of view as different choices across equivalent but *distinct* choice situations with various observable properties. What Dietrich and List call a frame is a set of properties of the choice-situation that are salient to the decision maker. These properties may be features of a particular option, of a menu of options, or of the environment over and above options. Importantly, the set of properties objectively identified by the observer and to which the subject assigns weights when they are motivationally salient constitute the **context** of the decision situations. According to that interpretation, agents subject to irrational framing effects display unstable evaluation of properties: their evaluation of a property changes across contexts and thus violate a principle distinct from extensionality, called Invariabilism. As we will see, Invariabilism is the view that the weight of a property should be constant *across* situations.[2]

More generally, the common thread within this sparse literature are the following two questions: How stable should our representations of outcomes be? And then, how robust should our attitudes be to such representational shifts? Note that these two questions apply both to changes of representation for the same situation and across different situations.

This dissertation examines both interpretations of irrational framing effects — as extensionality violation and as Invariabilism violation. In each case, I will evaluate the underlying account of decision theoretic representation along with its normative representational standards. In the remainder of this chapter, I set up the discussion by defining the key notions of representation and context more precisely. I then present the perspectives and desiderata of the account defended in this thesis. After reviewing its main claims, I sketch the outline of the dissertation.

[2]To keep the exposition simple, Invariabilism is defined here in terms of properties and not reasons. As we will see in chapter 4, properties can be treated as the reasons for preferring a particular alternative to another, and the evaluation of a choice situation under a frame corresponds to the weighing of these reasons.

1.3 Representation

As we have seen, assessing the rationality of framing effects led philosophers to conceive of different notions of decision theoretic representations. I will examine a number of them, and in particular descriptions, understandings, frames, and specifications. What all of these notions have in common is that they refer to a target reality — that I call a **decision situation**, which may be conceived of in multiple ways, and identify principles of adequacy between beliefs, desires and the said representations of the decision situations. *For instance, Kahneman and Tversky's Invariance principle states that two descriptions of the same decision situation should lead to the same evaluations and choice.*[3] While a decision situation is typically described in the vocabulary of philosophy of action (usually by beliefs and desires), the formal representation of a decision situation is expressed in the vocabulary of decision theory. Although other versions of the Invariance principle exist, all attempt to answer the question of how to represent a decision situation correctly.

If a notion of representation is to play a role in decision theory, it ought to offer insights that standard belief-desire theories fail to offer. One way to determine that role is to ask *what aspects of the activity of rational representation cannot or should not be characterized only by beliefs and desires?*

Indeed, if one is to determine rational standards of representation, one should provide criteria of demarcation between correct and incorrect representations from the point of view of practical rationality. Suppose that these criteria happened to be reducible to a theory of rational beliefs and desires. Since decision theory characterizes rational decisions in terms of beliefs and desires, normative decision theoretic standards could be formulated without any reference to representation. In that sense representation would be theoretically neutral for practical rationality. Suppose on the contrary that representational standards are not reducible to decision theoretic standards for beliefs and desires. If one is to show that representations are not theoretically neutral, one should offer examples of situations where decision theory is either insufficient to determine the rational way of representing the decision situation; or where our intuitions about the correct way of representing the situation cannot be derived from decision theoretic principles.

An analogous question raised by confronting theories of representation is the evaluative neutrality of rational representations. Framing effects are a case in point: advocates of the Invariance principle argue that the way we represent a given decision situation should not affect the way we evaluate the decision problem representing it. Conversely, those rejecting the principle argue that representations have an incidence on practical evaluations. They thus offer counterexamples to the principle where our intuitions about the rational evaluation of a situation derive from our

[3] "*Invariance [is] an essential condition for a theory of choice that claims normative status is the principle of invariance: different repre- sentations of the same choice problem should yield the same prefer- ence. That is, the preference between options should be independent of their description. Two characterizations that the decision-maker, on reflection, would view as alternative descriptions of the same problem should lead to the same choice-even without the benefit of such reflection. This principle of invariance (or extensionality [Arrow 1982]), is so basic that it is tacitly assumed in the characterization of options rather than explicitly stated as a testable axiom.*" (Tversky and Kahneman 1986, p. S253)

intuitions about how to frame the problem adequately.

Let's take stock. Upon asking what counts as a rational representation of a decision situation, we have seen that framing effects raised several specific philosophical issues. First, how stable should our representations of outcomes be? And then, how robust should our attitudes be to such representational shifts? Settling on these questions leads philosophers to also address what I call the questions of the theoretic neutrality of representation for normative decision theory, and of the evaluative neutrality of representation for the rational decision-maker.

Now, in order to distinguish between legitimate and illegitimate changes of representation across situations, the literature often relies on the notion of legitimate sensitivity to context. As we will see, the terms of sensitivity and contexts are used in more than one sense by the authors. To clarify these notions, let's first have a look at different types of rationality requirements for decision theory and representation.

1.4 Distinguishing rationality standards

The philosophical issues raised by the definition of the context of representation calls for distinguishing different types of normative requirements. An externalist requirement imposes rationality constraints that are external to and independent from the agent's actual attitudes.[4] For instance, the rational requirement to take into account all available evidence before reaching a judgment is externalist, while requiring that agents hold judgments that are consistent with all the evidence they recognize constitutes an internalist requirement of rationality.

Though this first distinction bears on requirements of decision theory, the following one bears specifically on rational representation. Standards of representation usually involve requirements of adequacy between the domain (or target) represented and the object doing the representing (the representation).[5] However, different views can be held on the nature of these two objects and their relation. One may for instance take the domain to be the external world represented by the agent, and the representation to be the "intentional content" of the agent's attitudes when representing the situation.[6] The decision situation can then be interpreted as the agent's environment, and the rational agent is required to appropriately respond to changes in her external environment when representing the situation. Although such representational standards make reference to objects that are external to the agent, they could either involve internal or external requirements of adequacy: For example, an internalist may require that the agent represent identically two external environments about which she has identical beliefs, while an externalist could expect objectively identical environments to be identically represented.

[4] Bradley 2017, pp. 13–14.

[5] Frigg and Nguyen 2021, § 1.

[6] The expression "intentional content" refers here to the content of representation or mental states. The term intention will not be used in that sense in the remainder of the dissertation. See (Jacob 2023) for further distinctions between intentionality and intentions.

By contrast, other views do without a definition of decision theoretic representations referring to some external reality that the agent would represent, and without the associated requirements about how the content of representation should represent a thing that is to be represented. As we will see, this dissertation embraces this second kind of view about decision theoretic representations, and will define a decision situation as a particular pair of options, along with the set of features that are relevant for the agent. We can already note one reason for defining practical representations without appealing to an external reality. Some decision situations may be deemed identical from an external point of view although they are not identical from the agent's standpoint; this is the case if the situations differ in a way that actually matters to the agent. However, external standards of representation will still require agents to represent them identically.

These two distinctions about rationality standards — one for decision theory, the other for decision-theoretic representation — shed light on two ways of defining context and context-sensitivity.

1.5 Context, context-sensitivity, and decision situations

I will set aside the linguistic notion of context (i.e. the parts of speech which shed light on the meaning of a particular sentence[7]) and the associated issues. I believe there are at least two different ways of defining a context in decision theory, one objective and external, the other subjective and internal. The former distinguishes the object of agents' attitudes from objects of reality, and defines a decision context as the set of objective features of reality which affect the agent's decision in equivalent pairwise choices. This definition is intuitive since it corresponds to the empirical method of identification of a context in the lab: context-sensitivity occurs when the presence of a certain feature of reality, relevant or not, impacts the decision-maker's choice.

The standard explanation of context-sensitivity is that people are not *seeing* that two pairs of options are identical. Though intuitive, the empirical approach introduces an observer's viewpoint that is external and objective, and determines what the context is. It thus presupposes the identity between two decision problems which would have given rise to the same choice if it were not for that irrelevant context-sensitivity. As we will see, this first definition of context is useful for evidencing context-sensitive beliefs when the identity between decision problems is uncontroversial, as this is the case with equivalent probabilistic statements. This is also the approach that Dietrich and List (2016) use to define context-sensitivity. In other cases, this approach raises a number of difficulties that can be avoided by using an alternative definition which only appeals to the agent's own standpoint. Then, a pair of choices is illegitimately context-sensitive if it is affected by the presence of any feature that is judged irrelevant by the agent's own lights. If so, the only way for the empirical approach to conclude to illegitimate representation is retrospectively: after being informed by the observer's point of view, if the agent

[7] Merriam-Webster 2024.

agrees that the decision situations are identical, then the illegitimate context can be identified. For normative purposes, this second definition of context is more appealing on two counts.

First, introducing an observer's objective point of view is unnecessary to account for the said notion of context. Indeed, one may do away with it by simply distinguishing on the one hand the agent's point of view, that is informed by the presence of this context feature and its effect, from the uninformed point of view on the other hand. If the agent was epistemically mistaken about the said identity, one can only conclude to epistemic context sensitivity. If on the contrary she believed the options are indeed identical up to an additional feature and she regrets her choice after being informed, she may be charged with illegitimate evaluative context-sensitivity. Alternatively, she may maintain that these options ought to be valued differently and the so-called contextual feature is not irrelevant for the particular decision situation. This last case supports the fact that an observer's objective point of view is also insufficient to identify legitimate context sensitivity.

I will elaborate and justify this internalist definition of context further in chapter 4; for now, let's notice that a challenge for this approach is to account for our intuitions about irrational framing effects independently from any reference to some external reality that a decision problem is intended to represent. If successful, this approach is appealing as it is agnostic about general philosophical issues of representation (e.g. discussions about the nature of mental or scientific representations). Additionally, it would keep requirements of practical rationality on representation separate from normative requirements stemming from these issues of general philosophy.

To be able to discuss views involving different definitions of context and representation, let's provisionally define a **decision situation** as the set of all possible features that may affect the agent's choice between a particular pair of options. The decision situation thus defines the domain of representation. I will reserve the term **context** of an option pair to refer to the set of features of the situation whose presence may affect the evaluation of consequences.

We are now equipped with a broad enough framework to articulate the different kinds of representational issues discussed in this dissertation. The same decision situation may be represented in a variety of ways. Schick (1991) and Bermúdez (2020) both claim that the same decision situation can legitimately give rise to different representations in virtue of the very nature of these representations (understandings for the former and frames for the latter). By contrast, a *pair of distinct decision situations* may lead to equivalent decision problems, as in the Asian disease example or in cases of context-sensitive evaluations of a consequence. Finally, over and above the problems of extensionality and illegitimate context-sensitivity, issues of representation may arise because a particular decision situation — or domain of representation — is not clearly defined: as we will see in chapter 3, it is not always obvious how extended and finely grained the specification of a decision situation should be.

In what follows, I will specify the perspective and desiderata guiding this work, determine the

scope of the research question, and informally spell out how the account I will defend here answers that question. I will conclude with a brief outline of the dissertation.

1.6 Perspectives

The spirit of this work is driven by two perspectives, the first of which is *Moderate Humeanism*,[8] the view that "*rationality only constrains our attitudes indirectly by disallowing certain combinations of beliefs, desires and preferences.*"[9] Applied to representational issues, Moderate Humeanism may broadly be interpreted as the claim that conative attitudes play a fundamental role in rationally representing a decision problem. Though I will not make arguments in defense of Moderate Humeanism, I make claims compatible with it, or possibly strengthening arguments in its favor.

Second, this thesis is guided by an interest in standards of rationality for bounded creatures like us. I will thus discuss how bounded agents with limited time and cognitive capacities. There are several ways of defining rational boundedness. The expression may refer to cognitively limited agents and to what is ideally rational for them. Alternatively, bounded rationality may be synonymous with "qualified or less than ideal rationality": then the limitation no longer bears on the agent's cognition but on the concept of rationality, and one may ask what is qualifiedly rational for an agent (with or without cognitive limitations). Although this dissertation discusses both notions, it only refers to bounded rationality in the former sense.[10] I will thus discuss how bounded agents may rationally represent decision problems in a manner adequate to their goals and constraints. By contrast with an idealized conception of human rationality and agency, I hope to highlight the specificity of bounded rationality standards: human bounded rationality is not only a matter of limitations on cognition. Indeed,

(i) Differences between an ideally rational agent and a human bounded one are not only differences in degree, but also differences in kinds and functioning (as the notion of conflict, sophistication, or temporal coordination suggest). Thus, rational representations of decision problems mainly make sense for bounded agents. Analogously, one may hypothesize that there are ways of shaping and structuring decision problems that are specific to human bounded agents like us.

(ii) Bounded rationality standards involve a specific relationship to temporal, cognitive, and motivational resources that affect normative standards of representations. A decision theoretic representation should be evaluated against two distinct standards of rationality: the internal and the external point of view of rational assessment.

[8] The term was coined by Broome (1999).

[9] Bradley 2017, p. 14.

[10] Boundedness as qualified rationality will be explicitly discussed in other terms, in particular in chapter 3, where normative standards of rational criticizability and blameworthiness are discussed. I thank Prof. Thoma for clarifying that issue.

1.7 Desiderata and scope of the dissertation

A convincing answer to the research question that is pursued here should comply with two descriptive desiderata. First, the account should convincingly explain how decision problems arise for bounded agents. In particular, it should relate the representation of a decision problem to the agent's context of agency. Consequently, a theory of representation should articulate descriptions in the vocabulary of philosophy of action, to descriptions in decision-theoretic terms. Second, the account should have explanatory and predictive power for boundedly rational agents: it ought to account for "normal" as well as "anomalous" representations and decisions. Indeed, determining the representation of a decision situation faced by the agent should contribute to answering the question: "Why did the agent phi?" This is due to the explanatory role of rationality in "normal" (non anomalous) cases. Then, the answer offered by a theory of representation will be of the form: "the agent represented the decision problem he is facing as R, and applied decision rules to it, which explains why he phied."

In addition to these descriptive constraints, I introduce several normative desiderata. First, the account should be normatively plausible, in the spirit of the "ought implies can" adage. I focus on requirements of rationality which are compatible with bounded agents' limited abilities, and which take into consideration the cognitive costs these constraints may require of them. Additionally, if decision theory is to account for some of the normative intuitions of folk psychology,[11] it should take into account the specificity of human agency. Such an endeavor is also worthy for those whose interest lies in general rationality standards, as they may infer more general hypotheses by contrast with an account of bounded rationality, either by lifting some of its assumptions or by rejecting certain of its claims in the case of unbounded agents.

Second, the representation of a decision problem should help answer the question: "Why did he not act rationally?" and thus contribute to explain where the irrational choice stems from. An act may be irrational because the agent knows what they should do but fails to act accordingly (akrasia). Then the representation of the decision problem is correct but some decision theoretic principle has been violated. Alternatively, irrational decisions may result from the agent's failure to consider the relevant features of the decision situation; if this is the case, the charge of irrationality bears on the misrepresentation of the decision problem: had the agent represented the problem correctly, he would have acted accordingly. This second kind of irrationality is at work in framing effects, and a satisfactory theory of representation is expected to explain what went astray.

Finally, the theory should contribute to account for how sophisticated agents may represent decision problems and choose. In particular, it should investigate the possible existence of representational strategies, which would allow agents to use representations as instrumental tools for their ends.

In addition to these desiderata, a few restrictions need to be made on the scope of the research

[11] This view is, for instance, defended by Pettit (1991).

question. This dissertation does not address epistemological issues of representation, or if so only indirectly (see subsection 5.3.2 for further distinctions). Since the concerns of decision theory are guided by practical reason, I will not investigate the conditions for epistemically correct representations. Additionally, I will not delve into linguistic considerations nor philosophy of mind. Some semantic issues will be discussed, though only in their relationship with practical reason. Last, although the representation of a decision situation can be viewed as a process, this dissertation does not offer action-guiding standards nor procedures of representation.

1.8 Claims

This dissertation defends the view that decision-theoretic representations for the bounded are instrumental tools to agents' ends. If decision theory accepts that rational agents are not omniscient beings that take into account all possible descriptions of a decision situation, it is reasonable to assume that decision problems are non-exhaustive and selective representations.

The negative part of this account is dedicated to showing that these selection issues for decision-theoretic representations are independent from those raised by the Invariance principle. Schick's account of understandings and Bermúdez's account of frames rely on the claim that conflictual decision situations justify a notion of representation that violates the Invariance principle. Both interpret framing effects as several alternating point of views on the same decision situation. I argue that conflict is not sufficient to justify such notions of representations. Situations of conflicts can instead be modeled without violating Invariance by treating consequences as intensional properties of particular options.

Standard decision theory overlooks two issues of representation, that I call the issues of the scope and granularity of decision problems. Indeed, a tension exists between on the one hand our incentives to isolate certain considerations to deliberate about a particular decision, and on the other the fact that our decisions are often connected through their consequences. An account of rationality for the bounded should provide guiding principles to isolate decision problems in a way that is both realistic and permissible.

Since they are selective representations, decision problems presuppose some prior choices and evaluative attitudes that can explain why the decision problem is specified in a certain way. A plausible explanation to our way of designing decision problems is the fact that we are creatures of projects. As such, our decision situations and choices are embedded within a structure of future intentions and plans. If this is correct, that means decision problems are also not found but constructed. Depending on our ends, these representations are more or less precise and extended. They may be the product of non-deliberative, reflective, or policy-based plans. This allows us to explain how we represent decisions requiring non reflective intentions, as well as cases where the agent reflectively deliberates about the most appropriate representation of the decision situation.

Although future intentions and plans are defined by Bratman relative to the agent's desired ends, I argue that a third evaluative attitude needs to be added to desires and future intentions to make sense of instrumental representations. I defend what I called a Caring Principle, *that consequences are objects that make a differential impact which is relevant for the decision-maker if she cares about it*. This principle is motivated by a number of arguments in favor of distinguishing desires from cares for representational purposes.

Combined with planning, the Caring Principle offers an account of instrumental representation that is a plausible solution to the problems of scope and granularity, as well as to Resnik's "Specification problem".

The account distinguishes external and internal standards of assessments of representations. Internal standards determine whether the agent or her intentions are blameworthy given her prior plans; external standards assess their rational criticizability independently from her prior plans. These requirements show how one should rationally filter options and set how finely grained states of the world should be. However, an intention-based account of representation can only address the objections raised at irrational intentions if it is grounded on the agent's cares instead of desires.

Instrumental decision problems are thus characterized as partial, selective representations that are shaped by the agent's cares and plans. This definition does not stipulate what framing effects are nor when they are irrational, but already gives a rationale of why a particular representation is relevant or not for the decision maker.

The definition of framing effects defended by this account relies on Deitrich and List's hypothesis that a decision situation can be "represented extensionally" by treating its relevant features as properties of the choice situation; then, these relevant features also constitute the set of weighable reasons of the decision situation. The context of the situation can then be meaningfully defined as the set of relevant reasons of the decision situation, and two kinds of context-sensitivities should be distinguished: first, whether a property is instantiated or not can depend on the context of the decision situation. Second, the evaluation of those properties can itself be context-sensitive. Applied to framing effects, the evaluative requirements raised the questions of (i) when the same consequence should be valued equally across contexts, and (ii) when equally valued consequences should play the same evaluative role across contexts.

To answer these questions I introduce five extreme views about practical reasons — Invariabilism, Atomism, Holism, Generalism and Particularism — and argue that all should be rejected, both on general grounds and in the case of bounded agency. These views thus cannot offer a criteria of demarcation for framing effects. Nevertheless, the discussion about the way we compose reasons and their weights can be used to determine when the instantiation and the evaluation of properties are legitimately context sensitive, via the notion of general and particular differences.

In that light, I defend a principle ruling out illegitimate framing effects as inconsistent choices in decision situations that are "evaluatively equivalent". Evaluatively equivalent decision situations

can be defined in terms of general and particular differences the agent cares about. Irrationality claims should be distinguished in three cases: in the lab via a formal description of a decision problem, in the lab via a practical decision situation, and in real-world situations. *Our normative intuitions about framing effects thus do not require violating the Invariance principle nor Invariabilism.*

I compare my account with Bermúdez's account of frames, the only theory of representation that also takes decision problems to be constructed and value driven. Both of our accounts contend that orthodox decision theory is over restrictive regarding framing effects, yet for different reasons. I then defend the present account on a number of grounds, and plead in favor of accounts of representation that acknowledge that all specifications of decision problems and their consequences are value-driven.

This dissertation gives an even greater role to evaluative attitudes in representing decision problems than Schick and Bermúdez: indeed, any decision-theoretic consequence represents a feature of the decision situation that has been selected based on some evaluative difference that this feature makes for the agent; either directly through cares, or indirectly through instrumental plans grounded on the agent's cares. This makes the decision-theoretic representations of this account partial, constructed, and value-driven though it does not violate the Invariance principle. Consequently, rational representations for the bounded are not theoretically neutral for decision theory: the way we represent a decision situation cannot be reduced to the standard model of beliefs and desires. They are not evaluatively neutral representations either, as they involve the selection of a set of relevant features. What guarantees the coherence and stability of and coherence of our representations is our cares and our plans.

1.9 Outline

Though most chapters of this dissertation were written as self-contained pieces, they all participate in the elaboration of a coherent account, primarily developed in chapters 3 and 4.

In chapter 2, I discuss one of the few current theories of representation articulating philosophy of action and decision theory. Schick's theory of understandings focuses on decision situations where agents assign conflicting beliefs and values to descriptions of the same outcome, act, or state of affair. In his alternative model of action, he claims that the agent's understanding of a conflicting situation explains why she acts on some desires and beliefs and not others. By doing so, he rejects a tenet of decision theory widely agreed upon: the Invariance, or extensionality. I argue that his account is too permissive from a normative perspective, and offer a model of conflict that does not violate Invariance if consequences are intensional properties.

In chapter 3, I defend the Care Principle: rationally admissible consequences are *objects that make a differential impact which is relevant for the decision-maker if she cares about it*. However, this principle of representation is not sufficient on its own to account for how we construct and represent decision problems, as it does not address what problem of the scope and grain of

decision problems. I hence defend a second principle of representation which claims that the decision problems we face are shaped by prior intentions and plans as Bratman (1999) defines them. By providing a filter of admissibility for options, prior plans offer a solution to the problem of scope mentioned earlier. A plan-based view of representation would state that only options that are inferred from our prior plans, through means-end reasoning, should be included in our decision problems. I then offer a systematic account of plan-based instrumental representations and show how it solves Resnik's problem of Specification.

In chapter 4, I investigate Dietrich and List's hypothesis that decision problems may be represented "extensionally" as a set of intensional properties corresponding to the set of reasons weighing in favor of a particular alternative. This hypothesis allows the authors to distinguish two kinds of context-sensitivities and to give a definition of irrational framing effects. To assess their account, I examine five views about the context-sensitivity of practical reasons and their weights, including Invariabilism mentioned above. I argue that these five views should be rejected, and thus cannot offer a criterion of demarcation for framing effects. Nevertheless, the discussion about the way we compose reasons and their weights can be used to determine when the instantiation and the evaluation of properties are legitimately context sensitive via the notion of general and particular differences. I then defend a principle ruling out illegitimate framing effects as inconsistent choices in evaluatively equivalent decision situations. I plead for distinguishing irrationality claims made in three particular cases: in the lab via a formal description of a decision problem, in the lab via a practical decision situation, and in real-world situations.

In chapter 5, I review Bermúdez's account of non-extensional frames and frame-sensitivity, and examine his claim that decision theoretic standards of rationality should be complemented with certain standards of reasoning with frames. After presenting his account of frames, I examine how it differs from the account of representation defended by this dissertation. Among the key differences, Bermúdez imposes externalist constraints on beliefs when the present Humean account does not. I then examine Bermúdez's claims about rational representations and look into two issues. I first argue that his canonical example (Agamemnon) is not sufficient to reject the Invariance principle nor to justify the introduction of ultra-intensional frames, consistently with chapter 2. What is more, irrational framing effects can be defined independently from factual neutrality as inconsistent choices across evaluatively equivalent decision situations, as defended in chapter 4. Second, I contend that the specification of factual propositions in a decision problem is not evaluatively neutral, which in turn violates the Due Diligence requirement. This pleads in favor of accounts that acknowledge that all specifications of decision problems and their consequences are value-driven.

In the last chapter, I first summarize the main conclusions of this thesis and provide an overview of its arguments. I then turn to the strengths and weaknesses of the account, and conclude by some open questions for further research.

References

Bermúdez, José Luis (2020). *Frame It Again: New Tools for Rational Decision-Making*. Cambridge University Press.

Bradley, Richard (2017). *Decision Theory with a Human Face*. Cambridge University Press.

Bratman, Michael (1999). *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge University Press.

Broome, John (1999). "Can a Humean be moderate?" In: *Ethics out of Economics*. Cambridge University Press, pp. 68–88.

Dietrich, Franz and Christian List (2016). "Reason-Based Choice and Context-Dependence: An Explanatory Framework". In: *Economics & Philosophy* 32.2, pp. 175–229.

Frigg, Roman and James Nguyen (2021). "Scientific Representation". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2021. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2021/entries/scientific-representation/.

Jacob, Pierre (2023). "Intentionality". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Spring 2023. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/spr2023/entries/intentionality/.

Merriam-Webster (2024). *Context*. In: *Merriam-Webster.com Dictionary*. URL: https://www.merriam-webster.com/dictionary/context (visited on 01/11/2024).

Pettit, Philip (1991). "Decision Theory and Folk Psychology". In: *Essays in the Foundations of Decision Theory*. Ed. by Michael Bacharach and Susan Hurley. Blackwell, pp. 147–175.

Resnik, Michael D. (1987). *Choices: An Introduction to Decision Theory*. University of Minnesota Press.

Schick, Frederic (1991). *Understanding Action: An Essay on Reasons*. Cambridge University Press.

Tversky, Amos and Daniel Kahneman (1981). "The Framing of Decisions and the Psychology of Choice". In: *Science* 211.4481, pp. 1–30.

Tversky, Amos and Daniel Kahneman (1986). "Rational Choice and the Framing of Decisions". In: *The Journal of Business* 59.4, S251–S278. (Visited on 01/11/2024).

Chapter 2

Understanding decision situations

Schick's theory of understandings and strongly intensional desires

Contents

2.1 Introduction

In chapter 1, I introduced the question raised by this dissertation: what are the rationality standards of decision-theoretic representations? Upon asking what counts as a rational representation of a decision situation, I examined several philosophical issues raised by framing effects. First, how stable should our representations of outcomes be? And then, how robust should our attitudes be to such representational shifts? In this chapter, I consider one of the few theories of representation of decision situations that address these issues: Schick's theory of action and

understandings.

Schick focuses on decision situations where agents assign conflicting beliefs and values about descriptions of the same outcome, act, or state of affair. Through a series of examples, he argues that beliefs and desires, that are traditionally considered the reasons for action, often fail to explain and rationalize agents' actions. In his alternative model of action, he claims that the agent's understanding of a conflicting situation explains why she acts on some desires and beliefs and not others. Understandings are psychological states that allow agents to see and thus value facts in various ways. Schick argues that these understandings "activate" certain reasons that motivate the agent to act accordingly.

Schick's theory offers interesting insights on issues of representation of decision problems. First, he supports the view that the same fact may be described by several distinct propositions. Second, he claims that rational agents may assign different values and probabilities to propositions corresponding to the same fact. By doing so, he rejects a tenet of decision theory widely agreed upon: the invariance principle (sometimes called extensionality principle). As agents understand situations differently, Schick argues, they come to select different descriptions of the decision problem leading to different choices. This leads him to a model of representation that selects only evaluations and beliefs about descriptions corresponding to the agent's understanding of a given fact, leaving aside other evaluations and beliefs that they may hold about that fact. This account is appealing as it makes room for a realistic conception of decision-making, where agents are not committed at any particular moment of time to including all of their beliefs and desires about a situation into their representation of the decision problem they face.

I agree with Schick's view that representations of decision problems are bound to be selective descriptions. Moreover, If decision theory accepts that rational agents are not omniscient beings that take into account all possible descriptions of a fact, it is reasonable to assume that rational representations can be selective about which propositions to include in the description of a decision problem. The same act or outcome may be described and valued in a variety of ways depending on the decision problem at hand and on the agent's ends. Along with Schick, I believe that this fact doesn't threaten to make decision theory to become a vacuous theory of rationality, as different representations of a decision situation lead to different prescriptions. However, I disagree with his account of understandings as rational representations and his analysis of conflict as a paradigmatic case of extensionality violation.

Schick's account nevertheless raises a series of important questions for decision theoretic representations:

- How do decision problems arise, and can we establish a set of principles that specifies how agents ought to construct their representations?
- May rational agents leave out some propositional descriptions of a given fact when they represent a decision problem, along with the beliefs and values associated with these descriptions?

- How can a theory of representation accommodate the fact that representations of decision problems are selective descriptions, without making rationality constraints vacuous?
- Should the representation of a decision problem include everything the agent cares about simpliciter, or only what she cares about for the decision problem at hand?
- If non-defeasible attitudes like values, ends or care are to play a role in the construction of decision-theoretic representations, how do we make sure they don't license "delusional representations" that describe the decision problem in a way that fit their ends, irrespectively of the agent's beliefs?

After briefly introducing understandings, I argue that the notion of understanding should be distinguished from awareness, so that arguments in favor of understandings should not be confused with arguments in favor of awareness. I then examine Schick's central claims: the fact that conflicted agents can be represented as agents with intensional attitudes, and that understandings are the best explanations to the way agents rationally settle among these attitudes. His account thus connects three notions: understandings, conflicts, and intensionality. To assess these claims, I will first focus on the relationship between understandings and conflict, and argue that the existence of rational conflicts does not make a compelling case for understandings. I then question the relationship between intensionality and conflict, and argue that situations of conflicts can be modelled without violating weak extensionality. In the last section, I draw out the implications of Schick's insights on self-delusion for a theory of representation.

2.2 Introducing understandings: Orwell in the Spanish War

Schick's most recurring example is drawn from Orwell (1954) "Looking Back on the Spanish War", where he recalls his experience as a soldier and describes the odd situation he faced when he was once confronting a fascist enemy:

> "*A man [. . .] jumped out of the trench and ran along the top of the parapet in full view. He was half-dressed and was holding up his trousers with both hands as he ran. I refrained from shooting at him [. . .] I did not shoot partly because of that detail about the trousers. I had come here to shoot at 'Fascists'; but a man holding up his trousers isn't a 'Fascist', he is visibly a fellow-creature, similar to yourself, and you don't feel like shooting at him.*"[1]

Schick's analysis of the situation can be described as follows: Orwell had both the beliefs that the soldier was a fascist enemy and a human being, and both the desires to kill the fascist and save the human being:

> "*Orwell wanted to 'shoot at Fascists' and he believed he now could do it. On the belief/desire theory, he had a solid reason for shooting. What then was it about those pants that got him to put down his gun? Orwell answers that question: 'a man holding up his trousers. . . is visibly*

[1] Orwell 1954, § III, p. 199.

> *a fellow-creature, similar to yourself'. I take it the pants were down to his knees, and that Orwell is saying that someone half-naked and 'visibly' human had to be seen as human. Before the man jumped out of the trench, Orwell had seen his firing at him as shooting at a fascist, which he wanted to do. The soldier's half-naked predicament was for him a wake-up call — like Adam's call to me. He then saw his firing at him as his shooting at a fellow-creature, and this he didn't 'feel like' doing. He had, of course, known all along that, under their pants, the fascists were human. He had never faced that fact, never fully confronted it, but how did his not having faced it weaken the force of his knowing it? And how could he want to kill a fascist and also not feel like doing it? How can a change in a person's seeings undercut what he wants to do?"*[2]

As suggested in the previous passage, Schick argues that standard beliefs-desires analyses are not sufficient to explain Orwell's choice of putting down his gun: he had reasons to shoot and not to shoot, and none of his beliefs nor desires changed along the way. However, a glimpse at the man's trousers made him change his mind. Schick's point here is that although Orwell's beliefs and desires did not change, his *understanding* of the situation did: this understanding connected the situation to the relevant belief (the soldier is a human being) and desire (I should spare human beings) in Orwell's mind. Note that for Schick, Orwell's choice of sparing the soldier *cannot* be explained in terms of the relative desirability of the two options: this is not the case that Orwell wanted more to spare the human being than to kill the fascist, for "*That would only give us more questions. Why should a man's holding up his pants have changed the relative strengths of Orwell's desires?*"[3]

2.3 Understanding is not awareness

One of the reasons why Schick's account is appealing is that it complements traditional reasons with a third subjective factor to explain choice, that provides a necessary condition for beliefs and desires to be activated. As such, the role of understandings seems similar to that of awareness: both concepts are used to account for the fact that agents may not act on attitudes they have, because of some cognitive state they are in at the time of choice which prevents them from acting on these attitudes. What is the relationship between awareness and understanding? This question matters insofar as if the two notions were to overlap, it would offer normative support to Schick's account. Indeed, it would be hard to blame Orwell for becoming aware of a particular proposition, or for becoming unaware, so to speak, of another one. In what follows I will review three decision-theoretic interpretations of awareness, and argue that Schick's notion of understanding cannot be interpreted as any of them.

Bradley (2017) distinguishes awareness from opinionation. Agents may be aware of a proposition without having particular beliefs or desires about it. Awareness can be defined here as the

[2] Schick 2003, p. 3.
[3] Schick 2003, p. 5.

consciousness or availability of a proposition before the agent holds opinions to it. Like understandings, awareness gives access to a proposition independently from the attitudes we may hold towards that proposition. For instance, Schick observes that "if someone has some understanding, he needn't believe the proposition that expresses it. He may be only considering the proposition, reflecting on it, debating it".[4] However, the analogy between awareness and understandings stops there. Understandings being selective, they allow agents to discard beliefs and desires they already have, by selecting some propositions and not others.

Consequently, one may have beliefs and desires without a specific understanding. This is the case for Orwell before he decides to shoot: he has certain beliefs and desires in favor of shooting the soldier, he does not act on them, as he doesn't understand the fact of shooting in a way that activates these attitudes. The existence of an understanding is thus not a necessary condition for opinionation. By contrast, If we follow Bradley's general account of awareness, one may not have beliefs or desires about a proposition they are not aware of. Awareness of a proposition is necessary for the agent to hold an attitude about a given proposition. Could we say of Orwell that he is unaware of some proposition relevant to his choice? I don't think so, as if he were, he would not be able to have beliefs and desires about that proposition.

So far, I have compared understanding to a restricted definition of awareness. According to that definition, we are not aware of a proposition or a prospect if we never encountered it before. However, a looser definition of awareness may describe agents as unaware of a proposition when their attention is no longer turned to that proposition — they may have forgotten about it or it slipped their attention. The parallel between understanding and attention is tempting in Orwell's case. It would explain his decision not to shoot by arguing that he turned his attention from a set of propositions to another. Moreover, this interpretation would be coherent with the fact that it was by looking at the soldier's pants that Orwell decided not to shoot: his attention was diverted from one description of the situation to another.

Although this interpretation of understanding as attention to some propositions makes sense in Orwell's case, it is at odds with other examples Schick uses to illustrate his concept of understanding.

In the rich aunt's case, a rich elder lady may have been killed by either her niece or her butler. Both have "motives" or incentives to kill her, as the niece could get richer out of the murder and knows it, and the butler wanted to punish the lady for sleeping with the gardener. Yet only the butler chose to kill the woman. Although both have beliefs and desires conducive to the choice of killing, why did one choose that option and not the other? Schick argues that the difference between the two agents lies in their representation of the problem: the niece believed that the killing would make her rich and desired to be rich, but she didn't *see (or understand)* the killing as an act of enrichment but as a murder. Therefore her wanting the money wasn't activated. By contrast, the butler saw the killing as a way to get revenge, and had adequate belief and desires

[4] Schick 1991, p. 82.

about it.

> *"The niece (it had crossed her mind!) saw the killing as murder, from which she backed off. The niece didn't see the killing of her aunt as something she wanted to do (it hadn't crossed her mind as an act of self-enrichment). So her wanting to be rich never got connected for her."*[5]

Both believed that killing the old lady would make them rich, and both desire to be rich, yet only one sees the killing as an act of enrichment. The different understandings of the prospect, Schick argues, leads them to different decisions. Now, if we relate the notion of understanding to that of attention, it would imply that what made the difference between the two individuals is their relative attention (or inattention) to some propositions. While the butler decided to kill the woman to get rich, the niece did not, simply because she did not turn her attention to that description of the prospect. I don't find the explanation convincing, especially as the stakes in this situation are so high. As stakes rise, agents' attention is less likely to be diverted. In such a simple situation, resorting to attention to explain differences in decisions is not plausible.

I have so far considered two senses of awareness that could not be reconciled with Schick's notion of understanding. I would like to turn my attention to a third acception of awareness introduced by Bradley (2017), that is relevant for representation of decision problems. In certain institutional settings, it may be required that some piece of information be removed from a case, and that it should not be taken into consideration to make subsequent decisions relative to that case. For instance, inadmissible evidence in a legal judgment falls in that category. In such situations, features of a prospect, or a prospect itself, may deliberately be removed from the description of the decision situation, on the ground that it should not be taken into account in the deliberation process. To take another example, recruiters may be required to ignore certain considerations such as race or age when assessing a job candidate.

Can Schick's concept of understanding be interpreted as a form of "deliberate unawareness"?. In the rich aunt's case, it may be argued that the butler deliberately chose not to consider the moral implications of the murder, or that the niece refused to take into account the financial benefit of that prospect. Although this interpretation is plausible, I don't think it captures what Schick has in mind. Orwell's example is a case in point: he still faces conflicting beliefs and desires as he decides not to shoot, and did not deliberately remove some of them from his attention. This example suggests that Schick does not only wish to account for deliberate exclusion of information or attitudes, but also for cases where agents are unconsciously prone to saliency effects — be them rationally permissible or not. Schick is explicit about this ambition:

> *"Understandings can be tacit. They can be fully unconscious: we needn't be aware of all understandings and may sometimes even resist admitting to ourselves how we are seeing things."*[6]

We have seen that Schick's notion of understanding cannot be reconciled with any decision-

[5] Schick 1997, p. 19.
[6] Schick 2003, p. 83.

theoretic concept of awareness. Consequently, arguments in favor of awareness don't apply to understandings: one cannot claim that the way understandings select some attitudes and not others is rational on the ground that understandings correspond to the state of awareness of the agent when deliberating and settling on a particular option. I believe this is due to the specificity of understandings, that simultaneously pick propositional descriptions and attitudes that the agent may already be aware of. Unlike awareness, understanding is an intensional notion that presuppose a distinction between facts and coextensional propositions describing these facts. Schick's arguments in favor of understandings hinge on his account of intensionality. Indeed, he claims that conflicted agents can be represented as agents with intensional attitudes, and argues that understandings are the best explanations to the way agents settle among these attitudes. His account connects three notions: understandings, conflicts, and intensionality.

2.4 General requirements on understandings set by Schick

Schick's understandings are part of a theory of action committed to two central normative claims. One is that the way we describe a fact may legitimately impact one's decision. This first claim relies on the idea that propositions are in a many to one relationship with facts, and that individuals should be allowed to consider some propositions and not others in relationship to a fact. In addition to the claim that our understandings of a fact is selective, Schick offers two criteria for what counts as an adequate understanding. As we will see, distinct understandings allow rational agents to value differently descriptions of a fact that are not logically equivalent. In addition to logical constrains, Schick does condemn some illegitimate uses of understandings on specific moral grounds. In substance, Schick deems unacceptable the understandings that are self-delusive, dishonest, or inauthentic.[7] Self-delusion consists in persuading ourselves that we have a certain understanding of a situation on the ground that it makes us feel better about our choices. A criminal may for instance convince himself that he is doing his job when committing a crime, although he does believe the crime is immoral. By persuading himself that this is a job and not a crime, he manages to feel good about a decision he was already committed to take although he didn't feel good about it. Schick condemns this way of understanding the decision situation on the ground that it is inauthentic. Consequently, an adequate understanding corresponds to a set of beliefs and desires that are logically consistent and morally acceptable.

Except for these two conditions, Orwell could have understood his action indifferently as the killing of a fellow human, as an act of resistance against fascism, or as both. This, in spite of the fact that he holds beliefs and desires associated with each of these understandings. More generally, Schick contends that some desire that is relevant for the decision-maker may rationally be "deactivated" as long as the agent doesn't have the understanding associated with that desire.

[7] "*Persuading ourselves to adopt an understanding on just this basis is a self-delusion (the analogue for understandings of a doxastic self-deception). It is a public relations stunt pulled by us on ourselves. An understanding contrived for this purpose might be said to be inauthentic. Or we might call it dishonest. This can now be extended. In the basic case above, we expect to do x and take a new view of that.*" (Schick 1991, pp. 151–164)

In his view, rationality does not have any bite on which understandings ought to be included in the representation of a decision problem, as long as it satisfies the logical consistency and the moral requirement.

Before delving into a more formal discussion of the invariance principle, I should already say that these rationality requirements on understandings are too permissive. Orwell should have rationally included both descriptions of his option in his representation, to the extent that both were relevant to his choice. True, this forces him to face a situation of conflict he may not be able to settle, but from a normative perspective, this is not problematic. Rationality allows individuals to be conflicted when the outcomes they face are incommensurable. More importantly, if the two outcomes are commensurable, cherry picking one at the expense of the other after deliberation would constitute a failing of rationality. This means that either the features of the conflicting decision situation are commensurable, in which case Schick's rationality requirements on representation are too weak; or these features are not commensurable, and the moral requirements imposed by understandings are too strict from a moderate Humean perspective: one may rationally represent a decision situation independently from the degree of immorality of their values and beliefs. I will come back to this issue in more precise terms in section 2.7 where I discuss the relationship between conflict and what Schick calls strongly intensional desires. But first, I will examine Orwell's case informally and argue that the existence of rational conflicts does not make a compelling case for introducing understandings as a third decision theoretic attitude.

2.5 The existence of rational conflicts is not a compelling argument in favor of understandings

One of Schick's motivations for such permissiveness is that it provides a model of conflicted yet rational agents. Conflicted agents are individuals who want both h and k, although they believe that "if h then not-k". This is problematic if these agents's beliefs and desires are closed under deduction, as is the case with Schick's account (see section 2.6 for a more detailed review): Under deductive closure of beliefs and desires, if I desire X and believe that "if Y then X", then I (should) desire Y. If deductive closure holds, k and not-k are simultaneously desired by the conflicted agent. By adding a proviso on the closure of desires, one may rationally be conflicted without having contradictory attitudes. Indeed, it allows Orwell to rationally want to kill the fascist and spare the human being, as long as he understands the option of shooting as only one of the two the two descriptions of the act is salient to him. Schick thus interprets Orwell's situation as follows. At first, the British writer is conflicted between two desires, but he has no specific understanding of these desires. Then, the killing of a fellow human becomes the salient description of his option, and this is what moves him to spare the German soldier. From the beginning to the end, Schick deems Orwell rational to the extent that he never had two contradictory understandings of the option simultaneously. This interpretation allows Schick to

model situations of rational conflicts, by the absence of specific understanding, and to explain how conflicted agents settle on an option when one understanding arises.

Situations of conflict often are described in terms of incommensurable desires. In that picture, agents are conflicted between two options when they desire both, yet can't manage to desire one more than the other. Schick rejects this model of conflict because it fails to explain why we settle on one of the alternatives. By introducing a third factor which gives causal force to beliefs and desires, he hopes to account for the dynamics of conflict settlement. However, I think the price to pay for this modelling strategy is high, as it may make the concept of understanding lose its substance.

To see this, let us consider Orwell's situation. If Schick wants to make his attitudes and choices rational, he must assert that Orwell had initially no specific understanding of the situation, and that one eventually became salient to him. Nevertheless, Orwell had killed fascists before, he understood these actions in those terms, and was now again on the battlefield for the same reasons. Is it still plausible to say that he didn't see the shooting of the German soldier as the killing of a fascist? I don't think so. The concept of understanding is defined as the subjective grasp of fact as a certain proposition, and I doubt that Orwell didn't grasp the shooting as a legitimate killing, be it before or after the action. Consequently, it must be the case that he had at least that understanding of his option. However, since he settled on sparing the soldier, he must also have held the alternative understanding of the act, as the killing of a fellow human.

Schick's interpretation is problematic here, as either Orwell simultaneously held two contradictory understandings, which is not rationally licenced by Schick's account; or that only one of the two descriptions (the sparing) was salient to him at the time of action, while the other (the shooting) was active at some point before acting. This saves both Orwell from being irrational by Schick's standard, and explains why he eventually spared the soldier. However, if this is the case Orwell's situation can just as simply be explained in terms of changing desires: he initially wished to kill the German, and at the sight of the pants changed his mind about it. I am not questioning the explanatory role played by that sight on Orwell's choice, only that explaining the conflictual decision by an external stimulus (the sight) changing the agent's understanding is not explanatorily more powerful than explaining it by the effect of that stimulus on changing desires.

The simplest explanation for Orwell's behavior is that he was torn between two incommensurable desires. His desires may have changed overtime; he may have settled for an option by resolving his conflict (this is unlikely as he didn't stop fighting fascism after that episode), or he may just have chosen one of the two alternatives without managing to sort out his desires. In the case of incommensurable desires, I don't think rationality has any bite on the conflicted agent's choice: one should be licenced to select any of the two alternatives without further constraint. Schick tries to do better than this kind of explanation, by introducing an attitude that will spell out how Orwell settled his conflict. Understandings thus serve two purposes for Schick: to explain

conflicts and their settlement, by activating certain beliefs and desires, and to capture the agent's subjective grasp of the situation. However, these two purposes can't be reconciled, since if an agent is conflicted, he must simultaneously grasp the situation in conflicting ways. Yet Schick's understandings can't rationally allow that, so the only way to make conflicted agents rational for him is to say that they hold conflicting understandings successively and not simultaneously. But then understandings don't do better than desires to explain conflicts, and the notion loses its explanatory substance.

Let's take stock. Understandings correspond to ways of representing decision situations before acting. The main originality of understandings is that they allow conflicted agents to differently value the same fact, option, or outcome under two descriptions. Furthermore, Schick argues that understandings make sense of rational decisions that could not be explained by beliefs and desires alone. I rejected this last claim in the present section. In the following sections, I argue that situations of conflicts can be modelled without supposing that the conflicted agent knowingly values the same option differently under two descriptions. In Schick's vocabulary, conflicted agents like Orwell do not violate the weak extensionality of desires for options. To make the argument precise, I first present Schick's account of rational intensionality.

2.6 Schick's account of intensionality: Material equivalence and coreportiveness

In order to allow for distinct understandings of a situation, Schick proposes to relax extensionality constraints on agents' attitudes. Roughly put, extensionality requires that agents be insensitive to the way events or facts are described. Facts here are defined as "whatever a true proposition reports that makes this proposition true". In Schick's account, agents may have different attitudes to the same fact, and this is done indirectly by coreportive and materially equivalent propositions.

Materially equivalent propositions, that are both true or both false, may report the same fact and nonetheless be valued or believed differently. Here material equivalence is to be distinguished from logical equivalence. Materially equivalent propositions have the same truth value de facto, whereas logically equivalent propositions necessarily have the same truth value. For instance "London is in England" and "Paris is in France" are materially but not logically equivalent. Coreportive propositions are defined as propositions whose material equivalence is not accidental, as it derives from some physical self-identities. For instance, "Cosmo Kramer is tall" and "Jerry Seinfeld's neighbor is tall" are coreportive, since Cosmo Kramer and Jerry Seinfeld's neighbor are physically identical. Let us note that in this framework, all coreportive propositions are materially equivalent, but materially equivalent propositions may not report the same fact.

Keeping the definition of coreportiveness in mind, understandings can be introduced as mappings from a fact to a proposition. We understand a fact in some particular way as we map

it to some proposition reporting that fact: "In an understanding, we map out a fact, we grasp it in terms of some proposition. Still, it is the fact that we grasp, not the proposition, that we then understand."[8] Whenever an individual understands some fact or option as a particular proposition, we say of this proposition that it is salient or compelling. Understandings are key to Schick's account, to the extent that they allow agents to have non extensional attitudes about a fact. For instance, one may assign two different values to the same fact F through two materially equivalent propositions p and q reporting that fact; if one understands F as p, then only one value of p is taken into consideration for choice.

To assess extensionality constraints, Schick introduces three degrees of extensionality, as defined below. Let p and q be propositions, and A be some attitude to these propositions. Schick defines weak, strict and strong extensionality as follows:

- A is *strongly extensional* if: (p if and only if q) $\Rightarrow$ ($A(p)$ if and only if $A(q)$);
- A is *strictly extensional* if: (p and q are coreportive) $\Rightarrow$ ($A(p)$ if and only if $A(q)$);
- A is *weakly extensional* if:

(p and q are coreportive and the agent believes it) $\Rightarrow$ ($A(p)$ if and only if $A(q)$).

The failures of strong, strict, and weak extensionality are respectively weak, strict and strong intensionality: for instance, an attitude that fails to be strongly extensional is weakly intensional. Schick observes that strong intensionality implies strict intensionality, which in turn implies weak intensionality. Consequently, these three kinds of intensionality are to be understood as three degrees of permissiveness for agents' attitudes. What degree of intensionality does Schick's account license for beliefs and desires? First, it allows agents to believe one proposition and not its coreportive counterpart as long as the two sentences are not known to be coreportive. Beliefs are thus strictly but not strongly intensional here. By contrast, desires violate weak extensionality: we may want one proposition to be true and not another although we believe them to be coreportive. Let's keep in mind that while Schick endorses strong intensionality of beliefs and desires for materially equivalent propositions, he remains committed to weak extensionality when it comes to logically equivalent propositions.

In addition to principles specifying intensionality constraints, we saw that Schick introduced principles of deductive closure for beliefs and desires. One stipulates that a rational individual ought to believe all deductive consequences of what he believes; the other imposes the same constraint on desires. A third commonly accepted principle of closure states that if a person wants h, believes k, and if m follows from h and k, then she must also want m. Schick relaxes this third principle by attaching a proviso to it: the closure holds only in the case where h and m are salient for the individual, i.e. if the individual understands some fact in terms of proposition h which he wants, and if the fact following from h and k is understood in terms of proposition

[8] Schick 1991, p. 83.

m. Roughly put, Schick only binds agents to deductive closure at the level of understandings rather than desires: they may hold desire about contradictory descriptions of a certain act as long as these descriptions are not simultaneously salient to them.

With these requirements in mind, I now argue that decision-theoretic consequences can be modeled as intensional properties of extensional options; if this is correct, situations of conflicts like Orwell's do not violate the weak extensionality of desires as Schick claims.

2.7 Conflicting consequences can be modeled as properties without violating extensionality

I argued in the previous paragraphs that Schick's argumentative strategy in favor of understanding fails, as it either makes conflicting agents inconsistent, or it renders the concept of understanding unsubstantial. Moreover, the rationality requirements bearing on understandings are too permissive and the moral ones too strict to constitute rationality standards of decision-theoretic representations. However, one may be worried that a model of rational conflict may still require violating the principle of weak extensionality for desires independently from Schick's requirements on understandings. What is the relationship between intensionality and conflict? In Orwell's case for instance, it may be said that he both desired to shoot and not to shoot, and that the degree of desirability of shooting varied depending on the proposition describing that act: he desired to "kill the nazi" and "spare the human being" simultaneously. In such cases, conflict occurs at the level of propositional descriptions of the acts or facts. Yet more standard situations of conflicts arise when our actions have some desirable and undesirable consequences that we have difficulties weighing.

I see two ways of representing such situations, one that involves strongly intensional desires and one that doesn't. In the first model (let's call it model A), the consequences of an act are defined as coreportive propositional descriptions of that act. Acts are analogous to facts in model A, and a conflict arises whenever the act can be described by several coreportive propositions whose desirabilities can't be weighed. Propositional descriptions and consequences play the same role here: an act can be described by as many propositions as there are consequences For instance, "killing the fascist" and "not sparing the human being" are here two descriptions of the same act. Since consequences are subject to variable desirability, so do coreportive propositions describing an action. In model A, extensionality violation arises any time an action involves consequences with different desirability.

Strictly speaking, what makes a situation conflictual is the inability to weigh consequences rather than their differing desirability. Nevertheless, this model makes intensionality central to the act of weighing consequences, as consequences are defined as coreportive descriptions of an action. Schick's theory of understanding relies on this model, and we can see why: the appeal of this approach is that it captures the fact that deliberation occurs by comparing different descriptions

or perspectives of the act considered: when it comes to fighting fascism, Orwell prefers to shoot, but he doesn't when it comes to sparing human lives. An act has ambiguous desirability to the extent that it can be seen in several ways that are not equally desirable. However, this modeling strategy has a price, as it requires sacrificing extensionality: any act involving consequences with differing desirability will violate the principle of weak extensionality of desires.

There is a second way of modeling conflict that may save weak extensionality of desires. In that model (model B), conflicted agents do not have contradictory desires for coreportive descriptions of an option, their desires for options being simply indeterminate. The main difference with the previous approach is that consequences are not defined as coreportive propositions describing the same fact but as properties of the option that make an evaluative difference. As such, consequences are intensional objects that are valued for their sense and not their reference: the consequence of flying to the morning star may be considered and valued independently from the consequence of flying to the evening star.

In model A, in order to establish the desirability of an act, one needed to consider the desirability of the sense denoted by the proposition describing this act exclusively. Orwell's desire to "kill the nazi" was exclusively determined by that sense. However, this is not the case in model B where the desirability of an act does not depend exclusively on the sense of the proposition describing that act, but also on other senses considered by the agent. Each proposition instantiates a series of properties, and the desirability of the act is indeterminate until the properties instantiated by that act are weighed.

Let's see how these models of conflict can be formalized. In model A, p and q are two propositions coreportive of the same act, whose references are noted "qua p" and "qua q" respectively. Model A asserts that an act is desired under a particular propositional description, and in conflictual situations, an act has different values under different propositional descriptions, noted $v(\text{act qua } p)$ and $v(\text{act qua } q)$:

$$v(\text{act qua } p) = v(p \text{ qua } p)$$
$$v(\text{act qua } q) = v(q \text{ qua } q)$$
$$v(\text{act qua } p) \neq v(\text{act qua } q).$$

By contrast, in model B, conflict does not imply inconsistent evaluations of the same act, simply that the value of the act is indeterminate as long as the weighted desirability of its consequences is not determined. For a given option pair, conflictual features are represented as properties of an act, noted P and Q:

$$v(\text{act}) = v(P \,\&\, Q) = v(\text{weighted evaluation of } P \text{ and } Q).$$

Weak extensionality violation is no longer characteristic of conflict. In model A, the value of the act is not indeterminate but overdetermined in conflicting cases. The price to pay for the second

model is that it requires a linguistic distinction between features that are desired in virtue of their own sense and those that are desired in virtue of other senses considered by the agent. For instance, it must distinguish between desiring the morning star "qua morning star" and the desiring the morning star "qua morning star and evening star". I think this distinction is useful, as consequences in decision theory are objects of the first kind. Indeed, if you are considering whether or not to go to Venus, you may include both the consequence of going to the morning star, and that of going to the evening star. These are distinct properties with possibly distinct desirability. Decision-theoretic consequences must therefore be intensional objects that can be desired for their sense rather than their reference.

Let's summarize the general argument here: I have argued that Schick presupposes a model of conflict where the consequences of an option are coreportive propositions describing that option. This model (A) treats options as facts and consequences as propositions describing an option. Consequently, any two consequences that are not equally valued will violate extensionality. In situations of conflict the agent then (knowingly) values differently two coreportive descriptions corresponding to different ways of seeing and valuing an option. Any such situation will be treated as a breach of weak extensionality by model A. Instead, I have argued that model B is more appealing as it treats conflicts as an inability to weigh consequences of the same action, and not as different propositional descriptions of the same act. In model B, conflict does not require the desirability of options to be strongly intensional.

An objection: cases of coreportive descriptions of consequences

Now, one may worry that by still treating consequences as intensional properties in model B, the problem of extensionality violation may reappear at the level of consequences.[9] In what follows I examine three possible senses in which the valuation of properties could be problematic: if properties have the same extension, if they involve coreferential names, or if the properties happen to be identical. In all three cases, I argue that model B and the weak extensionality of desires gives the correct intuitions, so that the objection raised above does not go through.

Suppose that the set of animals that have a kidney is exactly the set of animals that have a heart, so that the associated properties have the same extension. I may still value these properties differently, even when these properties are consequences of an option in a decision problem. For instance, I may desire to save animals with a kidney and not those with a heart; but upon learning that these two properties are coextensional, I will have to weigh the two properties to decide whether overall I want to save these animals or not. In that sense, coextensional properties can be valued consistently with the weak extensionality of desires. So far, I have considered properties that are coextensional in the sense that every object having one property also has the other property.

Let's now consider cases where the consequences of an option involve coreferential names.

[9] I am grateful to José Luis Bermúdez for raising this important objection that needed to be addressed.

Suppose you have the choice to travel to Hesperus, which you love. One of the options thus possesses the property "traveling to Hesperus" as a consequence. You then discover that Hesperus and Phosphorus are the same planet. The two previous properties are cotextensional since every option possessing one property also possesses the other. However, by contrast with the previous case, they only differ in their proper name. One could be tempted to conclude that the "traveling to Hesperus" and "traveling to Phosphorus" are here two coreportive descriptions of the same property. I don't think this is right, since it is not the case that the same property is desired differently under two descriptions or two understandings: unless you find traveling to Hesperus desirable only because the planet's name is Hesperus, you presumably value the properties "traveling to Hesperus" and "traveling to Phosphorus" differently for independent reasons that made you find one desirable and not the other: e.g., you traveled to different parts of the planet and didn't equally enjoyed yourself.

That said, it could be that some of the reasons why you find the property "traveling to Hesperus" desirable also count as reasons for finding "traveling to Phosphorus" desirable. For instance, you have never been to Phosphorus nor to Hesperus, and in each case this constitutes a reason for traveling there. Once you discover the identity between the two planets, you should only count these two properties (of having never been to that particular planet) as one and the same. Consequently, the evaluation of overlapping properties should not be counted twice. As we will see in chapter 4, the issue of double counting of the weight of properties is not specific to consequences involving coreportive expressions: generally speaking, properties of an option that logically overlap should not be weighed independently. Whether or not the properties overlap, you have to make up your mind about the overall desirability of the option given that it makes you both travel to Hesperus and to Phosphorus. Though this way of speaking is not natural, it is faithful to the idea that the consequences of an option are features in virtue of which we prefer one option to the other, just as we favor an alternative for a particular reason.[10]

The last case to consider is the one where two properties are identical. There is no clear agreement in the literature on the conditions of identity between two properties,[11] but we don't need to specify these conditions to understand the role of identical properties in decision-theoretic evaluations. Suppose that you learn that properties A and B of an option are identical. So far, you have been treating them and valuing them differently, and when an option had both properties, you weighed them as separate consequences. Now that you know these two properties are identical, valuing them differently would be inconsistent, and weighing them as separate consequences would be a case of double counting. You thus have to make up your mind about the all-things-considered value of the property and only treat it as a single consequence of the option.[12]

We have considered three cases: properties sharing the same extension, properties involving

[10] For more on consequences as properties and as reasons see chapter 4.
[11] Orilia and Paolini Paoletti 2022, § 3.
[12] Chapter 4 discusses a model of reason-weighing (Sher 2019) that addresses the issue of double counting.

coreportive proper names, and identical properties. In all of them, learning the logical relation between the two properties rationally constrain the agent to weigh two conflicting evaluations in order to make an overall evaluation of the option. As beliefs change, the consistency requirements of beliefs, including beliefs about the relationship between two properties, constrain the rational evaluation of the option possessing these properties. All of these cases can be handled by model B without violating the invariance principle.

Consequences are thus objects that we should read intensionally, while options are read extensionally. Indeed, as we construct the representation of a decision problem, we may assign to an option new properties that were not considered beforehand. For instance, some consequences are instantiated only in the presence of a particular pair of options (e.g. relational properties, such as fairness). When we talk about a particular option, though we may refer to it by a proposition (e.g. shooting, shooting the fascist) or a proper name (e.g. option A), it does not fix the sense of the proposition. The fact that I call Orwell's option "shooting the nazi" does not mean that I cannot assign to it consequences that are instantiated by another coreferential sense of that option. Consequently, strictly speaking, it is not the case that Orwell simultaneously desired to kill the nazi and spare the human being. He desired to kill the nazi "qua nazi", and spare the human being "qua human being". It may be more rigorous to say that he had reasons to kill the nazi, (namely that he was, as a nazi aiding the conquest of another country) and reasons not to kill him (namely that he is still human and humans should not be killed) or equivalently, reasons to spare the human being and reasons not to. These reasons correspond to the desirability of the properties instantiated by his two descriptions of the act, and all reasons must be considered before establishing the overall desirability of the act, regardless of the sense considered.

One could be worried that this model goes against Frege's intuition[13,14] about the truth conditions of attitude claims. In Frege's view, the truth of a belief claim depends on the relationship between the believer and the sense of the claim, not its reference. The same goes for desire, as you may desire the morning star without desiring the evening star. It might be inferred from Frege's view that, as with beliefs, what is desired is the sense of a proposition and not its reference. In other words, it could be inferred that the desirability of a proposition exclusively depends on its sense. I don't think this inference is correct. If I believe that the morning and the evening star are identical, what determines the desirability of the morning star cannot only be determined by the properties of the sense "the morning star". Frege's motivation for denying extensionality of beliefs and desires is the fact that propositions are not always believed to be coreferential. In this sense beliefs about identity statements play a crucial role in the intensionality of attitudes. Once we believe that p and q are coreferential, we no longer have a reason to have different attitudes to these propositions. Model B allows to retain Frege's intuition about the truth conditions of attitude claims, while refraining from violating extensionality when propositions are believed to be coreferential.

[13] Frege 1892b.
[14] Frege 1892a.

The appeal of this approach is that weak extensionality of desire is kept safe: it does not make Orwell rationally desire one statement to be true and not the other. Orwell may well be torn between options to the extent that the properties instantiated by the two statements are not equally desired or even are incommensurable, but this does not imply that he values differently the same option or fact under different descriptions. As long as the agent has not settled on the desirability and weights of the consequences of an act, the desirability of the act is indeterminate under any description considered. Once the agent has settled, the act has the same desirability under all descriptions. This does not mean that the act of weighing removes all indeterminacy from the agent's point of view: it is only after settling that the evaluative indeterminacy of options dissolves. Settling does not guarantee that the agent's decision is not arbitrary, and this should be expected. In cases of conflicts, one may rationally decide to pick an option without being able to settle on the relative weights of all consequences. These cases and associated philosophical issues (such as bootstrapping) will be treated in the next chapter. Cases where evaluative indeterminacy can be removed non arbitrarily will be addressed in chapter 4.[15]

So far, we have seen that Schick's account connects three key notions: understandings, conflicts, and intensionality. I first focused on the relationship between understandings and conflict, and argued that the existence of rational conflicts does not make a compelling case for understandings. I then turned to the relationship between intensionality and conflict, and contended that consequences with differing desirability do not make a case for licensing the violation of weak extensionality: we can — and should — model situations of conflicting desires without violating weak extensionality, by treating consequences as intensional properties and options extensionally. In the last section of this chapter, I turn to an issue of representation raised by Schick that is fairly independent from his considerations on understanding: the problem of self-delusion.

2.8 Self-delusion and representations: why relevant descriptions cannot be selected based on beliefs or desires

Self-delusion can be defined as the fact of justifying a choice we already made by selecting only one description of the act, on the ground that other descriptions are less desirable. While his account of self-delusion does not defuse the normative worries I raised about understanding, it does provide useful insights for a theory of representation. I think this notion of self-delusion has important bearings on a normative theory of representation. Schick rightly reminds us that we should not pick a description on the ground that it makes us feel better about ourselves. By doing so, he suggests that the desirability of a description does not determine whether or not that description should be included in our representation.

Schick does not formulate a precise normative principle to rule out cases of self-delusion. I think there are two possible ways of addressing the issue. One consists in condemning self-delusion on epistemic grounds, by saying that agents should not distort objective reality by picking

[15] I thank Prof. Thoma for raising these issues.

descriptions that fit their desires, and should thus include all descriptions they believe to be true of a fact. Applied to the criminal's case, it requires him to describe his act both as a crime and as a job, since he believes these descriptions to be correct.

I think this way of analyzing self-delusion is problematic, as it does not distinguish between relevant and irrelevant descriptions of fact. A large number of descriptions may be believed to be true of a given fact, and we can't realistically expect agents to include all of them in the representation of their decision problem. Consider for instance a criminal about to rob a bank. "Robbing the bank", "stealing private property", "getting richer", or even "doing one's job", may all be considered objectively adequate descriptions of the same act. However, some of them are more relevant than others for the decision at hand, and we can't expect agents like us to include all of them on the ground that all propositions are believed to be true. This epistemic condition is thus not necessary because of its normative implausibility, and may even be insufficient at a deeper level. From the agent's internal point of view, self-deception is not only experienced as retrospective acknowledgement of false and incomplete beliefs, but also by the fact that the delusional description does not fully reflect the agent's concerns and evaluations of the fact.

Although self-delusion cannot be prevented by anchoring descriptions to epistemic considerations, I believe it may be avoided by subjecting them to evaluative attitudes. The issue with self-deluding behavior is that desires are illegitimately used to select relevant descriptions. Consequently, the evaluative attitudes that filter relevant descriptions should not be desires, as they must account for the fact that we ought to include certain considerations in our decision problem irrespective of the desirability of these considerations. The caring attitude seems like a good candidate in that respect, as we may care about some outcome although we find it undesirable. For instance, we care about the dire consequences of our actions, no matter how undesirable they are. In the context of self-delusion, the normative intuition prohibiting self-deluding behavior can now be phrased as "agents should not exclude descriptions of an outcome if they care about that outcome under that description". The criminal cannot exclude the moral description of the theft he is about to commit if he cares about the moral value of his action. Analogously, Orwell ought to describe his act as the killing of a fellow human and as fighting fascism if he cares about both.

This proposal of using cares as a guide practical representations is only tentative for now and will be developed in detail in the following chapters. Nevertheless, we should already take note of two worries rising from that conjecture in the case of self-delusion. First, one may be concerned that by anchoring our representations into cares we are only pushing the issue of rational representation back to what we should rationally care about. Chapters 3 and 4 will address this worry in a moderate Humean way by only setting consistency requirements on cares, desires, and beliefs.

Second, one may be concerned that by basing our descriptions of facts on subjective attitudes, we dismiss objective considerations that should play a role in the way we represent a decision

problem. Indeed, it may be argued that agents should be blamed for not describing a situation the way it objectively is. If we ground representational adequacy on cares, the worry goes, we fail to account for these normative intuitions.

This concern can be formulated more clearly once we introduce a distinction between external and internal requirements of representation. This distinction mirrors one made by Bradley (2017) between external and internal requirements of rationality: the former demands that our judgments respond adequately to external reality, while the latter imposes consistency between judgments. An analogous distinction can be made between external and internal requirements of representation. While external requirements give a standard of adequacy between representations and objective reality, internal requirements impose consistency between our representations and our subjective attitudes. I think that if there are any external requirements on representations, they should constrain representations indirectly through judgments: a representation may be externally inadequate only to the extent that the judgments involved are not externally adequate. Indeed, if the representation of a fact is at odds with reality, it must be because some of the agent's beliefs or evaluations do not respond correctly to the way the world is. To see this, consider for instance an individual prone to hallucinations, who sees an apple every time he is offered a pear; he describes the fact of taking the apple as "taking the pear". This individual fails external requirements of representation, as he sees reality in an inadequate way. However, his misdescription of facts derives from the externally inadequate belief that the fruit in front of him is a pear. His judgments of the situation lead him to represent objective reality incorrectly. My point is that externally incorrect descriptions must come from externally incorrect beliefs. This is due to the fact that when agents associate a proposition P to a fact F, they must believe that "P is true of F". Consequently, a misdescription of a fact must result from an inadequate belief about the relationship between a proposition and a fact.

Analogously, a description may be externally inadequate because it derives from evaluative judgments that are not externally adequate. For instance, we may blame someone for describing a fact incorrectly on the grounds that he cares about the wrong features of this fact. The blame here lies in the fact that there is some external standard of what correct evaluative judgments are, and that the agent's evaluations are incorrect by this standard.

Although it is a subjective attitude, I think that caring precisely allows us to separate objective considerations from subjective ones when it comes to assessing agents' representations. What people ought to believe and care about are prudential matters that do not pertain to practical rationality nor representational requirements. A normative account of representation should focus on how agents ought to represent a decision problem independently of what they ought to believe or care about. In the case of self-delusion, it permits to identify internal requirements deriving from consistency between various attitudes, while keeping the issue of their external validity separate.

2.9 Conclusion

In this chapter, I have examined Schick's theory of understandings as decision theoretic representations. I have rejected it on two main grounds: first, the existence of rational conflicts does not make a compelling case for understandings. I then questioned the relationship between intensionality and conflict, and contended that situations of conflicts can be modelled without violating weak extensionality by treating consequences as intensional properties of a particular option. Schick's account nevertheless offers interesting insights on issues of representation of decision problems, and in particular on the relationship between self-delusion and representation. I distinguished between internal and external requirements of representation, and defended an internalist view of representation setting consistency requirements between the agent's attitudes while keeping the issue of their external validity separate.

Moreover, I agree with Schick's view that representations of decision problems are bound to be selective descriptions. If decision theory accepts that rational agents are not omniscient beings that take into account all possible descriptions of a decision situation, it is reasonable to assume that rational representations are selective. Schick made that claim in the case of propositional understandings of options; in the next chapter, we will see that rational representations raise other selection problems even when the invariance principle is respected.

References

Bradley, Richard (2017). *Decision Theory with a Human Face*. Cambridge University Press.

Frege, Gottlob (1892a). "On Sinn and Bedeutung". In: *The Frege Reader*. Ed. by Michael Beaney. Wiley-Blackwell, pp. 151–172.

Frege, Gottlob (1892b). "Über Sinn und Bedeutung". In: *Zeitschrift für Philosophie Und Philosophische Kritik* 100.1, pp. 25–50.

Orilia, Francesco and Michele Paolini Paoletti (2022). "Properties". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2022. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/spr2022/entries/properties/.

Orwell, George (1954). "Looking Back on the Spanish War". In: *A Collection of Essays*. Written in 1942. Garden City: Doubleday.

Schick, Frederic (1991). *Understanding Action: An Essay on Reasons*. Cambridge University Press.

Schick, Frederic (1997). *Making Choices: A Recasting of Decision Theory*. Cambridge University Press.

Schick, Frederic (2003). *Ambiguity and Logic*. Cambridge University Press.

Sher, Itai (2019). "Comparative Value and the Weight of Reasons". In: *Economics & Philosophy* 35.1, pp. 103–158.

Chapter 3

Instrumental representations of decision situations

A plan-based, care-based account of representation

Contents

3.1 Introduction

Our representations of decision situations are partial and selective. In chapter 2, I introduced the distinction between relevant and irrelevant descriptions, and suggested that while a large number of descriptions may be believed to be true of a given fact, we can't realistically expect agents to include all of them in the representation of their decision problem. Bounded cognition is not the only ground for thinking of decision problems as partial representations. In cases of deliberate unawareness, we may choose to remove certain features of a prospect from the description of a decision situation, on the ground that it should not be taken into account in the deliberation process. This is the case for inadmissible evidence in court for instance. Decision representations are thus selective not only because of our bounded cognition, but also because it may be necessary to remove some information from the decision problem.

However, making decision representations selective raises a number of issues. First, the descriptions included in the decision problem should be selected for legitimate reasons. For instance, selecting descriptions of options and consequences based on their desirability (or lack thereof) can be interpreted as self-delusive behavior that principles of rational representations cannot sanction. In other situations, the scope of the decision problem is unclearly defined. To see this, consider the following case.

Roxane is pondering whether or not to sign up to the gym. To that end, she may take into consideration various things such as benefits for her health, financial costs and so on. If she comes to the conclusion that the cost of a subscription is too high, she may stop deliberating and give up on her consideration of going to the gym altogether. She might as well consider other options available to her, such as finding a cheaper place to work out, cutting on other sources of spending, or even starting to earn more money. As Roxane progressively includes additional considerations and options, what was initially a simple decision question becomes a decision problem with unclear boundaries, which may connect with other decision issues as new features are taken into account. Then, it is not even clear whether we can talk about specific decision problems in isolation, or if these separate representations ought to be collapsed into a unique, exhaustive decision problem. How should Roxane rationally represent the decision problem she is facing? The outcome of her future deliberation considerably depends on the way she conceives of the situation.

While this issue would probably be irrelevant for an omniscient decision-maker, it does matter for bounded creatures like us. Our representations of decision situations are partial and selective. While a large number of descriptions may be believed to be true of a given fact, we can't

realistically expect agents to include all of them in the representation of their decision problem. As Roxane's example suggests, there is a tension between on the one hand our incentives to isolate certain considerations to deliberate about a particular decision, and on the other the fact that our decisions are often connected through their consequences. This poses a first question of how extended a decision problem should be; let's call it the problem of scope.

Now, suppose that Roxane's decision to subscribe to the gym depends on what the weather will be like for the rest of the season. If it is sunny, she may prefer to spend her spare time outdoors, and would be happy to work out if the weather is poor. Similarly, her budget for the gym may depend on whether she gets a pay raise or not. As the desirability of an option may be sensitive to particular states of affairs, agents have a rational pressure to take into account these distinctions into the representation of the decision situation. Yet again, a bounded agent cannot reasonably be required to form maximally specific descriptions before making rational decisions. This second example raises the issue of how finely-grained our representations of a certain decision situation should be; I will refer to it as the problem of grain of representation. For a given set of options, how fine grained should a decision problem be?

Both of these issues stem from the fact that our representations of decision situations are partial and selective. While a large number of descriptions may be believed to be true of a given fact, we can't realistically expect agents to include all of them in the representation of their decision problem. An account of rationality for the bounded should provide guiding principles to isolate decision problems in a way that is both realistic and permissible.

Standard decision theory does not address the issues of the grain and scope of our representations of decision problems. This is more generally the case for normative issues of rational rational representation. Without getting into technicalities, Savage's decision theory[1] cannot accommodate the problems of scope and grain as it only applies to what he calls "small worlds", that fully specify the relevant states of the world, consequences and options before deliberation. A theory of representation should thus provide principles of isolation of these small worlds, as well as a plausible story as to how such decision problems arise in our lives as bounded decision-makers.

In this chapter, I advocate a two-stage theory of representation. The first stage stems from what I call the Caring Principle, according to which admissible consequences are objects that make a differential impact which is relevant for the decision-maker if she cares about it. However, this claim is not sufficient on its own to solve the problems of scope and grain presented earlier. I then introduce a second principle of representation, Bratman's Principle, which claims that the decision problems we face are shaped by prior intentions and plans as Bratman (1999) defines them. Combined, these principles offer a plausible solution to the problems of scope and grain, as well as to Resnik's "specification problem". In the last section, I argue that though illuminating, a purely instrumental theory of representation based on intentions needs to be

[1] Savage 1954.

grounded on cares to be immune to various objections aimed at irrational intentions.

3.2 A care-based account of rational consequences

In this section, I explain what kind of things decision-theoretic consequences are, and then defend a first principle of representation. The Caring Principle states that *admissible consequences are objects that make a differential impact which is relevant for the decision-maker if she cares about it*.

3.2.1 Decision-theoretic consequences are objects which make a "relevant differential impact"

Decision situations arise in the perspective of future deliberation between options. In the simplest case, we consider some action and weigh its relevant consequences to comparatively assess the value of acting or not acting accordingly. What do we mean by the relevant consequences of an act? In textbooks or in the lab, the set of options and consequences is often imposed on the agent; however, in real life decisions, agents are required to figure out these relevant consequences for themselves before assessing them. *Before being formalized, decision situations (or informal decision problems) are thus not found but constructed.* Decision-theoretic consequences have an implicit specific property that is not shared with consequences in everyday speech: they are differential features of the act, in the sense that they arise from comparing the impact of performing some act with that of not performing it. If I consider going to the library, a wide range of consequences may be inferred from that act: I will be able to borrow books, to work in a quiet area etc. However, if going to the library is only considered in comparison with, say, going to the office, the outcome "being able to work in a quiet area" is no longer relevant, as both options offer the possibility of working in a quiet space. In simple cases where only one option is considered, the relevant consequences of that option emerge by contrast with status quo. When more than one option is considered, as in the last example, the relevant consequences are established by studying the differential impact of a given option with respect to the other options available. This characteristic of decision-theoretic consequences can be inferred from Consequentialism, a principle holding that actions should only be evaluated on the basis of their consequences. Identical consequences leading to identical evaluation of acts, the relevant consequences for future deliberation must be consequences that differ across acts. Relevant consequences are thus features of the act that make a difference for the agent — relative to status quo, or to other options, depending on the cases considered.

So defined, the relevant consequences of an act include any difference made by that act. Then, things that "make a difference" in decision theory matter to the decision-maker insofar as they are important differences. Frankfurt (1982) studies the relationship between the fact that things are important to us and the fact that they make a difference. He observes that contrary to our intuition, things do not have importance in virtue of the differences they make, as some differences are more important than others. He thus concludes that "*nothing is important unless*

the difference it makes is an important one."[2] I agree with this diagnosis, and believe it is useful for the present inquiry about what makes a decision-theoretic consequence relevant. A relevant consequence is thus not only a feature of our action that makes a difference relative to other options, but it is also necessary that the difference be important to the decision-maker. That difference is in fact relevant if it is important to her. This observation confirms the direction of the relationship between representation and values suggested in chapter 2: If the value of an outcome and the representation of that outcome are codetermined, then, from an internal perspective, values determine how an outcome should be understood, represented or described. In that respect, values are more explanatorily basic than internal decision-theoretic representations. We can now formulate the first principle of representation, or Caring Principle, as follows:

Caring Principle

Admissible consequences are objects that make a differential impact which is relevant for the decision-maker if she cares about it.

In the next section, I refine the notion of relevance described earlier, and propose that decision problems should rationally include properties which a differential impact the agent cares about. I argue that caring is the most plausible attitude on which to ground the notion of relevance.

3.2.2 Caring and rationality

Caring, desires and relevant consequences

So far, I referred to the property of "x being important to us", and that of "us caring about x" in an interchangeable way. Although the two notions usually go hand in hand, there is a significant difference between them. We can say that x is important to us if we believe that x holds some value to us. Unlike caring, importance is an epistemic attitude. One may care about something without believing so. Moreover, one may value some object without caring about it: for instance, I may believe that giving money to charities is important, although I don't care about it. What we care about is what exposes us to gains and losses.[3] In the previous example, not giving to charities may not impact me although it goes against what I value. Disagreement between what you care about and what you value thus tracks a divorce between conative and cognitive attitudes. It is unclear to me whether rationality compels agents to care about objects they value. On the one hand, changing your values to fit what you care about sounds dishonest or inauthentic, which may give a ground for precedence of values over care. On the other, it is not clear whether this issue pertains to either rationality or even agency. This is one way of interpreting Hume's claim about instrumental rationality: *"In short, a passion must be accompanied by some false judgment in order to its being unreasonable; and even then 'tis not the passion, properly*

[2] Frankfurt 1982, § 3, p. 259.
[3] Frankfurt 1982.

speaking, which is unreasonable, but the judgment."[4] Applied to cases of conflict between caring and evaluative judgment, rationality cannot require agents to modify the former attitudes in favor of the latter. Moreover, we have little control over what we care about, and if "ought implies can", we can't expect individuals to start caring about something because it holds value to them.

Both caring and importance differ from desiring. Indeed, one may intensely desire some object without caring about it. While desiring has to do with feeling, caring has to do with the "*more or less stable motivational structures that shape one's preferences and guide and limit one's conduct.*"[5] One key difference between caring and desiring lies in the fact that the former is more persistent in time than the latter, which is why it plays a more stable motivational and action-guiding role. This difference provides an argument in favor of using caring as a guide to representation. Indeed, if decision problems are mediums of deliberation for agents, it makes sense to make them rely on a persistent evaluative attitude that does justice to what agents care about in the long run, rather than on desires which can be temporary and whimsical. For instance, advocates of voluntary euthanasia require that the patient have an "enduring desire to die" in order for the euthanasia request to be valid, and thus insist that a "cooling off period" should be enforced after the patient's request to die. I believe this extreme example illustrates the connection between the persistence of a pro attitude, and the stakes associated with that attitude. The argument is even more compelling in the case of intertemporal choice, where agents ought to take into consideration changes in their conative attitudes over time. Caring allows us to achieve intertemporal coherence in such situations. By distinguishing between momentary and persistent evaluations, it enables rational agents to have sophisticated attitudes towards their desires. Finally, the distinction between caring and desiring accounts for the ambition of control underlying deliberation before decision-making. In decision theory, options can broadly be defined as "the things we choose between or rationally control".

If caring only amounts to persistently desiring, the distinction does not seem compelling enough to justify the introduction of a conceptually distinct attitude such as caring. However, two additional characteristics introduced by Shoemaker's account of agency[6] plead in favor of this distinction. While his main concern is with saving compatibilism from objections about identification, he provides interesting insights on the role of caring attitudes for practical reasoning. His main thesis is that reflective motivation is ultimately grounded on caring. To support it, he observes that motivationally efficacious desires to act stem from our caring about something that my act will help promote. Caring is thus explanatorily and causally basic with respect to efficacious desires. Moreover, when we try to motivate others, we insist on things they care about in the following way: "*if you truly care about X, you ought to* Φ*.*"[7] As caring exposes us to gains, losses, and the emotional states associated with them, we rationalize these emotions by putting forward such cares. These observations have an important implication for prescriptive

[4] Hume 1995, § 2.3.3.6 p. 415.
[5] Frankfurt 1982, § 3, p. 260.
[6] Shoemaker 2003.
[7] Shoemaker 2003, p. 91.

rationality: if prescription involves the possibility of correction, it is made possible by making agents aware of what they care about. This fact also holds for prescriptions based on rational imperatives, according to which our practical reasoning is guided by principles independent from our desires and commitments. When we obey rules of practical reason — and axioms of decision theory may be said to fall within this category —, we are motivated to do so if we care about being guided by reason and reason-related ideals and values. More generally speaking, caring is a marker of reflective deliberation. Our intentional and motivated actions carried out without caring about them are unreflective wanton activities. Conversely, in situations that matter to us, we care about what efficacious desire will drive our action and we engage in self-reflection, thus refusing to be guided by momentary impulses.

First, caring about something does not necessarily make it desirable since we care about gains as well as about losses. We care that some event won't happen or that we don't incur some losses. These facts make caring a better candidate than desires to guide our representations. Indeed, we previously saw that behaviors of self-delusion rely on the fact that we select features of representation based on their desirability. Caring avoids this pitfall, since undesirable features will still be included in our decision problem as long as we care about them.

This last point leads me to the most important distinction between desires and cares. The frustration of a desire may lead to a strong emotional reaction that we identify as a significant loss. However, this does not imply the agent cared about the object of desire, only that he cared about not having such a desire frustrated. If desires constitute the standard of instrumental rationality, the only way to address frustration is to satisfy the relevant desire; by contrast, cares don't disappear after being satisfied, and even persist when they are frustrated to the benefit of a greater care. One can thus rationally suppress a certain desire that is not endorsed by one's cares. This possibility matters particularly in the cases of deliberate unawareness where some feature of a prospect, or desire should be removed from the description of a decision problem.

To illustrate this point, consider a simple temptation case where the agent, once offered candies, has a strong impulse to accept them, though he only cares about his health. In the absence of deliberation, he may well eat the candies and cave to his impulse. The source of irrationality is thus the gap between what he cared about and what he desired. In a standard case of deliberation, he will realize what he cares about most, which will ultimately shift his desire from accepting to declining the candies. If that's the case, he may even consider a way to suppress (or not promote) this impulsive desire for candies. Note that this last possibility would not be achieved by an account substituting cares for second-order desires. Indeed, if the agent plainly desires not to cave to a certain desire, she may very well rationally decide to suppress that second-order desire so as to fully enjoy the satisfaction of the first-order one.

In the previous example, if the agent were still to desire accepting the candies after deliberation, he would exhibit weakness of will. This distinction between impulsive behavior and weakness of will roughly corresponds to Aristotle's two kinds of akrasia (propeteia and weakness). Conscious

or not, one of the purposes of deliberation is thus to access the evaluative standpoint against which desires are assessed.

Caring plays a motivational role in standard non akratic cases. As the agent cares about a particular consequence, she desires the gains and is averse to the losses associated with it. The converse implication does not hold: you may rationally desire that p though you don't care that p. By contrast, in non standard akratic cases we care that p though we don't desire that p. This asymmetry between caring and desiring grants an explanatory status to desires, and a normative status to cares: a choice can always be explained by a desire, whether or not it stems from a care.

So far, I have defended an account of rationality that defines rationally relevant consequences as follows. For a given pair of options A and B, consequence c is a rationally relevant consequence of A if c makes a difference that the decision maker cares about. In standard cases, the existence of such care is sufficient to play a motivational role and trigger the associated desire. Conversely, the agent's decision is deemed irrational when she akratically acts on a desire she cares not to have. That said, akrasia is not the only source of irrationality. While akrasia pertains to practical irrationality, irrational choices may stem from representational irregularities. In the next section, I explore further the relationship between rational representation and rational decision-making.

Akrasia, rational representation, and rational decision-making

When akratic behavior can be explained by an impulse, the impulse may trivially lead to an incorrect representation of the decision problem: in the previous example, the man's desire for candies made him forget about health considerations which matter to him. Conversely, an incorrect representation may lead to impulsive behavior. To see this, consider a modified version of Orwell's example introduced in chapter 2, that I call the Pragmatic Orwell case. Pragmatic Orwell truly cares more about fighting fascism than about sparing this human being. However, the vision of the man with his pants off makes him forget about these considerations and focus on the fact that shooting this man would only amount to killing a human being. Pragmatic Orwell would then be guilty of irrationality in virtue of his inadequate, impulsive representation of the decision problem.

This case should be distinguished from the candy example; indeed, pragmatic Orwell is not a victim of classical impulsive behavior. It is not an impulsive desire that directly moves him not to shoot, it is the fact that he only saw a human being in front of him (and not a human being and a fascist enemy). While the candy case instantiates an impulsive behavior unmediated by representation, pragmatic Orwell acts incorrectly in virtue of his inadequate representation. Folk psychology would then explain his choice via a hasty representation of the situation, which for a moment prevented him from forming a "pragmatic", all-things-considered intention. When the agent's representation is the locus of irrationality, a certain type of psychological explanations of the agent's choice are available, which adopt the agent's point of view in order to "make sense" of the irrational decision; and it is plausible that the agent's representation of the decision

problem fulfills this role.

So much for cases of incorrect representation and impulsive behavior. Alternatively, the agent correctly took into consideration the relevant consequences, yet failed to make a rational decision. This may be the result of either inadequate deliberation or of pure weakness of will, and only in these two cases can the agent be charged with violating *decision* theoretic and agential principles of rationality. Can weakness of will be the product of an incorrect representation of the decision problem? I believe it cannot be the case. Agents display weakness of will when they choose against their all-things-considered, best judgment. Consequently, the locus of irrationality of weakness of will cannot be representational.

I thus considered four possible combinations of irrational actions, depending on whether the irrationality stems from an impulse, from weakness of will, from an incorrect representation, or an incorrect decision. Summarized in table 3.1 on p. 53, these four cases do not constitute an exhaustive typology of irrational representations and decisions; other sources of incorrect representations will be introduced in this dissertation.

Table 3.1. Four possible combinations of irrational actions

<table>
<tr><th></th><th>Impulse (no deliberation)</th><th>Weakness of will</th></tr>
<tr><td>Incorrect representation</td><td>An incorrect impulsive representation leading to an irrational choice.

Pragmatic Orwell example</td><td>Impossible by definition of weakness of will.</td></tr>
<tr><td rowspan="2">Incorrect decision</td><td rowspan="2">An impulsive desire leading to an irrational choice, independently from representation.

Candies example</td><td>Correct representation and incorrect deliberation.</td></tr>
<tr><td>Correct representation, correct deliberation; inability to maximize (pure weakness of will).</td></tr>
</table>

Caring thus allows us to distinguish between impulsive choices stemming from incorrect representations, impulsive choices leading to an incorrect representation, and choices explained by weakness of will. Along with the role caring plays in evaluating and controlling desires, this fact provides grounds for a necessary principle of representation: For a given set of options, only *consequences which make a differential impact the agent cares about* are considered relevant to her decision problem. This principle makes representations of decision problems suitable for their role in effective deliberation, which consists in connecting the agent to the nexus of core attitudes which precedes the weighing procedure of deliberation.

3.2.3 The Caring Principle is not enough: the problems of the scope and grain of representation for the bounded

So far, I argued in favor of the view that consequences are objects that make a differential impact which is relevant for the decision-maker if she cares about it. This view allows agents to include consequences which emerge only in the presence of some alternative options and not others. However, this principle of representation is not sufficient on its own to account for how we construct and represent decision problems. Although it tells us what consequences to include for a given set of options, it does not indicate when an option is relevant to a decision problem.

Let's illustrate this issue with a simple chess problem. During a high-stake game, Rafaella considers moving her threatened rook. Although she sees three possible such moves, none of them secure her rook entirely. Worse, she realizes that partially securing his rook now may put pressure on other valuable pieces this turn and later down the game. These evaluative tradeoffs stem from independent epistemic (objective and subjective), evaluative, and representational uncertainty.

How should a decision theorist model Rafaella's choice problem to help her decide rationally? On the scale of a single turn, securing the rook makes sense, but how about the nth turn's consequences? Without getting into the technicalities of chess theory, the complexity of the game makes it currently insolvable in finite time. If extensively valuing all possible moves is off the table, what should the adviser do? A sensible intuition would recommend anticipating as many turns as possible before deciding. Assuming away the problem of limited time and cognitive resources, should this intuition be understood as a requirement of rationality? In other words, should the scope of our decision problem be as inclusive as possible?

I interpret the problem of scope as the difficulty to find a satisfactory way of picking the right frame for a decision problem, and a particular instance of the structural discrepancy between domain and image of representation of decision situations. Indeed, it introduces two contradictory intuitions that guide our representation strategies. On the one hand, the larger the domain of rational deliberation, the better off the agent after choice. On the other, promoting maximal scope of information goes against another strong normative intuition about human rational agency. Bounded agents often have to make quick deliberative decisions over decision situations that are too evaluatively complex to be extensive.

Rafaella cannot have complete evaluative preferences over the next move without knowing which particular move will contribute best to victory. On what grounds can she specify the same extensional description of the outcome as "securing her rook next turn" rather than "securing her bishop next turn"? Representational standards should allow her to settle for a description of her move that does not include all the possible consequences of all future turns: for creatures like us, decision-theoretic representations can be rational though less than maximally extensive. More generally, the selection of an adequate specification is not trivial because of representational conflicts between globally and locally appropriate extensions of a decision situation. This

corresponds to Rafaella's conflict between considering one game-turn in isolation, all of them or neither. In the gym example, it corresponds to the evaluative interdependence of two local decision situations.

Importantly, an analogous rationale is needed when it comes to determining how finely grained a decision situation should be constructed. Suppose Rafaella intends to secure her bishop for the next two turns. Her evaluation of the particular moves she makes to that end will vary greatly depending on her opponent's reaction, but also on the effect of her moves on the rest of the board. Her move will ultimately depend on how specifically she describes the anticipated consequences of her options.[8]

The adequate scope and grain of decision problems cannot be framed as an optimization program under constraints if we are looking for a justification for either of them. As we will see, Resnik's specification problem shows that such a solution is vulnerable to infinite regress as the representation question becomes a decision problem itself.

A care-based account of representation does not address the problems of scope and grain. It won't tell Rafaella how to frame her next moves before solving her choice problem next turn. Moreover, it does not offer a story of how particular decision problems arise. One could argue that deliberation is about figuring out the best option to get what we care about; yet this only raises more issues: we can't reasonably include all objects of care and options to bring them about in the same decision problem. This would amount to collapsing all decision problems to a single one, as in the gym problem. One could object that we should only include objects we care about most, but I don't think this is a plausible view of what decision problems are. We are often able to suspend judgment about things that are important to us to deliberate about less important questions; using a ranking of objects of care would force decision makers to exclusively focus on important issues. In the next section, I will argue that many decision problems we face are shaped by prior intentions and plans, which may thus provide us with an additional principle of representation.

3.3 The role of prior intentions and plans in representing decision

A possible solution to the issue of scope of representation may be found when we consider the role future intentions and plans play in our lives. Bratman (1999) offers a theory of agency that relies on these attitudes. In *Intention, Plans, and Practical Reason*,[9] he advocates the view that intentions to act are distinct mental states that cannot be reduced to beliefs and desires.[10] In this section, I will introduce the aspects of his theory of practical reason and agency I deem relevant

[8] A technical remark: In Savage's framework, the problem of grain applies both to the set of options and to the set of states of the world. The value or desirability of an option can be highly sensitive to particular uncertain circumstances, which are conditions over which the decision-maker has no control. Analogously, the value of an option may be sensitive to how precisely this option is described. See the next section for an example of the second kind.

[9] Bratman 1987.

[10] From now on, I may use the terms desire and desirability in the general sense of a conative attitude; in order to agree with Bratman's terminology, I temporarily suspend the previous distinction between cares and desires.

to representational issues; I will then defend an account of rational representation based on Bratman's theory, one of its merits being that it offers a satisfactory solution to the problems of scope and grain of representation.

3.3.1 Prior intentions and plans as filters of representation

Prior intentions and plans involve motivational commitments that make them less defeasible than desires. When I intend to take a plane tomorrow, I do not merely desire to do so, as I am in some sense committed to that action: in the absence of interference, I will take that plane tomorrow. Unlike desires, that are potential influencers of my choice, intentions are conduct-controlling pro attitudes. Intentions also play an important role in further reasoning and deliberation. Once I intend to take the plane tomorrow, I will not continue to deliberate about it unless I gather new information or if my desires change. In addition, we frequently reason from prior intentions to further intentions. After intending to take a plane tomorrow, I may deliberate about the appropriate means to get to the airport, intend to take a bus, then consider a particular bus company. Intentions are thus involved in reasoning in different ways: from intended ends to intended means, from general to more specific intentions. Besides, prior intentions and beliefs constrain my further intentions through consistency constraints: "It should be possible to do all that I intend in a world where my beliefs are true".

Bratman defines plans as complex future intentions to act. Intentions are typically parts of larger plans that answer two needs that we have as limited agents. First, the need to deliberate with limited time and resources requires that deliberation be future-oriented and that we plan our decisions in advance. Moreover, achieving complex goals requires intertemporal and interpersonal coordination that can only be possible by embedding our intentions in plans. This is why plans are partial, and hierarchical structures: when I plan to go to the airport, I may leave till later deliberation about the specific means and course of actions to get there. This allows me to isolate certain decisions and fill in the rest of my plan when I will have the time and resources to do so. In addition to being partial, plans are hierarchical in the sense that they are embedded structures. Plans about means are embedded in plans about ends; general plans embed more specific ones: my plan to go on vacation abroad will embed my plan to go specifically to Mexico. Consequently, we deliberate about some parts of plans while holding other parts fixed.

The theory has an important implication for decision problems. Prior plans' partial and hierarchical structure, along with the reasonings that connect them, are such that they *pose problem for further deliberation*. They constitute a background framework in which belief-desire reasons are weighed during deliberation, by providing a filter of admissibility for options. Indeed, Bratman thinks of practical reasoning as a *two-stage structure*. Given my prior plans, I first filter options that are specific, consistent means to achieve them. Only then do I weigh my beliefs and desires to select a particular option among those filtered in the first stage. This two-level structure of practical reasoning renders deliberation tractable: by giving a standard of relevance for options, it narrows the scope of decision and provides "a specific purpose of

deliberation, rather than a simple injunction to do what is best." Additionally, it reconciles the rational demand of satisfying our (rational) desires in the long run with the necessity to take into consideration our cognitive boundaries.

According to this view, decision problems arise in the context of prior intentions and plans. I believe we can infer from it normative principles of practical representation that will address the issue of scope for decision problems. When I deliberate about going to the gym, my deliberation is already embedded in a framework of prior plans. For instance, I may consider that option because I already planned to get in shape. My priors plan constrain the options I will further consider, and consistency across my plans will also prevent me from unlimitedly extending the scope of my options: going to the gym may raise financial issues which give me an incentive to consider further options, like earning more money or adjusting my budget; however, if my current career choice and budget strategy are already part of my prior plans, I can legitimately hold these plans fixed, and isolate the question of going to the gym. Moreover, this instrumental view of decision problems allows us to capture the distinction between alternative options that are compatible with our prior ends (e.g. finding an alternative, cheaper place to work out), and those that may conflict with them (e.g. earning more money). Before offering a systematic account of this view, let's turn to other essential features of Bratman's theory: intentional commitment and reconsideration.

The claim that our goals determine our focus is well supported empirically. Various experiments have highlighted the phenomenon of inattentional blindness, according to which our attention (and our vision in particular) selects objects based on what we aim at. Simons and Chabris (1999) set up a series of experiments, where subjects were asked to perform a task that requires visual attention, like counting the number of passes between basketball players, and observed that subjects systematically failed to detect major changes that occurred in their visual field (in the basketball case, a gorilla walking through the court!). The intuitive interpretation of such experiments is that because attention and perceptions are cognitively costly, we focus on the few specific things that relate to our goals, either because they facilitate or hinder them. Conversely, objects elude our attention when they are not connected to our goals. This line of research suggests that descriptively, the focus of our attention is selective, and driven by what we aim at, which in turn depends on your concerns and values. If this is correct, one could still question the normative status of this model of selective attention and its implication for a theory of representation of decision problems.

3.3.2 Intentional commitment and reconsideration

While intentions and plans influence the way we shape and isolate decision problems, they are not irrevocable for all that. We may and sometimes ought to reconsider our plans, in the light of new information or changes in our desires. This fact raises an objection to the previous proposal to isolate decision problems based on prior plans. In the gym case, I suggested that we hold our prior plans fixed, and that we legitimately isolate the question of going to the gym from that

of my career choices. However, why should I not reconsider my prior plans instead? After all, my plan to get in shape provides me with new considerations that I did not take into account when I initially made my career choices. Reconsideration (or absence of reconsideration) may be deliberative or non deliberative. Expecting bounded agents to systematically reconsider their plans in a deliberative manner is unrealistic. All in all, it would amount to requiring them to form an exhaustive decision problem at all times to assess whether their prior plans are legitimate.

Most of the time, we do not take the time to deliberate about reconsidering our prior plans and intentions: either we stick to our plan in a non deliberative way, or we reconsider it without deliberating. The prevalence of non-deliberative, non-reflective (non) reconsideration is based on two facts about human agency. First, reconsideration has a cognitive cost that agents implicitly take into account. Ideally, we should reconsider our plans only in situations where shifting plans is beneficial and this benefit exceeds the cost of reconsideration. Second, reconsidering plans too often jeopardizes the benefits of intertemporal coordination. Reconsideration should thus include the opportunity cost of failing to achieve a complex goal — although we may have legitimate reasons for it —, as well as the implications for other plans connected to the one under reconsideration. Importantly, the very *act of deliberating as to whether to reconsider or not* amounts to *implicitly reconsidering*: "*in deliberating about whether to reconsider, we already implicitly reconsider*."[11] This new deliberation weighs the initial plans, as well as new options, and the various costs mentioned earlier. Finally, note that the scope of reconsideration can be more or less wide, depending on the intention's hierarchical position in our plans, and its degree of generality (as opposed to specificity). Indeed, we may reconsider some parts of our plans while holding other parts fixed.

3.3.3 Bratman's normative account of practical rationality

Three fundamental dimensions of normative practical rationality

Bratman's account of rationality makes three fundamental distinctions between dimensions along which assessments of practical rationality are made. According to the first distinction, the object of rational assessment can either be an *intentional action or the agent herself*. The second distinguishes two points of view from which rationality assessments are made, depending on whether it is internal or external to the deliberating agent. Finally, Bratman discerns two standards of normative assessments, rational criticizability and rational blameworthiness.

Differentiating the *agent's rationality* from her *intentional actions* allows for the possibility of an agent's being rational in her intentions and actions though she independently fails to comply with some rationality standards. Agential rationality is evaluated from a perspective *external*

[11] Extended quote: "*So in deliberating about whether to reconsider this intention he thereby already implicitly reconsiders it. It is not that my theory precludes such deliberation. Rather, the theory shows us that when the agent has such a belief about the consequences of reconsideration, the mere act of deliberating about reconsidering the prior intention itself may well amount to an implicit reconsideration of that intention. And if we assume that this implicit reconsideration is itself nonreflective, it will be subject to the two-tier theory*" (Bratman 1999, chapter 5)

to her point of view when deliberating, and is thus not of use for the agent when deliberating about what to do next. In evaluating an intentional action, the *internal* perspective takes the agent's prior plans as given and, while the external perspective does so independently from prior plans, intentions. Both perspectives determine the best course of action based on the agent's beliefs and desires, but they differ in the set of admissible options considered: the internal viewpoint restricts the scope of options to those relevant to the agent's prior intentions and plans; the external viewpoint takes into consideration all options available, i.e. all actions the agent believes she can perform. The external perspective allows Bratman to avoid a certain kind of intentional bootstrapping, which consists in rationally sanctioning intended means to irrational prior intentions; conversely, the internal perspective takes into account the cognitive costs required to determine the best alternative independently from prior plans.

When assessing an agent's attitudes and habits, we may pronounce the agent blameworthy with respect to some rationality standards only if it was in her power to make changes in order to comply with these standards. Consequently, one may be *rationally criticizable* without being *blameworthy*. This distinction allows agents to be assessed against (and to aspire to) certain rationality standards, independently from their power to have done otherwise. Imagine that Taliah is driving from Stockholm to Paris, and wrongly decides to take a southern route via Hungary, instead of the fastest northern route via Germany. As she leaves Stockholm, the set of decisions that will lead her to Hungary are criticizable to the extent that she could still exit the southern route, head towards Germany, and make better time; yet given her initial plan to pass by Hungary, she no longer considers options that don't fit that purpose, and thus cannot be blamed for sticking to her — initially faulty — plan. This picture of practical rationality evaluates the agent's blameworthiness from the internal, plan-constrained perspective. Conversely, from the external perspective, rational criticizability applies to any attitudes, habits, or actions she believed she could perform, though they may have become inadmissible from the internal standpoint. That way, the theory gives normative consideration to both belief-desire reasons (which are potential, defeasible influencers of action), and to intentional means-end reasonings (which are conduct-controllers).

Formal constraints on rational intentional choice

With the previous distinctions in mind, we can apprehend the formal constraints that Bratman's standards impose on intentional choice, depending on whether it is the product of deliberation (the simplest case), of general policies, or of unreflective habits.

Deliberative intentional choices are internally assessed by consistency standards of means-end reasoning and prior intentions, after which reason-weighing and rational choice standards apply to choice. By contrast, the external perspective prescribes choosing the best option available to the agent given his beliefs and desires at the time of deliberation, and so, independently from any constraints from prior intentions.

Unreflective, *non-deliberative choices* are choices produced by our general habits, whose standards differ from those assessing *deliberative decisions*. First, unreflective choices are externally rational only if it was rational not to reconsider that intention before choice (see rational (non) reconsideration below). Second, habits themselves can be subject to external and internal assessment. Because habits are general and recurring, Bratman assesses them according to a two-tier approach analogous to rule-utilitarianism, which maximizes the long-run expected impact of a given habit relative to an appropriate threshold. Internally, the set of options (here habits) is plan-constrained, unlike from the external point of view.

Habits of reconsideration (of past plans, habits or any other intentional dispositions) play a privileged role in bounded agency, as they determine the stability of our plans, the firmness of our intentions, but also our dispositions to give up on our plans and intentions when it is beneficial. In order to avoid bootstrapping issues, and because we rarely deliberate about the degree of stability of our plans, habits of reconsideration are only assessed externally, i.e. independently from the agent's deliberative point of view. Then, an agent displays reasonable stability only if his reconsideration habits have a long-term expected satisfaction above a certain threshold.

Finally, general policies of intentions are defeasible general intentions to act under a set of recurring circumstances, such as the policy of cleaning your hands before eating. A policy-based intention is deemed rational only if:

(i) It is in general rational to intend to A when C arises;

(ii) It is rational not to reconsider the general intention;

(iii) It is rational to apply it to the particular case.

These various constraints on rational intentions are summarized in tables 3.2 and 3.3 on pp. 61–62.

3.3.4 From Bratman's normative theory of rational agency to an account of rational representation

We now have a sufficiently detailed account of intentions and plans to elaborate a view of decision-theoretic representation based on it. In what follows I will try to define representations as collections of objects which are (a) intentional, (b) instrumental, (c) plan-constrained, (d) have an intention-like structure, and (e) have a plan-like structure. Indeed, we can conceive of collection of objects playing a functional role similar to future intentions and plans, without being attitudes themselves; they only share their essential properties (commitment pressure, partiality, hierarchy) and can then be used as mediums for our intentions and plans.

To illustrate this idea, it can be useful to think of maps as such objects. Maps are instrumental, intentional objects, in that they are the product of an intentional activity, constitute means to the end of guiding us around the territory or of reaching an intended destination. More importantly, maps can exert pressure on the agent's future actions: unless new geographic information is

Table 3.2. Rationality standards for future intentions and plans (1/2)

<table>
<tr><th></th><th>External standpoint:
• independent from prior plans
• external assessment of the agent</th><th>Internal standpoint:
• plan constrained
• relevant for agent's practical reasoning about what to do</th></tr>
<tr><td>Rational deliberative choice</td><td>Among all options the agent could perform, the one best supported by her belief-desire reasons.</td><td rowspan="2">(i) Means-end coherence and consistency with other prior intentions and beliefs (which determine the set of admissible options).
(ii) Best-supported option among admissible ones.</td></tr>
<tr><td>Rational unreflective choice</td><td>Rational if it was rational for the agent not to reconsider her choice (see rational reconsideration).</td></tr>
<tr><td>Rational policy-based choice</td><td colspan="2">(i) Rational strategy to intend to A when C arises.
(ii) Rational not to reconsider the general intention.
(iii) Rational to apply to this particular case.</td></tr>
</table>

Table 3.3. Rationality standards for future intentions and plans (2/2)

	External standpoint: • independent from prior plans • external assessment of the agent	**Internal standpoint:** • plan constrained • relevant for agent's practical reasoning about what to do
Rational unreflective (non) reconsideration	Two-tier long term expected impact of reconsideration *habits*: Habits of reconsiderations are *reasonable* if they exceed a certain threshold of long-term expected satisfaction.	
Rational deliberative (non) reconsideration: • Second order deliberation including costs and benefits of reconsideration. • Cognitively costly and thus rare.	Reconsider only when the benefits of reconsideration outweighed its gains, independently from prior intentions.	Reconsider only when beneficial to reconsider given the agent's prior plans.
Rational policy-based reconsideration	Particular case of rational policy-based choice.	

acquired, you will endeavor to move about consistently with the information provided by such a map. A map is plan constrained, if it exclusively includes paths determined by prior plans. A collection of maps may have a partial and hierarchical structure. Partial, when an initially sketchy map may be further specified as new information is gathered (this is the case for most maps); hierarchical, when specific maps can be embedded in more general ones (think of the zooming analogy on digital maps), and when mapped paths about the journey can be embedded in mapped paths about the destination.

Then, just as a "theory of rational mapping" would determine which collections of maps are admissible, an account of representation establishes which specifications are permissible. Here, representations are considered at two distinct levels. As Bratman conceives it, the specification of a decision problem is a structure filtering admissible means to our ends. I share his intuition that it is often because of our prior plans that particular decision problems arise: as we ponder about possible means to a given end, the choice of means poses problems for the decision-maker. Moreover, instrumental decision problems may be more sophisticated if alternatives serve several goals simultaneously. Then, the decision-maker should consider options achieving both ends, short of which he should start weighing their relative values. However, a specification can also be apprehended on a different level, as the result of an *intentional choice* over a set of possible representations. Specifications are then viewed as intentional, instrumental *acts* to further goals. Qua *Bratmanian structures*, they are subject to the internal consistency constraints described in subsection 3.3.3 (see internal requirement (i) for both deliberative and unreflective choice in table 3.2 on p. 61). Qua *intentional choices of specification* serving our goals, they are subject to the general standards of intentional actions, both external and internal.

We can now consider collections of decision-theoretic representations which possess properties (a) to (d). Let's call them by simplicity instrumental representations or specifications. Such objects are conceived to satisfy certain ends, and resulting from an intentional activity, guide the agent's future decisions. Moreover, they exclusively include options which constitute means to a prior given end (as described in subsection 3.3.1). They share a partial structure when certain features of the representation can be left to future specification. They are hierarchical when representations of means can be embedded in representations of ends, and when specific representations can be embedded in more general ones. Due to their similar structures, intentions and instrumental representations can be assessed against identical standards of rationality, depending on whether they are produced unreflectively, deliberately, or by a representational policy. The table at the end of this section specifies the standards of rationality applying in each case, illustrated by an example.

As you are driving, the traffic light ahead turns orange. Though you do not have time to think about the best way to frame your decision situation, you instinctively perceive it as: I can either brake and avoid an accident, or speed up and risk it. This is a case of unreflective specification, produced by habits and dispositions. A specification is internally rational if it is consistent with your ends and prior intentions. Since you intend to stay alive, failing to consider the possibility

of an accident would be internally irrational. Externally, its rationality also depends on the rationality of the agent's underlying representational habits in the long run, and their expected impact on her satisfaction. Though most of our specification decisions are unreflective, we sometimes use policies of specification: "when driving, only consider life and death matters" or even "when several specifications are available, pick the one requiring the smallest amount of cognitive resources". Then, Bratman's policy-based standards of rationality apply. Finally, a specification choice may stem from deliberation about how to represent a decision problem. Such deliberation involves reconsideration of past plans, either reflective or unreflective: if reflective, I argue that it should involve the emotional, cognitive and temporal costs of each plan reconsidered. The reason for this is that specifications now constitute deliberate means to the agent's ends; as such it should best serve the agent's goal, in particular by being efficient in its use of agential resources. For instance, suppose a company deliberates about reconsidering its prior objectives, in order to specify a new decision problem: how to react to a market change? Should they bother reconsidering all of their strategy or only part of it? Once the company has determined the perimeter of the prior strategy that should be reconsidered, several specifications are still available (just as several means may satisfy the same end). Then, each remaining specification is assessed via its instrumental value and the agential costs/benefits associated with it.

For now, we have a systematic account of instrumental representation, where each specification is both considered as a Bratmanian filtering structure, and as the product of an intentional action, as tables 3.4 and 3.5 on pp. 65–66 illustrates. In the next section, I will apply more precisely these standards to an important problem for decision theory: Resnik's specification problem.

3.3.5 Instrumental representations offer a solution to Resnik's specification problem

These claims shed light on a philosophical problem for decision theory which Resnik (1987) named the "choice problem of specifications". Broadly put, choosing the best problem specification (or matrix representations) amounts to choosing between decision tables, which in itself is a second-order decision, requiring another problem specification, and thus leading to an infinite regress of decision problems. Resnik concludes that "*any application of decision theory must be based ultimately on choices that are made without its benefits. [. . .] Then, such immediate decisions are not rational, and because of the regress, no decision is rational.*"[12]

Resnik defines these "choices made without their benefits" as immediate decisions, as opposed to deliberative decisions. With this distinction in mind, one can reformulate the specification problem as a rationality trilemma (similar to Agrippa's trilemma[13]): Either a specification is the product of no previous decision, in which case it is not rationally justified; alternatively, if it stems from a deliberate decision, the specification problem is only pushed one step further. If it finally results from an immediate decision, the specification is not rationally justified either. Resnik

[12] Resnik 1987, Chapter 1, p. 11.

[13] See (Comesaña and Klein 2019, section 5) for a detailed introduction to Agrippa's trilemma.

Table 3.4. Rationality standards for instrumental representations (1/2)

How is the representation/specification produced?	Example	Rationality assessment
Unreflective specification	You suddenly brake at the orange light because you instinctively specified the decision problem as: braking (no accident) not braking (accident).	**Internal:** Means-end coherence and strong consistency of the specification with prior intentions (determines the relevant admissible specifications). **External:** Two tier long term expected impact of specification habits.
Policy-based specification	"When several specifications are available, pick the one requiring least cognitive resources." Disposition to brake at an orange traffic light may have different consequences in different situations.	**Internal:** Means-end and prior intention consistency. **External:** (i) Rational specification strategy when *C* arises. (ii) Rational not to reconsider the general intention. (iii) Rational to apply to this particular case. Two tier assessment of nonreflective (non) reconsideration of our prior specifications: long term expected impact of the specification.

Table 3.5. Rationality standards for instrumental representations (2/2)
(The blue boxes refer to joint requirements.)

Deliberative specification	Reconsideration of past plans	Unreflective reconsideration	As a car behind you rushes into you, you instinctively speed past the orange traffic light, reconsidering the specification of the situation described earlier.	**Internal:** Reasonable plan stability based on two tier assessment of reconsideration habits. **External:** Ideal plan stability based on whether reconsideration is beneficial at any time.
		Deliberative reconsideration	A company *deliberates about reconsidering* its prior objectives to specify the new problem: how to react to a market change? Should they bother reconsidering all of their strategy or only part of it? The expected costs of reconsideration are taken into account.	**Internal:** Means-end and prior intention consistency. **External:** Two tier assessment of reflective (non) reconsideration.
	Deliberation about the agential costs and benefits of each new specification available after effective reconsideration		The company has determined the perimeter of the prior strategy that should be reconsidered. Several specifications are still available (just as several means may satisfy the same end). Each remaining specification is assessed via its instrumental value and the agential costs/benefits associated with it.	**Internal:** • Means-end and prior intention consistency. • Best specification among the screened ones, based on its instrumental and agential value. **External:** Among *all* specifications the agent could conceive of, favor the one best supported by her instrumental and agential values.

sketches an answer to the trilemma, by arguing that neither deliberative and non deliberative decision-making is always rational. When stakes are high enough, weighing pros and cons is beneficial; in cases of limited resources (including cognitive and time resources), deliberation would not be rationally recommended. Importantly, he observes that "*we do not decide between immediate and deliberative decision-making on a case by case basis.*" When driving a car, I do not deliberate about how to decide whether I should brake quickly or not. Instead I rely on implicit policies of letting my choices be instinctive in such situations. Resnick objects that we should have a policy for reassessing policies, creating a new regress, but does not provide a further solution to it.

In chapter 3, we saw that using caring as a guide to rational representations raises an important concern, as it pushes the problem of rational representation back to what we should rationally care about and causes regress problems. Resnick's specification problem raises similar concerns, and I hope to address both of them in what follows.

The account elaborated in the previous sections helps clarify and solve Resnik's problem. A first argument regards the general structure of the regress argument. As he concludes that the specification regress leads to the irrationality of all specifications, Resnik presupposes that a decision can only be rational if the associated specification is itself the product of a rational decision. However, one may conceive of a decision problem built on past irrational decisions, which nevertheless can be considered rational *for a given the specification*. This is typically the case for instrumental decision problems, which may be the product of an irrational intention (and decision) though the instrumental part of the plan is rationally conceived and executed from an internal point of view. Conversely, a future intention may be rational though the ensuing plan is not. This distinction between "local" and "global" rationality in fact corresponds to Bratman's internal/external distinction. Consequently, specification can be assessed locally from the internal perspective, or globally from the external perspective.

Secondly, we can now differentiate deliberative from non-deliberative decision specifications.

Deliberative decision specifications result from deliberating about reflectively constructed in the perspective of deliberation. A company deciding how to react to market changes may deliberate about how to represent the problem at hand. This deliberation may question two aspects of representation. First, it may involve reconsidering all previous plans and commitments which constrain the scope and grain of representation, in which case the external and the internal points of view coincide. The company may reconsider all previous commitments, and may include all the options relevant to a given past plans and habits. Second, at the higher order of reflection, the agent may investigate the tradeoff between the (cognitive) costs and (instrumental) benefits of initiating a reconsideration of your past plans. Typically, our representations are reflectively constructed when the stakes involved are high enough. Note that reconsidering one's plans does not imply falling over again into Resnik's specification regress: the newly formed plans and intentions will constrain the specification of the associated decision problem. Moreover, the

rationality of the reconsideration itself is not subject to specific deliberative standards. This is due to the fact that deliberation about whether to reconsider or not amounts to reconsidering; this reconsideration is unreflective and thus subject to the two-tier standards for nonreflective reconsideration.[14]

Non-deliberative decision specifications arise from non-reconsidered plans and habits. Most of our decision problems fall into this category. When buying milk at the supermarket, the number of options and tradeoffs one may face is such that it would be overdemanding to reconsider our habits every time a new product appears on the shelf. Analogously, we cannot reconsider whether to gather further information about each product in order for our representation to be more exhaustive. Similarly, changing supermarkets (and thus reconsidering) in order to have a wider set of options cannot be rationally required of bounded agents. In such cases, the specification is determined by our previous plan; and whether or not this specification is the best available (the first step of Resnik's regress problem) pertains to external standards of rationality.

From the internal perspective, the question of the adequacy of the plan-constrained specification is answered:

(i) Given a particular specification, by assessing whether it allows you to reach your ends in a satisfactory way. You may have found an all-things-considered better option by finding a larger supermarket, but this one is good enough for the purpose of bringing milk back home for breakfast. The solution I defended in subsection 3.3.6 to the problems of scope and grain is a special case of this general instrumental principle of representations.

(ii) In general, by assessing the long term impact of our representational habits: if you never consider cheaper options at the supermarket, it may have financial consequences you care about. This two-tier standard is, as defended by Bratman, satisficing rather than maximizing. The long term impact of our representational habits is then evaluated against a satisfaction threshold. This view coheres with the previous distinction between satisficing standards for search (representational) problems and maximizing standards for decision problems.

In both cases, the standards of adequate specification are not maximizing but satisficing, and thus differ from standard decision-theoretic standards (maximization principle). Consequently, a rationally justified specification is not necessarily the product of rational decision-making. It may be either justified by satisficing means-ends reasoning, or by the long term impact of the representational habits leading to that specification.

Let's take stock of what I have argued so far. First, rationally justified decisions may be based on irrational ones, in virtue of Bratman's distinction between internal and external rationality standards. Second, specifications resulting from deliberation do not face a justification problem,

[14]Technically, deliberation about specification and deliberation about reconsideration of past commitments are not the same thing. Nevertheless, when stakes are high enough, our incentives to deliberate about specification and incentives to deliberate about reconsidering past plans go hand in hand.

as they reconsider all previous plans and policies, in which case the specification is the product of rational decision-making (the external and internal standards then coincide). In the non deliberative case, the adequacy of a particular specification can be justified by satisficing means-end standards: a specification is then adequate if it includes satisficing means to the agent's end. A more specific version of this general principle is the solution I defended earlier to the problems of scope and grain under certain structural assumptions. Moreover, specifications resulting from non reconsidered representational habits, are assessed by the satisficing, long term impact of these habits.

With these distinctions in mind, Resnik's specification problem can be solved thanks to the fact that policy-based, deliberative, and immediate specifications are subject to distinct rationality standards. The fact that immediate specifications may be rationally justified provides a stopping point to the justification regress. Indeed, an immediate specification is not determined by an anterior choice between a set of possible specifications. Its justifiability does not rely on higher order or anterior justifications, but on consistency considerations internally, and on long term expected impact externally. These standards are not maximizing but satisficing: they do not demand to adopt the best specification habits, but good enough ones. Nevertheless, Resnick raises another objection which may be immune to my last argument. Independently from their respective justifiability, each type of specification strategy offers advantages and disadvantages depending on the stakes, as well as on the cognitive, temporal and emotional resources available. However, Resnick argues, "*we do not decide between immediate and deliberative decision-making on a case by case basis*", instead we rely on implicit policies of letting my choices be instinctive in such situations. Then, we should have a policy for reassessing policies, creating a new regress. Note that the typology of specification strategy does not address this worry, as the new regress does not stem from a specification choice, but from deciding whether or not my specification choice should be reflective.

I don't think this new regress is problematic either for the present account of representation. As we saw earlier, Bratman's policies have their own rationality standards, as do nonreflective (non) reconsideration of our policies. And as in the immediate case, these standards are not maximizing but satisficing. As long as these standards are respected, both immediate specifications (in the first regress), and nonreflective (non) reconsideration of our policies (second regress) offer a final stopping point which does not require further rational justification.

3.3.6 A tentative, instrumental solution to the problems of scope and grain

In section 3.2, I introduced what I called the problems of scope and grain representation, which both question the legitimacy of small-world decision problems in the light of the fact that most of our choices are interdependent in some ways. I illustrated that point with the gym example, where the consideration of a gym subscription ultimately depends on your financial or even career choices. Now that we are equipped with a tentative account of instrumental representation, let's see how it answers the problems of scope and grain. First, Bratman's

planning theory has an immediate implication on the issue of scope. By providing a filter of admissibility for options, a plan-based view of representation would state that only options that are inferred from our prior plans, through means-end reasoning, should be included in our decision problems. In the gym example, unless you reconsider your anterior commitments, subscriptions which don't cohere with your prior financial and professional plans will not be taken into account. To the question of grain of representation, this view would suggest that states of the world should be individuated according to their relevance in our mean-end reasoning: if different states induce different instrumental desirability with respect to our prior intended ends, they should be individuated accordingly.

The appeal of this account is that, contrary to other principles of individuations, agents are not required to differentiate states of the world every time they lead to a different desirability of outcome. Only those states leading to different desirability with respect to prior ends are required to be individuated. For instance, if I already planned to go on a trip to Paris, and I am considering whether to go by train or by plane, I will differentiate states of the world which might affect my chances of reaching my destination (say the eventuality of a storm, or a train strike), but I won't be (internally) required to distinguish circumstances that will affect the desirability of the journey itself: even I prefer traveling first class, the account does not prescribe that I discern cases where there are no more first class seats and available.

Moreover, means may either be selected on the basis of their degree of instrumental sufficiency or necessity with respect to an end. Although this process of selection is usually thought of as an epistemic one, we will see in chapter 4 that it raises problems of irreducible evaluative uncertainty.

This instrumental principle of individuation allows us to drastically reduce the number of acts, consequences, and states of the world, while making the decision problem sufficiently exhaustive and fine-grained to guarantee that I succeed at fulfilling my end. Its obvious drawback lies in the fact that I may be rationally allowed to omit circumstances that may gravely impact the desirability of my options. If I am prone to flight sickness in case of bad weather, the principle won't compel me to take it into consideration as I individuate the state space. This way of determining the grain of representation may appear unrealistic as it neglects our desires. However, it allows me to distinguish between desires I commit to by giving them the status of ends from mere penchants by which I don't want my decisions to be affected. Moreover, many desires actually are implicitly intended ends: in the previous example, I may have implicitly intended to avoid being flight sick as much as possible; if so, coherence between my prior ends will require that I factor it in when individuating states of the world. Finally, the distinction between internal and external standards captures the fact that habits of reconsideration (and thus of representation) can be criticizable (for not serving your desires in the long run) though the bounded planning agent did not have it in his power to do otherwise. One merit of this account is that it captures the pragmatic intuition we have about instrumental rationality under bounded cognition. When we lack time and resources to consider all details of a situation, we

prioritize by only dedicating our attention to features that will be decisive to the completion of our goals.

Although Bratman's filter provides a principle of isolation, it does not entirely solve the problem of scope just yet. If I initially planned on a particular budget, and now intend to get a gym subscription to be in shape, my two intentions can potentially conflict. The natural solution offered by planning is to stick to my initial budget and only consider gym subscriptions which agree with it. However, I may partially reconsider the said budget for the sake of my health: as we've seen earlier, reconsidering consists in "partially lifting the filter of admissibility of options", and can be more or less comprehensive. The default strategy of never reconsidering amounts to giving precedence to past goals over new ones, and one can easily understand why it is not a satisfactory long term habit of reconsideration. Permanent reconsideration is not sustainable either given its cognitive cost, and the opportunity cost created by the absence of commitment. So all in all, Bratman offers a spectrum of possibilities to address the problems of scope and grain, some probably more adequate than others depending on the situation, but all internally rational nonetheless.

In subsection 3.3.5, I introduced a second level of analysis of decision problem specifications, which emerge as we consider representation as an intentional action. When this act is not reflective, only the long term impact of our habits of reconsideration is under rational scrutiny. However, my main concern in this section regards the scope and granularity of deliberative specification of choices and policies: when we deliberate about the best way to represent a problem, how should we trade agential resources against the satisfaction of our goals? The systematic account of specification I proposed in section 3.3.4 does stipulate that agential resources should be taken into consideration in such cases; for instance, Pareto-inferior specifications with respect to goal satisfaction and agential resources should not be chosen. However, little more can be said in the general case. Can we try to characterize these more "adequate" solutions any further?

The gym example illustrates the fact that satisfactory policies of reconsideration should take into account the particular stakes (determined by cares in the present account) raised by possible options: ceteris paribus, the more we care about certain intended ends, the more they should resist reconsideration. In the long run, failing to give more inertia to options with higher stakes will prevent the agent from getting what they want. Cares may constrain our propensity to reconsider our prior ends via various policies. The most radical one is a care-based lexicographic policy, which exclusively takes into consideration the goal she cares most about and reconsiders any other ends. In the gym example, if the agent cares most about her budget, she will reconsider her plan to work out if she only finds options clashing with the budget she committed to. A less extreme policy would constrain the frequency at which certain goals should be achieved or reconsidered: ceteris paribus, the more you care about an end, the more frequently you should take it into consideration in your decision problem.

Stakes are not the only criteria that should constrain reconsideration policies. The supermarket example suggests that strategies of reconsiderations based on resource management play an important role in bounded rationality. We (reflectively or not) choose not to reconsider the scope and grain of representation when we believe it would be too costly in agential resources.[15]

By considering specifications as filters and as deliberative actions, we've eliminated a number of possible specifications; yet the scope and grain remain largely underdetermined by these constraints. One reason for this is that I have apprehended deliberative representation as a *one-shot, static action*. By considering more sophisticated representational strategies such as dynamic ones, agents can benefit from iterative feedback and handle the issues of scope and grain more efficiently, as the following two examples suggest.

A week before leaving for a year-long trip, you offer to your dear friend to have a drink on Tuesday. He replies that he is busy that day and suggests another day of the week. The exchange goes on, and you come to the conclusion that there is no day of the week where both of you are available. As you think you have reached a deadend, you realize you could fit a quick lunch between two meetings on Thursday, and eventually manage to say goodbye to your friend. This example illustrates the kind of reasoning at work as you are trying to figure out adequate means to your end, that the scope of your means is defined (here, a meeting with your friend some day next week, or no meeting), but that the *grain* of the options and outcomes is undefined yet. In such cases, the example suggests that we go from coarsest to finest grain until an option constitutes an adequate means to your ends: at first, you may only consider the days of the week where you are entirely free, then the ones where you are available in the morning/afternoon, and then consider shorter time slots to find a shorter time slot you could both agree on. Note that in this example, a shorter time slot would only partially satisfy your end: you would much rather have taken the time to see him for a whole day. Nevertheless, a two hour lunch seems good enough to you, and you'd rather do so than not see your friend altogether.

After the abduction of a man, the police inspector in charge of the investigation decides to set a search perimeter around the location where the victim was last seen. The search does not make any headway and the inspector decides to extend its perimeter. Analogously to the previous one, this second example illustrates the fact that we go from less to more extended scope of options as we seek adequate means to our ends. Again, one could imagine that extending the scope of the search has a price, such as reduced accuracy or pace of the search. Yet as suitable means become hard to find, the inspector may accept to lower his demands regarding some dimensions of his end.

These examples support two important claims about the way grain and scope of representation relate to prospective means-end reasoning. The first claim regards decision situations where

[15] Pettit's notion of "virtual deliberation" fits this picture of human agency well. According to Pettit (2010), deliberation may have an active or a virtual influence on our decisions. Virtual control only triggers conscious deliberation when "something goes wrong" in the plans we've been unreflectively executing. Virtual deliberation can be interpreted as a reconsideration policy which can be applied to a variety of situations, and whose rationale is grounded on considerations of resource management.

the space of options is (i) nested for grain (ii) measurable for scope. Then, bounded agents may start with considering a coarser grain and more restrained scope of options, states, and outcomes. Depending on how *satisfactory* these represented means are for our end, agents may either settle for one of the means so described, refine the grain and extend the scope of the means, or else reconsider their initial end. Second, our ends are not always structured in a binary, all-or-nothing way. This leaves room for tradeoffs between the instrumental value of your means, partial as it may be, and the agential costs involved in iterating the refinement procedure. As we construct a decision problem at the service of an end, increasing the quantity and quality of relevant information has a cognitive price, and may decrease the instrumental value (or even desirability) of the options.

A dynamic procedure can be decomposed into several steps along which new information is gathered and used by further steps of the procedure. While iterative dynamic procedures repeatedly test candidate solutions until one is found; recursive dynamic procedures determine the solution by breaking down the problem into smaller subproblems, until it reaches a solvable or already solved subproblem. The plan-based filtering of admissible options can be used statically, while the lunch meeting and manhunt examples are cases of recursive problem solving. Indeed, they consist in incrementally refining the grain of representation and of expanding its scope, until an instrumentally satisfactory solution is reached. Finally, this procedure of refinement can be constrained by our strategies of agential resource management in various ways: by introducing a resource-based stopping point after which no further refinement is made; by fixing a limit on the number of future recursions to be made; or even on the increment of refinement between two successive recursive steps.

Bratman's theory can thus be used to build a systematic account of instrumental representation that offers an intuitive solution to Resnik's specification problem, and to the question of the scope and grain of decision problems. Furthermore, it sketches the structure of more sophisticated, iterated representational strategies for bounded agents under certain assumptions. In the next section, we will see that illuminating as it may appear, a purely instrumental theory of representation based on intentions needs to be grounded on cares to be immune to various objections aimed at irrational intentions.

3.4 Intentions and plans raise challenges which can be addressed by the Caring Principle

Illuminating as it may be, using Bratman's instrumental account as a general framework of rational representation gives rise to a few difficulties. In this section I will present two of them. First, it overlooks the differences between the instrumental and the non instrumental value of a particular state of affairs. Besides, it indirectly gives too much importance to desires through intentions, which both can be arbitrary and capricious. I then argue that *an intention-based account of representation can only address the objections raised at irrational intentions if it is grounded on*

the agent's cares.

3.4.1 The instrumental view of rationality presented by Bratman fails to distinguish the instrumental from the non instrumental value of ends

Several issues may arise from holding an exclusively instrumental view of rationality. If agents are not required to reconsider their plans, they may start caring more about the means than about their prior intended end. Various criticisms of instrumentality rely on this phenomenon: as we attempt to achieve some goal, the means we employ to that end acquire the status of ends themselves. Commitment and lack of reconsideration then compel us to promote those means regardless of the initial end they were intended for. While omniscient decision-makers would be immune to these issues, bounded individuals precisely use plans to isolate decision problems and to suspend judgment about the reconsideration of their prior plans. However, by forming intentions about the means to their ends, they may start to value these means independently from their initial purpose. If for contingent reasons, the means are no longer fit for their ends, agents may still pursue them because of their commitment to those means, which become ends in themselves. These phenomena can be explained by limited attention and memory, combined with our propensity to embed and isolate smaller plans into larger ones. Unfortunately, Bratman's account of planning does not allow decision-makers to distinguish ends that are intrinsically valuable in the decision-maker's eye, from means that acquired the status of ends although they only hold instrumental value. The only way to question these means in Bratman's view is via reconsideration. However constant reconsideration cannot be rationally required of agents. More importantly, reconsideration will only allow the agent to weigh the means against prior ends, without taking into account the fact that these means were of no intrinsic value for the decision-maker in the first place.

To illustrate this point, consider for instance a bounded planner caring about health, who intends to find a treatment against some disease. Among the alternatives he faces, suppose some are efficient against the disease but with heavy side effects, while others are less efficient but with substantial benefits for health. According to Bratman's two-stage structure of practical reasoning, only the efficient treatments will be considered. However, this leads the agent to omit the fact that the value of such treatment is bound to its benefits on health. Had the agent taken that fact into consideration, he would have weighed the merits of each treatment not on their efficiency against a specific disease but on their overall impact on health, which was the initially intended goal. Reconsideration will then weigh the value of finding an efficient treatment against side effects on health, rather than its overall effect on health: as finding a treatment has become a goal, it acquires intrinsic value. Bratman does impose consistency constraints across plans, which should force agents to only consider means that agree with all of our prior commitments; however, as soon as a means is intended, it becomes part of prior commitments and is indistinguishable from intended ends. In the previous example, health and finding an efficient treatment have both become goals on an equal footing.

One way to distinguish intended means from intended ends is to require that agents be aware of their higher order ends when evaluating the merits of their means: in our example, by keeping in mind that finding a treatment, before being an end, is already a means to the goal of preserving and enhancing health. This creates a hierarchy between ends, in which higher order, non instrumental ends take precedence over intended means. An immediate issue with this solution is the threat of infinite regress of ends: when should we stop considering higher order ends in our evaluation of subsequent means? One could address that worry by resorting to the existence of foundational, non instrumental ends: it could be argued that health is an end that should *prudentially* be promoted for itself. The requirement would then be that when foundational ends are available, they should take precedence lower order instrumental ends. However, it is not clear whether we are always aware of our non instrumental ends, and it seems overdemanding to expect that agents establish their non instrumental ends before making decisions. Instead of making it an unconditional requirement of rationality, one could think of it as a model of rational sophistication, which puts pressure on agents to take into account higher order ends as much as they can, without blaming them with irrationality for all that.

3.4.2 Capricious and arbitrary intentions

Allowing intentions to play a role in rational representations raises a number of issues for normative decision theory. We have seen that by filtering options and outcomes, intentions and plans partially constrain the scope of our decision problems. One rationale for this lies in the fact that intentions and plans are more stable than desires, allowing us to coordinate intertemporally and achieve complex goals. However, one may worry that intentions may be unstable and capricious just like desires. Consider the case of a young pianist forming the intention to become a professional musician. This intention should entail some form of commitment towards the goal he is planning to achieve. Suppose yet that the young man's intention is capricious: although he knows what it takes to become a professional pianist, he fails to follow through with his plan. According to our current theory of representation, his initial intention should lead him to screen options that constitute means to his end. Yet one could wonder whether this filtering principle is misguided here: if he is indeed unable to stick to his plan, the capricious pianist should consider other career options available to him before deliberating about the relevant means to his goal. In other words, decision problems cannot safely be defined by intentions if these attitudes are as unstable as desires.

There are several ways of interpreting this example. First, it could be argued that the capricious pianist didn't think through his plan and the sacrifices it took to achieve it. In that case, his initial intention is not in question: on the contrary, he should have considered more carefully the means necessary to his end, and reconsidered them if they were too costly or unrealistic. Alternatively, it could be the case that while he did consider the implications of his goal, he did not intend the means to achieve that end. Then, I believe it is safe to say that he did not fully intend to become a pianist: one cannot intend an end without intending the known and necessary means

to that end. If this is correct, the man's wish to become a pianist is not a full-fledged intention. Thirdly, the pianist may have intended to commit to his plan despite the sacrifices, but couldn't follow through because of weakness of will. If this is the case, the issue is with agency, rather than with his intention or the representation of the decision problem he will construct based on that intention (about the appropriate means to that end). Consequently, none of these three interpretations threaten an intention-based account of representation.

A more significant worry for representation lies in the fact that intentions may be arbitrary, in the sense that they may be disconnected from the agent's true concerns. To illustrate this point, consider an alternative version of the capricious pianist (let's call him the arbitrary pianist), who cares about becoming a football player, but forms an intention to become a professional musician in an ad hoc way. For the sake of the argument, let's assume that while he does care about becoming a football player, he did not yet intend so, nor did he form a plan about it. Consequently, his intention to become a pianist does not conflict with his prior intentions and plans; the problem is simply that while his intention may rely on a desire, it does not connect with his deeper concerns. This example raises the worry that if intentions may be arbitrary, they cannot legitimately ground rational representation. The arbitrary pianist, the worry goes, should have deliberated about the adequate means to promote what he cares about, rather than what he arbitrarily intended. As odd as such an individual may seem, I don't think he is rationally at fault if his intentions and plans are not inconsistent. What feeds our worry is the *commitment* we implicitly make to attend to what we care about, at least beyond a certain threshold. As long as the arbitrary pianist did not make such a commitment, his intention and subsequent choices should not be deemed irrational. Nevertheless, one could argue that it is a necessary requirement of rationality to commit to attend to what we care about. For now, I will set this issue aside, and examine other aspects of the relationship between caring and intentions in the context of rational agency.

3.4.3 Caring guarantees self-identification by pinning down attitudes that are external to the agent; it thus gives a legitimate anchor to our intentions

In order to understand the relationship between caring and identification for our present concern, let me elaborate on the role caring plays on agency. Classical compatibilist accounts of agency assert that free agents are agents whose actions coincide with what motivates them to act. Contemporary compatibilist accounts (Frankfurt (1982) and Watson (1975) to name a few) recognize that this condition is insufficient to guarantee free agency, since our effective motivation or desire may sometimes not correspond to what the person truly wants, as in cases of compulsion or addiction. In such situations, the desire can be considered external to the psychological self; compatibilist accounts then seek to define free agency as a sort of harmony between the various psychological elements of the self. However, models of agency that ground identification on higher order desires (as Frankfurt's) or on evaluations (as Watson's) fail to account for the notion of identification: in Frankfurt's case, it is unclear why higher order desires

should be granted special authority to determine what counts as internal or external desires. Watson's theory faces a similar challenge, as rational agents may not identify with what they judge best. As introduced in section 3.2, Shoemaker's solution to the problem of identification relies on the fact that at the time of decision, what we care about and makes us emotionally invested holds special authority to determine which psychological elements we should identify with and which we should deem external to ourselves. When a motivationally efficacious desire is independent or inconsistent with what I care about, it becomes external to me.

Whether or not we agree with this account of free agency, the relationship between objects of care and identification has substantial implications for normative rationality. Conscious biases of rationality can be considered external to the self, in the sense that agents no longer identify with them. In cases of deliberate unawareness, individuals do not identify with certain beliefs and desires that would affect their judgments and choose to remove them from their representation (see chapter 2 for a further discussion of deliberate unawareness). The same goes for phenomena of time bias or temptation. Applied to such situations, Shoemaker's account suggests that what makes agents identify or dissociate from biased attitudes is what they care about. I believe that this dependency of rational attitudes on cares also applies to intentions. Among the issues I previously raised about intentions, the fact that intentions may be arbitrary and that intended means become indistinguishable from intended ends (despite having no intrinsic value) may be interpreted as problems of identification between the agent and his intentions. These difficulties seem to dissolve once intentions are anchored to the agent's objects of care: if intentions do not conflict with what the agent cares about, they no longer are arbitrary; if intended means are assessed with respect to cares, they can be distinguished from intended ends and be assigned the value they deserve.

Consequently, The Caring Principle introduced in section 3.2 allows instrumental decision problems to include features that are not filtered by the agent's instrumental future intentions. Indeed, if before deliberation an object of care arises that was not selected for its instrumental value, the agent will be rationally criticizable (as explained in section 3.3).

Finally, some hope remains to address the problem of hierarchy of ends mentioned earlier. Clearly, we care more about some things than about others, so objects of care may be partially ordered. However, it seems that caring has a justificatory role for ends which desiring doesn't possess. Indeed, we often appeal to the fact that "we care about x" as an ultimate explanation to our plans, and so, quite independently from the degree to which we care about x. By contrast, in natural language, desires seem to play a less convincing justificatory role, and so quite independently from their intensity. Though the issue of hierarchy of ends and the threat of regress persist, it seems that caring plays a role to mitigate them.

Rosner (1998) introduces an example that is illuminating with respect to the relationship between intention and care. A niece has to decide whether to euthanize her aunt. She is aware of two possible motives of doing so: to relieve her aunt of the pain and to get rid of the old lady. While

the two ends lead to the same action, she cares about being moved by the first end but not by the second. Just like desires, intentions may thus be subject to endorsement and identification, and caring seems to be a legitimate way of grounding our pro attitudes. While Bratman argues that the rationale of intentions and planning is to get what we desire, Rosner's example, along with the difficulties I introduced in this chapter suggest that rational agency should be achieved by anchoring intentions to our objects of care.

This last claim is supported by another important concern. Along with Bratman, I argued that belief-desire reasons are more defeasible than intentions, which gives legitimacy for a hybrid account in which reasons play a subsidiary role with respect to plans. However, this argument no longer goes through if the legitimacy of plans themselves is grounded on their instrumental value at getting what we desire. Since as we have seen, intentions may be capricious, rigid and in many cases, simply the expression of unreflective desires, this normative priority of intentions over reasons loses its plausibility, and the intention-based account its general appeal. By requiring our intentions to be care-based, we avoid the pitfall of a theory of representation promoting the maximal satisfaction of agents' most capricious desires independently of whether they care about such desires.

In subsection 3.2.2, I considered four possible combinations of irrational actions, depending on whether the irrationality stems from an impulse, from weakness of will, from an incorrect representation, or an incorrect decision. Combined with instrumental planning as presented in section 3.3, the possibility of controlling desires that are not endorsed by one's cares can be rationalized by this account of representation. Yet one may view in this definition of representation a more specific way of overcoming temptation. Indeed, one may conclude further that the very act of representing decision situations that way constitutes an effective means to overcome temptation. However, whether or not correct decision representations are efficient motivational tools is for the most part an empirical issue. Consequently, this account allows us to make conceptual sense of that possibility but does not commit to it.

3.5 Conclusion

Let's take stock. In chapter 1, I introduced the question raised by this dissertation: what are the rationality standards of decision-theoretic representations? Upon asking what counts as a rational representation of a decision situation, I examined several philosophical issues raised by framing effects. First, how stable should our representations of outcomes be? And then, how robust should our attitudes be to such representational shifts?

Chapter 2 then examined Schick's theory of understandings as decision theoretic representations, that allowed conflicted agents to represent the decision situation in a selective way that violated the invariance principle. I rejected it on two main grounds: the existence of rational conflicts does not make a compelling case for understandings, and situations of conflicts can be modelled without violating weak extensionality. Schick's account nevertheless offers interesting insights

on issues of representation of decision problems, and in particular on the relationship between self-delusion and representation. I distinguished between internal and external requirements of representation, and defended an internalist view of representation setting consistency requirements between the agent's attitudes while keeping the issue of their external validity separate.

Along with Schick, I argued that decision problems are selective representations: if decision theory accepts that rational agents are not omniscient beings that take into account all possible descriptions of a decision situation, it is reasonable to assume that rational representations are selective. While Schick made that claim in the case of propositional understandings of options, I argued in the present chapter that two major issues of selection arise from what I called the problems of the scope and the granularity of decision problems. Indeed, a tension exists between on the one hand our incentives to isolate certain considerations to deliberate about a particular decision, and on the other the fact that our decisions are often connected through their consequences. While an omniscient decision-maker would probably not face the issue, an account of rationality for the bounded should provide guiding principles to isolate decision problems in a way that is both realistic and permissible.

I then advocated a two-stage view of representation. I first defended what I called the Caring Principle, that admissible consequences are objects that make a differential impact which is relevant for the decision-maker if she cares about it. I defended the introduction of cares in addition to desires and offered a model of motivation, control, and akrasia to distinguish irrational representations stemming from impulse and from weakness of will.

However, this first stage is not sufficient on its own to solve the problems of scope and of granularity. I then introduced a second principle of representation which claims that the decision problems we face are shaped by prior intentions and plans as Bratman (1999) defines them. Combined, these principles offer a plausible solution to the problems aforementioned, as well as to Resnik's "specification problem". Finally, I contended that an intention-based account of representation can only address the objections raised at irrational intentions if it is grounded on the agent's cares.

This second stage introduced another kind of irrational representations in addition to impulses and weakness of will: instrumental representations (either unreflective, policy-based, or deliberative) failing to satisfy the requirements of specification presented in section 3.3.4. Moreover, it distinguished external and internal standards of assessments of representations. The former are standards of rational criticizability, while only the latter constitute actual standards of rational blameworthiness. These requirements show how one should rationally filter options and set how finely grained states of the world should be. In the following chapter, we will see that the value or desirability of an option can be highly sensitive to particular uncertain circumstances, which are conditions over which the decision-maker has no control, but that he may rationally take into account. While the account of instrumental representation defended

in this chapter interpreted such cases as cases of reconsideration of the specification of the decision situation, the following chapter will explore the issues involved with context-sensitive instantiation and evaluations of consequences.

References

Bratman, Michael (1987). *Intention, Plans, and Practical Reason*. Cambridge: Cambridge, MA: Harvard University Press.

Bratman, Michael (1999). *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge University Press.

Comesaña, Juan and Peter Klein (2019). "Skepticism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2019. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2019/entries/skepticism/.

Frankfurt, Harry (1982). "The Importance of What We Care About". In: *Synthese* 53.2, pp. 257–272.

Hume, David (1995). *A Treatise of Human Nature (1739-40)*. Past masters. InteLex Corporation.

Pettit, Philip (2010). "Deliberation and Decision". In: *A Companion to the Philosophy of Action*. Ed. by Constantine Sandis and Timothy O'Connor. Blackwell, pp. 252–258.

Resnik, Michael D. (1987). *Choices: An Introduction to Decision Theory*. University of Minnesota Press.

Rosner, Jennifer Amy (1998). "Reflective Evaluation, Autonomy, and Self-knowledge". PhD thesis. Stanford University.

Savage, Leonard J. (1954). *The Foundations of Statistics*. Wiley Publications in Statistics.

Shoemaker, David W. (2003). "Caring, Identification, and Agency". In: *Ethics* 114.1, pp. 88–118.

Simons, Daniel J. and Christopher F. Chabris (1999). "Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events". In: *Perception* 28.9, pp. 1059–1074.

Watson, Gary (1975). "Free Agency". In: *The Journal of Philosophy* 72.8, pp. 205–220.

Chapter 4

Context-sensitivity, framing effects and invariant attitudes

Contents

4.1 Introduction

In chapter 2, I examined a first theory of representation which defines understandings or seeings of a fact on the model of Frege's notion of sense. This theory has the merit of accommodating

cases of legitimate framing effects by rejecting the principle of weak extensionality of desires. Agents may rationally value the same consequence or option differently under different frames as long these frames correspond to different ways of seeing or understanding the consequence. However, I argued that if an option is conceived as a set of consequences or properties of the decision situation that the agent cares about, the arguments raised by Schick against weak extensionality of desires are no longer compelling. Properties with the same extension do not necessarily need to be valued equally: you may value the property of being the morning star differently from the property of the evening star, just like you may value the property of having a kidney and the property of having a heart differently. Moreover, Schick's theory is too permissive since it does not offer a clear criterion to distinguish legitimate from illegitimate framing effects. Nevertheless, Schick raises central questions for decision-theoretic representations, and in particular, whether the value of a consequence should remain invariant.

In this chapter, I discuss the hypothesis that a decision situation can be "represented extensionally" by treating its relevant features as properties of the choice situation. These properties can be associated with a particular option, a set of them, or even to the situation itself. This idea was introduced by Dietrich and List (2013) who define frames as a set of salient properties associated with the decision situation. Interestingly, instead of defining properties of a choice situation extensionally as the set of alternatives that have that property, Dietrich and List choose the intensional definition of properties as descriptions or labels attached to an option. What motivates this choice is that two distinct properties of an alternative may have the same extension. Thus properties as descriptions of an alternative are not reducible to the set of all alternatives possessing these properties. This approach is consistent with the model of conflict offered in chapter 2, where I defined consequences as intensional objects that make an evaluative difference to the decision-maker, and argued that, contra Schick, decision-theoretic consequences are valued for their sense and not their reference.

So far, framing effects have been viewed as the effects of changing frames of the same "objective" situation or feature of reality. Both Tversky and Kahneman's formulation of the invariance principle[1] and Schick's formulation of weak extensionality[2] can be understood that way. Dietrich and List's approach is objective as well, but differs in its definition of framing effects. The appeal of their framework lies in the fact that motivationally salient properties correspond to reasons and to a weighing function. Framing effects are then conceived from the experimenter's point of view as different choices across equivalent choice contexts explained by context-sensitive attitudes to reasons. A context here is defined as a particular set of feasible options, or a menu

[1] "*Invariance [is] an essential condition for a theory of choice that claims normative status is the principle of invariance: different repre- sentations of the same choice problem should yield the same prefer- ence. That is, the preference between options should be independent of their description. Two characterizations that the decision-maker, on reflection, would view as alternative descriptions of the same problem should lead to the same choice-even without the benefit of such reflection. This principle of invariance (or extensionality [Arrow 1982]), is so basic that it is tacitly assumed in the characterization of options rather than explicitly stated as a testable axiom.*" (Tversky and Kahneman 1986, p. S253)

[2] See the definition in section 2.6.

of available options[3]. This kind of approach makes sense when one attempts to empirically identify a certain kind of framing effect, and to explain what in the experimental setup objectively triggered a change in the agent's perception of the situation and associated motivation. Schick has similar intuitions when he asserts that salient elements of the agent's environment may prompt different understandings, like Orwell's change of mind at the sight of the soldier's pants.[4]

Dietrich and List formally define a framing effect as a choice reversal due to a change in the set of motivationally salient properties induced by the context of choice. Importantly, context is defined as the set of properties objectively identified by the observer and to which the subject assigns weight when the properties are motivationally salient. These properties may be features of a particular option, of a menu of options, or of the environment over and above options. Normatively, their approach of frames contrasts with Schick's in that they raise different questions in relation to extensionality and the invariance principle. While chapter 3 asked whether the value of a particular option or consequence should be constant across coreferential descriptions of the same situation, this chapter asks whether the value of an option or the weight of a consequence should be constant across different decision situations. As we will see, the view that the weight of normative reasons should be constant across situations is called invariabilism.

If representing a decision problem consists in assigning features to a particular decision situation, these features may be general, in the sense that they may be relevantly instantiated in more than one situation. For instance, options in different contexts may display the same property of being healthy. Yet whether or not these features are instantiated may itself be context-sensitive. One may find it impolite to take the last apple available when offered fruits on a tray.[5] Then, the feature of impoliteness is instantiated in certain choice contexts but not others. The evaluation of the associated options will also be sensitive to the context of choice. In this light, we can distinguish the rationality requirements bearing on the instantiation of a property from those bearing on the evaluations of these properties. One may then argue that framing effects are irrational in virtue of illegitimately context-sensitive instantiation of properties, or because the associated evaluations are illegitimately context-sensitive.

The first part of this chapter focuses on conditions of evaluative context-sensitivity. In the Asian disease example, the two formal decision problems offered to the subject respectively display the general features "saving" and "not letting die" which are supposed to be equivalent, and to play the same evaluative role in the two choice situations. Yet the agent can be blamed for irrational framing only if the following rationality conditions hold:

(i) the same consequence ought to be valued equally across contexts;

(ii) if equally valued consequences ought to play the same evaluative role across contexts.

[3]This definition of context should not be confused with the one introduced by Bermúdez (2020) when he speaks of intensional and ultra-intensional contexts (see chapter 5).

[4]Orwell 1954.

[5]The original example was given by Sen (1993).

When do (i) and (ii) apply? Dietrich and List's framework does not convincingly distinguish legitimate from illegitimate cases of context-sensitive evaluations and thus when (i) and (ii) should hold. However, by associating frames to extensional properties that are also reasons, it subjects frames to the normative requirements of reasons. Consequently, the second part of this chapter examines five views about the evaluative context-sensitivity of practical reasons addressing (i) and (ii), including invariabilism mentioned above. I argue that these five views should be rejected, and thus cannot offer a criterion of demarcation for framing effects. Nevertheless, the discussion about the way we compose reasons and their weights can be used to determine when the instantiation and the evaluation of properties are legitimately context sensitive via the notion of general and particular differences. I then defend a principle ruling out illegitimate framing effects as inconsistent choices in evaluatively equivalent decision situations. I plead for distinguishing irrationality claims made in three particular cases: in the lab via a formal description of a decision problem, in the lab via a practical decision situation, and in real-world situations.

The last section of this chapter is not intended to introduce new ideas or claims but to illustrate how the claims made in the present chapter can be made more precise formally. Sher's framework of reasons[6] offers a procedure of aggregation by composition of reasons and their weights, and can be interpreted as a theory of context-sensitive evaluations of decision-theoretic consequences. Illegitimate framing effects can then be characterized as a failure to correctly weigh together all considerations relevant to the agent, consistently with the principle I defend. More importantly, it is a non-trivial theory of rational context-sensitivity that manages to reject the five extreme views about reasons by striking a middle ground between them.

4.2 Frames as sets of properties of a choice situation

How does context affect decision-making? Dietrich and List's mainly descriptive framework attempts to answer this question by amending rational choice theory with two structural modifications. First, they claim to systematically explain choice by appealing to reasons-based preferences. Relying on an earlier paper[7] they propose that choices are made based on preferences over bundles of motivationally salient properties rather than simply preferences over primitive options. These properties, along with a fundamental preference relation, constitute reasons with particular weights which, combined, explain choice.

In addition to their reason-based theory of preferences, Dietrich and List wish to account for the fact that options are not perceived the same way by decision-makers and modelers, and that context may affect the properties that are motivationally salient to the agent. To that end, they distinguish two kinds of dependence between motivation and context: context-variance and context-relatedness. Context-variance captures the fact that an agent may care about

[6] Sher 2019.
[7] Dietrich and List 2013.

different properties in different contexts. In that case the context change will typically trigger a change in the agent's motivation and perception. For instance, one may care about the pleasure brought about by actions only when these actions don't have moral consequences. By contrast, a motivationally salient property is context-related if it is not an "intrinsic property" of the option but one which depends on features of the context. This arises in three cases: if the property depends on the relationship between the associated option and other options, as the property of "being the second largest fruit available". Alternatively, the property may exclusively depend on other options and not on the associated one, such as the property of "offering a banana among the available options". Finally, the property may depend on a feature of the context that is "over and above the feasible set of alternatives". For instance, the fact that reggae music is being played in the background during choice.

After formally describing their framework, the authors offer an axiomatic characterization of the choice functions that may be explained by these various context-dependent reasons. They then propose to explain a series of psychological choice phenomena (such as framing effects, reference-dependent choice, attraction/compromise effects) which Rational Choice Theory fails to explain. These phenomena can be sorted into two classes: rationally bounded, when the context dependence is deemed illegitimate, and rationally sophisticated when it is not. Without committing to strong normative claims, the authors suggest that bounded rationality typically involves subconscious context-variance, while conscious context-relatedness would pertain to sophisticated rationality. Before assessing these claims, let's review the specific choice behaviors they claim to explain with their framework.

4.2.1 An illustration

In their canonical example, four individuals get to choose between fruits of different sizes and kinds. Bon-vivant Bonnie always picks the largest fruit; Polite Pauline primarily avoids picking the last fruit of its kind, and only then cares about size; Chocoholic Coco is indifferent between fruits when chocolate-covered fruits are unavailable, and chooses the largest one otherwise, as "the smell of chocolate makes him hungry". As Pauline, Weak-willed William is polite in the absence of chocolate-covered fruit, yet he is greedy as Bonnie otherwise. These four individuals embody the four possible combinations of behavior produced by context-variance and context-relatedness (as illustrated in table 4.1 on p. 86). Indeed, Bonnie only cares about size, which is neither context-variant nor context-related; while Pauline, "being the last fruit of its kind available" is a motivationally salient property which is context-related, as it depends on other options available, but not context-variant, as she cares about the property in all contexts. By contrast, William is polite only when chocolate is not available, making his motivation both context-related and context-variant. Finally Coco values size only when chocolate is available: his motivation is context-variant but not context-related.

While Bonnie illustrates classical context-independent rationality, Pauline's behavior is sophisticated to the extent that her motivation is related to context but stable across contexts: though

Table 4.1. Context-variant and context-related motivations

		Context variant motivation?	
		Yes	No
Context-related motivation?	Yes	Only polite when no chocolate available (William)	Always polite (Pauline)
	No	Largest fruit when chocolate available (Coco)	Largest fruit (Bonnie)

politeness depends on other alternatives available, she always cares about being polite. By contrast, William and Coco's motivation is unstable as it varies with context.

According to Dietrich and List, the example suggests that context-variance is irrational in virtue of two characteristics: the instability of the motivationally salient property (in William and Coco's case) and its subconscious influence over choice. Before analyzing the relationship between consciousness, stability, and rational motivation, I want to examine the source of alleged irrationality in William's case more closely. William displays akratic behavior to the extent that he cares about being polite but "loses his inhibition" in the presence of chocolate. If this is the only source of irrationality in his case, it stems from the gap between properties he cares about and properties he is actually motivated by. Indeed, if he cared more about enjoying his snack than about social conventions, he would not be described as weakly-willed or irrational (see Shoemaker (2003) and chapter 3 of my dissertation on this point). If this is correct, it has an important implication for a normative account of representation: it should include properties that agents care about rather than those that are motivationally salient for the agent. By doing so, decision-makers are constrained to include what is relevant to their choice though they may later display weakly willed behavior. In fact, this requirement may also apply to a descriptive theory, if it is to capture the difference between Weakly Willed William, and his doppelganger, who cares more about fruit than about politeness in the presence of chocolate.

In the following sections, I will argue that unstable motivation and subconscious influence from context are not necessary nor sufficient conditions for irrational context dependence. To that end, I will assert that Coco's choices may in fact be rational; I will then focus on cases of illegitimate framing effects to diagnose the source of irrational context dependence.

4.2.2 Irrational context dependence does not stem from unstable motivation

Two objections can be made to the previous argument. First, one may argue that William is not only irrational in virtue of his weakness of will, but also because of the instability of his motivation. Unstable motivation or unstable reasons may be problematic when the decision-maker believes

that certain elements of the context should not impact his evaluation of the options. This can typically be the case in abstract situations, such as political or moral choice problems. However, even then, the irrationality of the choice does not stem from the agent's unstable motivation. In fact, it is generally not true that unstable motivation is irrational. Let's consider again William's doppelganger, who cares more about fruit than about politeness in the presence of chocolate. If William is irrational in virtue of his unstable motivation, so should his doppelganger: the presence of chocolate triggers a change in his motivation which makes it unstable across contexts. I think that motivational stability is relevant for explanatory and predictive purposes: it allows us to "rationalize" a greater range of behaviors with fewer explanatory variables. However, it is unclear that it plays a role in determining what counts as rational in the normative sense. To see this, let's consider Coco's case (or alternatively William's doppelganger's). From a third-person, descriptive point of view, Coco displays unstable motivational dispositions to choose across contexts. However, for a first-person normative perspective, many of our rational choices stem from context-dependent evaluations. For instance, this is the case when we treat particular circumstances as exceptions to a general evaluative rule.

If Coco's evaluations are truly affected by context, he will not regret his choice. It is hard to see then why he would be deemed irrational on the sole ground of being unstable in his motivations. As we will see, context-sensitive evaluations do play a major role in human beings' experience of the world. If this is correct, there are legitimate grounds for treating agents' evaluative context-sensitive rationality in virtue of the possible context-sensitivity to world facts. This view is supported by empirical investigation of what Tversky and Kahneman (1991) call the "psychophysics of hedonic experience". When discussing the normative status of loss aversion, a particular instance of context dependence, the authors remind us that the physical functioning of pleasure and pain should be taken into account:

> "*Questioning the values that decision-makers assign to outcomes requires a criterion for the evaluation of preferences. The actual experience of consequences provides such a criterion: the value assigned to a consequence in a decision context can be justified as a prediction of the quality of the experience of that consequence. Adopting this predictive stance, the [context-dependent] value function can be interpreted as a prediction of the psychophysics of hedonic experience. The value function appropriately reflects three basic facts: organisms habituate to steady states, the marginal response to changes is diminishing, and pain is more urgent than pleasure. The asymmetry of pain and pleasure is the ultimate justification of loss aversion in choice. Because of this asymmetry a decision-maker who seeks to maximize the experienced utility of outcomes is well advised to assign greater weight to negative than to positive consequences.*"[8]

Unstable motivation is intuitively viewed as a source of irrational context dependence because of prudential or moral considerations. William is only polite when chocolate is unavailable; however, etiquette dictates that politeness should not depend on such factors. John may choose

[8] Tversky and Kahneman 1991, p. 1057.

the moral option only when reggae music is being played; ethical common sense would not license this kind of fickleness. Both cases may be seen as instances of irrational influences of context. However, unlike externalist accounts of rationality, the Humean view I am adopting here only requires one to act on a motivation if it is subjectively endorsed. For instance, John's choice is irrational only if he cares not to let his moral conduct depend on the presence of music. Etiquette, ethical common sense, and other moral or prudential considerations, provide substantial prescriptions as to how to value acts and outcomes, and how this value may or may not depend on other factors. However, these prescriptions should be dissociated from prescriptions of rationality. Motivation may well depend on any form of context as long as it does not conflict with the agent's other attitudes.

4.2.3 Irrational context dependence does not come from the subconscious influence of context

In the previous paragraphs, I argued that the source of irrationality in William's case lies in his weakness of will rather than in his unstable motivation. I resorted to Coco's case to defend that view. One may raise a second objection, and argue that what makes both individuals irrational is the fact that the influence of context on their motivation is subconscious. Dietrich and List (2016) seem incline to make that point, as they justify the subrational character of most context-variant motivations:

> "*One might hypothesize that human beings have better conscious access to how they perceive the options in a given context K and therefore to the properties in $M(K)$ [i.e., the set of motivationally salient properties in that context] than to the context properties that affect what $M(K)$ is (i.e., those properties which, in an empirical study, might be significant explanatory variables for M). Some changes in $M(K)$ might be due to subconscious influences, as in framing or nudging effects.*"[9]

> "*Since framing effects are usually thought to be subrational or subconscious, we may take a framing effect to involve a choice reversal whose source is context-variance, not context-relatedness.*"[10]

Is awareness a condition of rationality? At first glance, it does not seem like a necessary one: Coco may well be unaware of what causes his change in motivation, as long as his choice best serves the satisfaction of what he cares about, it is hard to judge him irrational. However, being aware of what triggers your motivation may be useful in two cases. First, if the motivation is misleading in that it does not best serve the ends you care about. Suppose Coco is actually misled by the presence of chocolate, so that his desire for a bigger fruit will not grant him any further satisfaction. Then being conscious of that bias allows him to anticipate it and correct it. Consequently, being aware of illegitimate motivational triggers is instrumentally useful to

[9] Dietrich and List 2016, p. 199.
[10] Dietrich and List 2016, p. 201.

control for this first kind of bias. Second, a triggered motivation may satisfy a desire though the agent does not wish to have that desire. For instance, Coco might prefer, in the presence of chocolate, to satisfy his hunger rather than not to, but would prefer not to be hungry at all for some further reasons. Then, awareness of his motivational triggers would be instrumentally useful to control his environment and avoid developing certain desires he wishes not to have.

While awareness is useful in the previous two cases to anticipate and correct biases, the irrationality of those biases does not stem from a lack of awareness. In the first case, it comes from a misleading desire, and from undesirable temptation in the second. Being aware of either of them does not make them any more rational, it simply allows the agent to hold more sophisticated attitudes to try to best serve what he wants.

4.2.4 What makes context dependence irrational? Deliberation and akrasia conditions

Dietrich and List's model elegantly expresses the distinction between motivational and perceptual context-dependence. Their model of desirability-induced salience can be seen as the descriptive, third-person perspective which a normative theory of representation ought to complete. To make a claim of irrationality in case of a framing effect, one must first establish that the two decision problems are identical or extensionally equivalent. My view is that this first step requires taking into account the agent's evaluative attitudes as primitives, and defining irrational context-dependence with respect to evaluative concerns such as cares. Then, the descriptive and normative perspectives would coincide only if motivation and deliberative evaluations are aligned.

From the normative standpoint, desires are understood as defeasible motivational attitudes, which may stem from akratic or non-akratic behavior. To illustrate this point, consider a simple temptation case where the agent, when offered candies, has a strong impulse to accept, though he only cares about his health. In the absence of deliberation, he may well eat the candies and cave to his impulse. If he were to still desire accepting the candies after deliberation, he would exhibit weakness of will. Ex post, the source of irrationality is thus the mismatch between evaluative concerns and motivations which deliberation can prevent in non akratic cases. This distinction between impulsive behavior and weakness of will roughly corresponds to Aristotle's two kinds of akrasia (propeteia and weakness of will). Conscious or not, one of the purposes of deliberation is thus to access the evaluative standpoint against which desires are assessed.

Introducing deliberation and akrasia makes it possible to identify impulsive context-sensitivity as a non-deliberative misevaluation of options that will be regretted ex post. Applied to Coco's situation, the rationality of his context-sensitive desire for chocolate will depend on the alignment of his evaluative concerns with his motivations rather than by sudden unconscious motivational changes. I believe this is right, as the former dependence relation accurately tracks the distinction between an impulsive, non-deliberative Coco, and a continent hedonistic Coco whose desirability

is adequately context-sensitive. By contrast, William displays weakness of will after deliberation if he caves to temptation though it is not the option he came to value most. Because of weakness of will, he nevertheless chooses the option he least cares about. In comparison, William's doppelganger does care about politeness and pleasure, but contrary to William, the increased satisfaction he gets from fruits outweighs his other concerns. William is not subject to a framing effect, since he is not facing identical decision problems framed differently. However, context still affects his choice in an irrational way.[11]

An objection may be raised against the view that context variance is legitimate under certain conditions. One could argue that this claim only applies to hedonic considerations, like in Coco's case: rational or not, desires for options may generate some satisfaction, and as long as the agent sufficiently cares about pleasure or satisfaction, her choice will be rationalized. In decision problems with higher stakes, the objection goes, pleasure is a negligible factor, and what decisively matters is the value of options independently from any hedonic consideration. Furthermore, the fact that "the psychophysics of hedonic experience"[12] is sensitive to frames is not sufficient to make context variance legitimate.

In reply to that objection, I will argue that carriers of value may vary with context. Pleasure and pain are cases in point, and not negligible ones. If the first-person evaluation of pain is context-dependent, it is hard to see why this fact should not be taken into account by a normative theory[13] of rationality. What is more, this fact may apply to other values. Dietrich and List gave an example of context-related politeness: the property is only instantiated in certain contexts. However, politeness may well be context-variant too. For instance, one may find it less necessary to be polite in a situation of emergency than under more usual circumstances.

In what follows, I examine more precisely the conditions under which these context-sensitive evaluations of properties can be said legitimate. If framing a decision situation consists in assigning a set of properties to a particular option pair, the framed properties can be treated as the reasons that the decision-maker weighs while deliberating. Dietrich and List's framework allowed us to distinguish two kinds of context-sensitivity: either the instantiation of a property or its weight can depend on the set of other properties available. Framing effects were then defined as a particular case of context-invariance. I argued that context-variance was, on its own, insufficient to conclude to irrational context-sensitivity. Can we still derive conditions of irrational context variance and context relatedness? For instance, under which conditions are Coco's evaluations illegitimately context-sensitive? This brings us back to the scope of the two rationality conditions introduced earlier, i.e.:

(i) When should the same consequence be valued equally across contexts?

[11] For a systematic account of deliberation and representation of decision problem, see (Bratman 1999, chapter 2).

[12] Tversky and Kahneman 1991.

[13] For instance, phenomena of local habituation or increased tolerance for painful stimuli may rationalize certain apparently inconsistent intertemporal choices. I suspect that it may be relevant to interpret cases such as Quinn's self-torturer's case (Quinn 1990).

(ii) Should valued consequences play the same evaluative role across contexts?

Since the set of properties are reasons counting in favor of an option of a particular option pair, rational context-sensitivity of properties and their weights ought to satisfy the conditions of rationality of practical reasons. In what follows, I turn to the literature on reasons to examine how reasons and their weights ought to combine, and how they can offer criteria to distinguish legitimate from illegitimate context-sensitive evaluations.

4.3 Legitimate context dependence, additivism and invariabilism of reasons

4.3.1 Composing reasons and the structure of the space of reasons: the five views

Remember that the frame of a decision situation is here defined as the set of instantiated properties of the decision situation that correspond to the reasons in favor or against one of the alternatives. Then, the rationality standards of frames ought to be consistent with those of reasons and their weights. The literature on normative practical reasons distinguishes various structural views about the adequate way to combine and weigh reasons. In this section, I examine five of them: *strict atomism*, *strict holism*, *invariabilism*, *generalism* and *particularism*. After drawing their implications for context-sensitivity, I argue that an account of practical reasons for the bounded should reject each of these views. I then offer an alternative principle of identification of irrational framing effects.

By constraining the context-sensitivity of the weights of reasons, these views also constrain the set of rationally admissible frames across decision situations. For instance, all of these views presuppose that two option pairs that instantiate an identical set of reasons must be treated equivalently by weighing each reason equally across situations. As before, reasons are defined as properties, and a context is the set of relevant properties instantiated by an option pair.

For a given option pair, an overall reason is a reason that is composed of all the reasons in favor or against one of the options.[14] Suppose that on a particular day, you prefer taking the bus to work rather than walking there because you are tired and the bus is free on that day. Your tiredness is thus *part* of the overall reason why you prefer to take the bus today. In that sense, your overall reason for preferring the bus is in part *composed* of the fact that you are tired.

Reasons that are composed of no further reasons are called atoms. Atomic reasons are often supposed to have two attractive properties: additivism and separability. A set of reasons is additive when the weights of these reasons can be added up to obtain the weight of the overall reason. A set of reasons is separable when the weight of the conjunction of any reason is fully determined by the weight of its conjuncts. its strongest form, **Strict Atomism** is the view that reasons are always separable and their weights always additive,[15] so that the overall weight of

[14] Brown 2013, p. 780.
[15] Sher 2019, p. 144.

reasons in favor of an option is fully determined by the weight of some fundamental atomic reasons.

Strict atomism approaches reason-weighing by first assigning weights to atomic reasons, from which the weights of all other reasons can then be derived. In order to determine the weight of your overall reason to prefer the bus over walking today, atomism requires first determining the weight of the atomic reasons of which overall reason is made: presumably here, R_1 = "you are tired" and R_2 = "the bus is free". One appeal of strict atomism lies in the fact that it corresponds to an intuitive conception of deliberation where the overall weight of reasons is derived from more basic weights. Moreover, atomism makes it easy to locate the source of disagreement between two weighing measures: for a given decision situation, two people disagreeing about the weight of any particular reason must rationally disagree about the weight of some atomic reason as well.

In section 4.2, we saw that Coco picks the largest fruit when chocolate is available and the smallest otherwise. Let's suppose that the fruit size (R_3) and the presence of chocolate (R_4) are both atomic reasons. Then, strict atomism requires that their weights be separable and additive. Separability implies that the contribution of R_4 to the overall reason is independent from R_3. When it is available, chocolate is present in both alternatives; by additivity, the presence of chocolate cannot be a reason for either of them, since its contribution to each option cancel each other. Consequently, Coco's preferences are not rationalized by strict atomism when his reasons are taken as atoms.

More generally, if a frame corresponds to a set of instantiated properties, endorsing strict atomism implies that the evaluative role of a frame is reducible to the evaluative role of the atomic properties it instantiates: any evaluative difference induced by a pair of frames reduces to differences in atomic properties. This reducibility is questionable. First, properties of the overall decision situation are not always reducible to properties of the associated options, as properties of the menu of options, or even of the decision situation itself (the environment, or background) may have independent evaluative force. Then, the fact that two options have identical properties when each option is considered in isolation is not sufficient to conclude that they are equivalent, as the overall context in which they are instantiated may make an evaluative difference. These objections to atomism motivate holist views about reasons.

A reason in favor of an alternative is said to be a whole one if it is part of no other reason. Sher (2019) and Dancy (2004) define holism as the negation of strict atomism. Here, I introduce a strict version of holism to distinguish it from that weaker version. I call **Strict holism** the view that reasons are never separable and their weights not additive: all reasons reduce to the overall reason of a pair of alternatives[16]. Then, determining the weight of each reason in isolation when it is the only reason available is never sufficient to determine the weight of that reason when

[16]Though Brown does not define strict holism, his definition of isolationism "the weight that a property has as part of a larger whole is the weight that it would have in isolation" is logically implied by strict holism (Brown 2013, p. 797).

other reasons are present. For instance, the weight of R_1 can be determined by figuring out how much your tiredness makes taking the bus preferable to walking only when it is the only reason that obtains. However, when both R_1 and R_2 are present, the weight of the overall reason made of R_1 and R_2 cannot be derived from the weights of R_1 and R_2 in isolation.

Invariabilism is the view that reasons have the same weight in all situations where they are present. The invariabilist thus claims that the weight of a normative reason is never context-sensitive.[17] This view seems to account for Dietrich and List's rejection of context-variance: in Coco's case, invariabilism entails that the weight he assigns to the size of fruit should be independent from the presence of chocolate. Part of the intuition behind invariabilism is that general moral principles apply irrespectively of the choice situation.

As I argued in the section 4.2, one may value the same property to different degrees in different contexts, like politeness may matter to different degrees in different situations which invariabilism does not seem to accommodate either. We thus have to reject the invariabilist view that the weight of a reason in a given choice context is always equal to the weight it would have in isolation. Consequently, conditions (i) (the same consequence ought to be valued equally across contexts) and (ii) (equally valued consequences ought to play the same evaluative role across contexts) introduced earlier do not hold in general. More importantly, this means that framing effects cannot be deemed irrational on the sole ground that they imply context-sensitive evaluations of identical properties. In the next section, I introduce the notion of general and particular differences, and offer an alternative principle regimenting framing effects.

Strict holism and strict atomism constrain reasons in virtue of the relationship of parthood, or composition of reasons. These requirements matter for framing effects as they bear upon the sensitivity of a reason to the presence of other reasons of which it is a part, and which are part of them. The last two views that I wish to examine are views about the relationship between general and particular reasons. While **Generalism** denies that reasons may hold strength independently from particular or context-specific consideration, **Particularism** asserts that all reasons ultimately reduce to particular ones.

Ethical particularism asserts that the weight of a reason (defined as a feature of an option pair) is always a function of its particular context of evaluation, while ethical generalism rejects the idea that the weight of any reason ultimately depends on the particular context at hand. What is under scrutiny here is not the ethical debate, but the analogous one for practical rationality when reasons are defined as properties of the options: how are general and particular evaluations of decision problems related? Though I am not committed to either of these views at an ethical level, the analogous debate for rationality is intimately connected to the problem of representation for bounded decision-making.

Weighing practical reasons involves evaluating both full particular descriptions (particulars) and general properties or states of affairs: general reasons do inform our choices in a given decision

[17] Brown 2013, p. 797.

situation, but they do not always determine our evaluations of particulars. I thus reject both evaluative particularism and generalism about reasons for action, as I rejected strictly holistic or atomistic accounts of reasons. Both claims rely on similar arguments: reasons are simple, cognitively accessible properties of a pair of alternatives, which can be inductively inferred from previous choice contexts and evaluations, and affect the evaluation of future choice contexts via deliberation.

So far, I argued against these five extreme views as general standards of practical rationality. In what follows, I make a specific argument against these views once we focus on the evaluative standards of bounded practical rationality.

4.3.2 The argument of boundedness against the five extreme views

If one takes bounded rationality standards seriously, it should allow for the possibility of the four kinds of evaluations resulting from these four kinds of reasons :

(a) Retrospective evaluative experience of a particular: in some contexts, particular evaluations are more fundamental than general ones. General evaluations are then inductively inferred from particular ones. For instance, I may have enjoyed a particular muesli once, and inferred that I did because I enjoyed the mix of yogurt and honey in general.

(b) Prospective evaluation by means-ends reasoning: in some contexts, general principles or reasons determine particular evaluations. The adequacy of general evaluations to a particular one is also inferred by evaluative induction. If I want to remain fit, I may evaluate and identify my meal options along the healthy/non healthy categories, as a healthy meal may be a necessary general means to the end of being fit, leading me to decide to eat this particular healthy meal.

(c) Retrospective evaluative experience of a whole: in some contexts, the whole action (or set of relevant reasons) is evaluated independently from its parts. The evaluation of the parts is then inferred from the evaluation of the whole. In the muesli example, I identified (honey yogurt) as the whole which I value, and inferred that this evaluation is (tentatively) independent from other ingredients in the muesli. The evaluation is tentative because it may change dramatically depending on other reasons available, one thus assumes that no further individual reason will (decisively) affect the evaluation of the whole.

(d) Prospective weighing: in some contexts, reasons are additive and determine the value of the whole. Here as well, the agent must assume that the additivity of such reasons is not affected by further reasons. This ceteris paribus assumption of additivity may also be inductively grounded. For instance, I may know that having coffee and grapefruit juice in addition to my muesli will improve my breakfast in an additive way.

These evaluative inferences seem to correspond to how we proceed when we try to reach an

reflective equilibrium.[18] The notion refers to the outcome of a deliberative process that examines "*our moral judgments about a particular issue by looking for their coherence with our beliefs about similar cases and our beliefs about a broader range of moral and factual issues*."[19] Though it has been mainly used in ethics and inductive logic, it also captures the everyday practice we have when trying to determine what we ought to do. For our present purpose, it illustrates well the complexity of coherently articulating general and particular evaluative judgments, and gives additional grounds to make the evaluative attitudes (a) to (d) permissible.

An omniscient agent would dispense with these evaluative inferences, as she would know the decisive contribution of all general and particular reasons to the weight of the overall reason for an alternative. Inferences (a) and (b) are needed when one does know the exhaustive set of relevant reasons of a particular decision situation, or cannot determine exhaustively the relations between the general and particular reasons of the situation. Inferences (c) and (d) are only useful to agents that do not know whether a reason constitutes an atom or a whole: reasons being defeasible, bounded agents have to make assumptions about the decisive wholes and atoms of the decision situation. When reasons are non additive, the overall weight of the complete set of reasons in a given context cannot only be determined by the weights of reasons in isolation. Then, the agent's particular evaluation of the whole option may play a fundamental role in determining the weights of individual (yet dependent) reasons.

However, none of the extreme views about reasons introduced earlier (generalism, particularism, strict holism and strict atomism, invariabilism) can account for these four kinds of evaluations. Consequently, bounded rationality ought to reject the extreme views.

Strict generalism and particularism are over-constraining views of rational reasons. Nevertheless, the notions of general and particular differences can be fruitfully used to characterize decision-theoretic consequences. In chapter 3, I argued that consequences are features of the choice situation making a differential impact the agent cares about. When decision theorists discuss abstract decision situations, they typically characterize options by sets of general properties in virtue of which the options differ. Yet actual decision-makers may be sensitive to particular differences between options that cannot be expressed by some general property. In the next section, I examine the notions of general and particular differences and defend a principle identifying evaluatively equivalent decision situations that distinguishes legitimate from illegitimate framing effects.

4.3.3 A principle of identification of evaluatively equivalent decision situations

By defining framing effects as context-variant evaluations of reasons, Dietrich and List (2016) presumably meant to refer to reasons as general properties which can be identified by an outside observer. In the lab, the formal decision problems submitted to subjects are general descriptions.

[18] The notion was elborated by Goodman (1955) to justify rules of logical inference and by Rawls (1951, 1971) to justify principles of justice as fairness.

[19] Daniels 2023, Introduction.

In that sense, any reason holding motivational force must be a general one. A principle ruling out irrational framing effects outside of the lab should take into account the fact that a reason and its weight may be specific to a particular choice situation. One should thus distinguish general and particular difference-making features of a decision situation. While features making a general difference between two options can be present in two distinct decision situations, features making particular differences cannot. Consequently, two pair-wise decision situations can be treated equivalently only if neither of the four options makes a particular difference the agent cares about, and if the set of general differences instantiated by an option pair is identical to the set of general differences instantiated by the other option pair.

This condition guarantees that each option makes identical differences in its own decision situation. It is yet insufficient to require treating two decision situations equivalently: we have seen that the evaluation of options may depend on the menu of options as well as on features over and above the options themselves. Consequently, we also need to suppose the absence of relevant differences across decision situations. Combined, these conditions offer the following principle ruling out illegitimate framing effects in context-equivalent decision situations.

Principle of identification of irrational framing effects in context equivalent decision situations

Consider two particular decision situations each associated with a pairwise choice. In each case, the decision-maker may identify relevant differences between the options that are either particular to their context or that are general in that they hold across several choice contexts.

Further, suppose that

(a) each of the decision situations can be fully characterized by general differences between options, and these differences are the same in both situations, and

(b) the two decision situations do not differ in any particular way the agent cares about, so that these general differences equally apply to both situations.

Then, the two choice situations ought to be treated equivalently, and only frames yielding equivalent choices across situations will be rationally admissible.

Let's see how this principle applies to a simple case of illegitimate framing effect. Suppose you are regularly doing grocery shopping at the same place and you consider buying yogurts. The store usually offers three kinds of yogurts A, B and C. Yogurt C is always available; Yogurt A and B are sometimes out of stock. You believe A and B do not differ in any aspect you care about; the only difference you notice is the percentage of fat displayed: Yogurt A has a tag indicating 20% fat while the tag on yogurt B indicates 80% non-fat. You do believe that these percentages amount to the same quantity of fat. However, when A is available you prefer to buy C over A, while you favor B over C when B is available. If the only difference you care about between the

two choice situations is the amount of fat in the yogurts, the principle prescribes to treat them equivalently, making your preferences irrational.

This first example illustrates the principle in a simple framing effect case; let's now turn to a more elaborate case of evaluatively equivalent situations. As you intend to buy a car, a first car dealer offers you a choice between two cars A and B. The one you overall favor is bigger and cheaper. A second car dealer offers you a choice between cars C and D. C is bigger than D but more expensive. However, since you only care about the size of the car and not its price, the choice between A and B and C and D are context-equivalent: within each choice situation, one car is bigger than the other, and you care about no other difference across the two cars. Moreover, the car dealer's offers do not differ in any relevant way either. Then, failing to prefer C to D when you prefer A to B is a case of illegitimate context-sensitivity.

Suppose now that C is bigger than D, just as A is bigger than B. However, the second car dealer includes a free parking subscription for any purchase in his store. Because of the parking difficulties in your city, you never consider buying bigger cars; yet with the included subscription this is no longer a concern and you go for the bigger car. Though in both situations, the only difference between cars is size, the difference across choice situations is sufficient to make them evaluatively non-equivalent.

This principle does not involve any requirement bearing on the semantic relation between frames and some underlying objective reality that these frames would represent. Unlike the extensionality and invariance principles we have seen so far, it does not define framing effects as multiple framing of the same decision situation or of objectively identical decision situations. I see this as a strength of this normative account. Instead, I defend a view of framing effects as a particular kind of illegitimate context-dependence, where *evaluative equivalent decision situations* are not treated equivalently. Evaluative equivalence is defined here as the absence of relevant differences across situations.

General differences and their weight exist in virtue of the fact that the agent cares about them. If she cares about no particular differences between option pairs, the principle requires her to treat these choice contexts equivalently. Then, an extensionality principle is not needed. Any pair of choices inconsistent with her evaluations of differences will be ruled out by the principle I defended above. The irrationality claim does not bear upon the identity of options, but on the identity of cares for differences across options within and across decision situations. Identical cares for general differences imply equal evaluations of the associated differences. If there is no particular difference the agent cares about, the options ought to be valued equally.

4.3.4 Issues of empirical identification of framing effects in the lab

The principle introduced above is formulated from the agent's perspective, and derives from structural consistency constraints between beliefs, evaluations of general and particular differences (across and within choice situations). In the lab, the identification of illegitimate frame

sensitivity raises additional issues. The empirical scientist's usual solution is to hypothesize that the descriptions are equivalent for all relevant matters, and that this equivalence is endorsed by the decision-maker. In the Asian disease case, she must suppose that the difference between "not saving" and "letting die" is negligible for the purpose of identification. Consequently, what is part of the context, and what is not cannot rigorously be determined by the modeler without further hypotheses about the agent's evaluative dispositions underlying her individuation of the situation.

A second issue with the empirical identification of equivalent descriptions in the lab comes from the fact that they are general descriptions of a particular situation that do not exist. The fictional nature of the experiment is not in question here; however, descriptive equivalence between propositions makes sense when the descriptions refer to the *same particular* target. For instance, the descriptive equivalence between "not saving" and "letting die" may be determined both by their equivalence as general descriptions and as particular ones. More generally, descriptive equivalence depends on the set of general and particular differences that matter to the subject.

In cases where the subject is offered formal descriptions of decision problems, since no particular difference can be inferred between pairs of formal decision problems, the only relevant differences the observer can establish will be general. Then, assumptions must be made about the evaluative equivalence of properties (e.g., between saving and letting die). Misevaluation may come from inconsistent beliefs or evaluations (epistemic and evaluative framing effects, respectively).

When the observer does not resort to formal descriptions to ask the subject to make decisions, further hypotheses must be made to guarantee that the conditions of the principle apply. She must additionally hypothesize the set of relevant general properties characterizing the agent's frame of each choice situation, and that no particular differences lie across and within choice pairs.

Let's see how the principle applies to a formalized decision problem in the lab, in the case of. wealth-dependent evaluations of monetary gains.[20] John values differently a gain of $1000 when he has a million dollars and when he doesn't. The value of an additional $1000 varies with John's level of wealth. Is John subject to a framing effect? If so, is it irrational?

The fact that a good is context-sensitive does not necessarily means that it is represented differently in different contexts. John's sensitivity to wealth-levels does not necessarily imply that his *understanding* of the $1000 is different in the sense intended by Schick (1991). It may simply be the case that John has extensional but richer attitudes to money than what the functional form suggests. In John's case, monetary gains are not separable from background wealth. Formally, the value functions representing an agent's evaluation or experienced satisfaction level as a function of quantities of a particular good may have more than one argument: gains and losses, wealth level, etc. Moreover, the functional form itself may be complex enough as in the case of non-linear utility of money.

[20] I thank José Luis Bermúdez for underlining the relevance of endowment-effects for issues of framing sensitivity.

The fact that these evaluations cannot be expressed in a simple functional form is not an argument on its own to charge them with inconsistent evaluations. In fact, our theoretical understanding of how we should value financial prospects may be improved by criticizing the prescriptions endorsed by too constraining functional forms of utility. Although von Neumann and Morgenstern's theory[21] presupposed a linear utility of money, they contributed to clarifying the concept of decreasing marginal utility of money. Similarly, Buchak's decision theory[22] disconnects the concepts of risk aversion and of decreasing marginal utility of a good by opposing orthodox views that endorse simpler forms of utility functions.

In the previous example, John sees an evaluative difference between the same financial gain in different contexts of wealth. This difference is enough for him to treat the two choice situations non-equivalently, and reject the hypothesis of a framing effect. That being said, provided that he only cares about gains and absolute levels of wealth, the logical relation between these two features does constrain how John's preferences can coherently be context-sensitive. For instance, for a constant level of background wealth, John ought to prefer greater gains to smaller ones. Conversely, in choices that involve the same gain but different wealth level, he should prefer alternatives with greater background wealth.

This example illustrates how the principle applies in the case of a formalized decision problem in the lab. The requirement that equivalent decision situations ought to be treated equivalently can be conceived as a particular instance of a more general requirement of coherence between on the one hand, the agent's evaluative attitudes, in virtue of which the decision situations are framed as a particular set of relevant features, and on the other, the choices resulting from the evaluations of these relevant features as reasons that can be weighed.

4.3.5 Conclusion

This chapter explored the idea that a decision situation can be "represented extensionally" by treating its relevant features as properties of the choice situation. Then, these relevant features also constitute the set of weighable reasons of the decision situation. That hypothesis allowed us to distinguish between properties associated with a particular option, with a set of them, those attached to the situation itself. The context of the situation can then be meaningfully defined as the set of relevant properties of the decision situation, and two kinds of context-sensitivity should be distinguished. First, whether a property is instantiated or not can depend on the context of the decision situation. Second, the evaluation of those properties can itself be context-sensitive (what Dietrich and List (2016) respectively called context-relatedness and context-variance). In this light, the rationality requirements bearing on the instantiation and the evaluation of options as bundles of represented properties offer grounds for excluding certain ways of representing options. Applied to framing effects, the evaluative requirements raised the questions of (i) when the same consequence should be valued equally across contexts, and (ii) when equally valued

[21] Neuman and Morgenstern 1944.
[22] Buchak 2013.

consequences should play the same evaluative role across contexts.

This chapter answered these two questions by arguing that since the set of properties are reasons counting in favor of an option of a particular option pair, rational context-sensitivity of properties and their weights ought to satisfy the conditions of rationality of practical reasons. I thus turned to the literature on reasons to examine how reasons and their weights ought to combine. I introduced five extreme views about reasons (invariabilism, atomism, holism, generalism and particularism) and rejected them both on general grounds and in the case of bounded agency. These views thus cannot offer a criteria of demarcation for framing effects. Nevertheless, the discussion about the way we compose reasons and their weights can be used to determine when the instantiation and the evaluation of properties are legitimately context-sensitive, via the notion of general and particular differences. I defended a principle ruling out illegitimate framing effects as inconsistent choices in evaluatively equivalent decision situations. I defined equivalent decision situations in terms of general and particular differences the agent cares about, consistently with the definition of decision-theoretic consequences offered in chapter 3. In that light, I pleaded for distinguishing irrationality claims made in three cases: in the lab via a formal description of a decision problem, in the lab via a practical decision situation, and in real-world situations.

4.4 A marginalist illustration of the present account of rational context-sensitivity

The last section of this chapter is not intended to introduce new ideas or claims but to illustrate how the claims above evaluative context-sensitivity (or context-variance) made in the present chapter can be made more precise formally. Sher's framework of reasons[23] offers a procedure of aggregation by composition of reasons and their weights, and can be interpreted as a theory of context-sensitive evaluations of decision-theoretic consequences. Such a theory would endorse a marginalist view of evaluative context-sensitivity that is not central to my account; yet it has the merit of being a non-trivial theory that rejects the five extreme views about reasons by striking a middle ground between them.

4.4.1 The model

Sher offers a formal axiomatic model of reasons which draws from probability theory, attempting to make sense of how weights behave quantitatively when they are combined. To that end, he relates the weights of reasons to the expected differential impact they have on preferences. For a given option pair, the evaluative impact of a reason can be quantified as the difference it makes to the relative value (or betterness) of one alternative over the other. The value of an alternative is quantified by a value function V.

[23] Sher 2019.

Sher takes reasons to be propositions, and considers their evaluative impact from a position of **deliberative uncertainty**, where the agent does not know whether these propositions are true, to a state of knowledge where she knows that a particular set of reasons obtains. Reasons can then either be interpreted as evidence for the fact that a particular alternative is preferable to another, or as justifications of these preferences. As you intend to buy a car, you investigate what features make a car more desirable than another in your eyes. If getting to know that a car is electric makes it preferable to the other car, the fact that a car is electric is a reason to choose it over a non-electric one.

As with traditional uncertainty, the expected value $\mathbb{E}[V_o]$ of option o can be defined under deliberative uncertainty, as well as the expected value of o conditional on knowing that reason R obtains, $\mathbb{E}[V_o \mid R]$. Crucially, the **weight of a reason** is defined as the quantitative effect that coming to know that this proposition is true would have on the agent's preference between the two options. For a given option pair, the weight of a reason $w_{a,b}(R)$ can then be expressed as a double difference of expected values:

Weight of reason R relative to alternatives (a, b).

$$w_{a,b}(R) = (\mathbb{E}[V_a \mid R] - \mathbb{E}[V_b \mid R]) - (\mathbb{E}[V_a] - \mathbb{E}[V_b])$$

Here $\mathbb{E}[V_i \mid R]$ is the expected value of option i, were I come to learn that reason R obtains; $w_{a,b}$ thus captures the change of the relative desirability of a and b between two states of deliberative uncertainty: one where I don't know if reason R obtains and one where I do. Consequently, the weight of a reason is relative to (i) a pair of options, and (ii) the values of these options in a position of deliberative uncertainty.

Importantly, the conjunction of all relevant reasons counts as a reason in favor of a over b if and only if a is comparatively better than b conditional on knowing that all relevant reasons obtain. Consequently, all weighing functions $w_{a,b}$ are such that the conjunction of all relevant reasons $\bigcap_{i=1}^{n} R_i$ in favor of a over b has a positive weight just in case the expected value of option a conditional on $\bigcap_{i=1}^{n} R_i$ is greater than the expected value of b conditional on $\bigcap_{i=1}^{n} R_i$:

$$w_{a,b}\left(\bigcap_{i=1}^{n} R_i\right) > 0 \Leftrightarrow \mathbb{E}\left[V_a \,\middle|\, \bigcap_{i=1}^{n} R_i\right] > \mathbb{E}\left[V_b \,\middle|\, \bigcap_{i=1}^{n} R_i\right].$$

This equivalence between the weights of all relevant reasons combined, and the comparison of expected values allows Sher to define a complete transitive ordering of options based on the weights of their relevant reasons. For all pairs of actions (a, b), a weak order $<$ can be defined over options such that:

$$a < b \Leftrightarrow w_{a,b}\left(\bigcap_{i=1}^{n} R_i\right) > 0.$$

This implies that the agent prefers the action which, under deliberative uncertainty, has greatest expected value conditional on relevant reasons, or equivalently, the action most supported by the weight of the agent's reasons.

One of the great attractions of Sher's framework lies in the fact that the machinery of probability theory allows combining non independent propositions through conditioning. The conditional weight of R_1 given that R_2 obtains is defined as follows:

$$w_{a,b}(R_1 \mid R_2) = (\mathbb{E}[V_a \mid R_1 \cap R_2] - E[V_b \mid R_1 \cap R_2]) - (\mathbb{E}[V_a \mid R_2] - E[V_b \mid R_2])\,.$$

Conditional weights capture the differential impact on value, of a reason (here R_1) conditional on other reasons (here R_2). While the weight of "independent" reasons can be obtained by summation, a simple formula connects the weights of dependent reasons. In the case of two dependent reasons: $w_{a,b}(R_1 \cap R_2) = w_{a,b}(R_1) + w_{a,b}(R_2 \mid R_1)$. Moreover, option a is preferred to b if and only if:

$$\mathbb{E}[V_a \mid R_1 \cap R_2] > \mathbb{E}[V_b \mid R_1 \cap R_2],$$

or, equivalently,

$$w_{a,b}(R_1) + w_{a,b}(R_2 \mid R_1) > 0.$$

So far, we have seen that the weight of a reason is defined as the effect of that reason on the expected value of options. How are weights determined? Sher's model accommodates two competing philosophical views here: either reasons have weights in virtue of the fundamental values they bear upon; or conversely, the value of options can be determined by the fundamental weights of the underlying reasons. When we resort to intuitions about the significance of a reason, we take the property of being valuable as providing a reason (and thus a weight) to take a certain action. But we may take our intuitions about the weights of reasons as given, and derive the implicit values associated with them. The model easily allows it by reversing the formal relationships between these objects.

Although Sher's work is mainly concerned with normative ethical reasons, I believe the distinction previously made about the possible directions of the relationship between values and weights also holds for practical reasons. In fact, I think it may be an illuminating way of understanding the distinction between Humean and substantive views of rationality. In broad terms, the Humean view can be interpreted as taking desirability (or other conative attitudes) as the rational standard of value, and as fundamental with respect to reasons and their weights. Reasons thus have weights to the extent that they have a differential impact on the desirability of the options. Conversely, substantive views of rationality will argue that some reasons have weights independently from their effects on desirability. Then, the standard of value no longer is desirability but something like rational goodness or well being.

4.4.2 Implications for evaluative context dependence

First a technical remark. Sher chose to model reasons as propositions and not as properties of the alternatives, like Dietrich and List's did. This definitional difference may raise the worry that the insight raised by one framework may not apply to the other. I don't think this worry is grounded. As Dietrich and List (2013) show,[24] any property of an alternative can be defined as the proposition that the alternative has this property; conversely, any proposition can be associated with the property of satisfying that proposition, without any loss of generality. For instance, suppose I am deliberating about which car to buy and that the fact that a particular car is blue is a reason for me to buy it. The "blueness" of the car is a property of the alternative, and it corresponds to the proposition that "this particular car [or alternative] is blue". As we saw in chapter 2, using properties rather than propositions allows us to clarify the issues of extensionality raised by propositions, since properties can be coextensional while distinct.

Let's now turn to the implications of Sher's model for the five views introduced earlier. Let's notice first that Sher's conception of reasons and weights rejects Strict atomism, Strict holism, and Invariabilism. When reasons are independent, their weights are separable and additive, thus violating Strict Holism. Yet non-independent reasons have non-additive and non-separable weights, contra Strict Atomism.

Moreover, for a given option pair, the weight of a reason R_1 is in general not independent from the set of relevant reasons (R_i) of that pair: the conditional weight $w_{a,b}(R_1 \mid \bigcap_{i=2}^{n} R_i)$ is different from $w(R_1)$ when these reasons are not independent. If we define, as in subsection 4.4.1, the context of a decision situation as the set of relevant reasons of the option pair, the weight of a reason is not the same in all contexts where it is present. Sher's account thus does not support invariabilism. Instead, it offers a marginalist rule of composition of reasons specifying how weights combine to add up to an overall weight. According to **Marginalism**, the weight of a reason R for an alternative in a given context is the difference between (i) the weight of the overall reason for that alternative, and (ii) the weight of the overall reason for the alternative once R has been removed from the context. Marginalism thus defines the weight of a reason in its context as the marginal effect of adding it to that context.

The overall weight of two reasons $w(R_1 \cap R_2)$ can be decomposed as (i) the weight that reason R_1 would have in isolation $w(R_1)$, and (ii) the weight of R_2 conditional on R_1:

$$w(R_1 \cap R_2) = w(R_1) + w(R_2 \mid R_1).$$

[24] "*This definition is standard when X is a set of possible worlds or states of the world. If X consists of other objects (e.g., bundles of goods), subsets of X are conventionally called properties. It is a mere terminological convention that we speak of propositions rather than properties in this paper, and our analysis could equally be formulated in terms of properties. Setting aside subtleties, we can associate any property in the conventional sense with the proposition, as understood here, that a given object has that property (where the proposition is true of the objects that have the property, and false of all others). Conversely, we can associate any proposition as defined here with the conventionally understood property of satisfying the proposition. Our present terminology has some independent advantages, including (but not restricted to) the representability of propositions by sentences.*" (Dietrich and List 2013, p. 132)

The second term $w(R_2 \mid R_1)$ can be interpreted as the marginal contribution of R_2 given that R_1 obtains. This marginalist law of composition may be read in two different ways: either unconditional weights can determine conditional ones, or the weight of all relevant reasons can be determined by a sum of conditional weights. Since weights are defined by their differential impact on the value of alternatives, one may either obtain the overall weight of reasons by determining the weight of its components; or some of its conditional components may be derived by measuring their marginal effects on overall weights. The appeal of marginalism lies in the fact that, while rejecting strict atomism, it accommodates the idea of a weighing procedure to obtain overall weights.

Sher's decision theory is not exposed to trivialization since it excludes certain sets of preferences. When two propositions overlap (when one proposition implies the other for instance), the associated reasons are not independent, and adding their weight will create an issue of "double counting". To see this, suppose I am considering retirement plans with various properties. P_1 is the property of offering a full pension for at least 20 years, P_2 is the property of offering a full pension for at least 30 years. In such a case, the weight I will assign to the conjunction of P_1 and P_2 will not be the same as the sum of their weights in isolation. In such cases, Sher's model offers a way of consistently weighing reasons without falling prey to double counting.

At the beginning of this chapter, I argued that Coco's preference for big fruits only in the presence of chocolate was permissible. This is easily accommodated by Sher's framework. It may even be the case that the presence of chocolate carries no weight in isolation, although it leads to a reversal of preference between fruits. Formally:

$$w(\text{small \& chocolate}) = w(\text{small}) + w(\text{chocolate} \mid \text{small}) = w(\text{chocolate}) + w(\text{small} \mid \text{chocolate})$$
$$w(\text{big \& chocolate}) = w(\text{big}) + w(\text{chocolate} \mid \text{big}) = w(\text{chocolate}) + w(\text{big} \mid \text{chocolate}).$$

Supposing the weight of chocolate in isolation to be null: $w(\text{chocolate}) = 0$, then

$$w(\text{small \& chocolate}) < w(\text{big \& chocolate}) \quad \text{IFF} \quad w(\text{small} \mid \text{chocolate}) < w(\text{big} \mid \text{chocolate}).$$

I defended the view that unstable motivation alone is not a source of irrational context-sensitivity. I also insisted that rational context dependence is often mistaken for the effects of substantial prescriptions as to how the value of a property may or may not vary with other factors: for instance, morality may prescribe that the moral option be chosen whether or not reggae music is being played in the background. These substantial prescriptions, typically moral or prudential, should not be confused with rational ones.

Instead, I offered a principle defining evaluatively equivalent decision situations based on general and particular differences the agent cares about, which constitute the relevant reasons of each decision situation. This principle thus specifies when two decision situations can be characterized by the same sets of reasons. Since the principle involves conditions of the legitimate

identification Sher's model is compatible with the principle I defend. Though it presupposes that weights are codependent, they are uniquely defined by their functional form. Consequently, situations with identical sets of reasons should be equally valued. Then, agents are rationally required to value a reason identically across contexts they deem equivalent with respect to other reasons, short of which they may be blamed for inconsistent evaluation or for akratically responding to these evaluations. Since Sher's framework takes as given the set of relevant reasons of a situation, it can be coherently associated with the principle. Combined, they specify how the way we represent decision situations ought to cohere with our evaluative attitudes, across equivalent decision situations (through the principle), as well as across non-equivalent ones (through coherent evaluations, as in cases of double counting).

Finally, Sher's model offers insight as to how prudential prescriptions may play a role in ruling out certain kinds of context dependence. If the agent endorses such a prescription, he will accept that the weight of a reason ("stealing is wrong", or S) should be independent of another putative reason ("reggae is being played in the background", or R), so that $w(S \mid R) = w(S)$. In the Asian disease case, a prescription possibly endorsed by the agent would be that the difference between saving and not letting die should not affect the moral value of options. Again, this kind of prescription is not rationally binding unless the agent himself endorses it.

Now that we have defended a principled distinction between rational and irrational framing effects that is non-trivial, let's see how it compares with another principled theory of frames, Bermúdez (2020).

References

Bermúdez, José Luis (2020). *Frame It Again: New Tools for Rational Decision-Making*. Cambridge University Press.

Bratman, Michael (1999). *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge University Press.

Brown, Campbell (2013). "The Composition of Reasons". In: *Synthese* 191.5, pp. 779–800.

Buchak, Lara (Nov. 2013). *Risk and Rationality*. Oxford University Press.

Dancy, Jonathan (2004). *Ethics Without Principles*. New York: Oxford University Press.

Daniels, Norman (2023). "Reflective Equilibrium". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2023. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/fall2023/entries/reflective-equilibrium/.

Dietrich, Franz and Christian List (2013). "A Reason-Based Theory of Rational Choice". In: *Noûs* 47.1, pp. 104–134.

Dietrich, Franz and Christian List (2016). "Reason-Based Choice and Context-Dependence: An Explanatory Framework". In: *Economics & Philosophy* 32.2, pp. 175–229.

Goodman, Nelson (1955). *Fact, Fiction, and Forecast*. Harvard University Press.

Neuman, John von and Oskar Morgenstern (1944). *Theory of Games and Economic Behavior*. Princeton University Press.

Orwell, George (1954). "Looking Back on the Spanish War". In: *A Collection of Essays*. Written in 1942. Garden City: Doubleday.

Quinn, Warren S. (1990). "The Puzzle of the Self-Torturer". In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 59.1, pp. 79–90.

Rawls, John (1951). "Outline of a Decision Procedure for Ethics". In: *The Philosophical Review* 60.2, pp. 177–197.

Rawls, John (1971). *A Theory of Justice*. Harvard University Press.

Schick, Frederic (1991). *Understanding Action: An Essay on Reasons*. Cambridge University Press.

Sen, Amartya (1993). "Internal Consistency of Choice". In: *Econometrica* 61, pp. 495–521.

Sher, Itai (2019). "Comparative Value and the Weight of Reasons". In: *Economics & Philosophy* 35.1, pp. 103–158.

Shoemaker, David W. (2003). "Caring, Identification, and Agency". In: *Ethics* 114.1, pp. 88–118.

Tversky, Amos and Daniel Kahneman (1986). "Rational Choice and the Framing of Decisions". In: *The Journal of Business* 59.4, S251–S278. (Visited on 01/11/2024).

Tversky, Amos and Daniel Kahneman (Nov. 1991). "Loss Aversion in Riskless Choice: A Reference-Dependent Model". In: *The Quarterly Journal of Economics* 106.4, pp. 1039–1061.

Chapter 5

Frame-sensitive decision-making

Bermúdez's account of decision theoretic representations

Contents

5.1 Introduction

Let's take stock on what we have done so far. Taking framing effects and the invariance principle as the paradigmatic example of controversial decision-theoretic representations of a decision situation, I turned to Schick's theory of understanding and argued that although I disagree with his normative account of understanding and framing effects, *it raised a series of representational*

challenges for decision theory. Among them, the question of how decision problems arise, of how they may constitute partial and selective representations of a decision situation, and of the normative status of framing effect and extensionality.

In response to the first two questions, I advocated in chapter 3 incorporating cares and instrumental plans to the standard belief-desire model of agency upon which decision theory traditionally relies. Indeed, Bratmanian plans account for a number of our intuitions on the way bounded agents isolate decision problems and their components. Their partial hierarchical structure largely accounts for the structure of decision problems. Combined with cares, they provide a legitimate conceptual apparatus to think of what I called the problems of scope and grain of decision problems.

In chapter 4, we saw that frames can be "represented extensionally" by treating the relevant features of the decision situation as properties of the choice situation that correspond to reasons in favor of a particular option over another. Then, the problem raised by framing effects was no longer whether coreportive descriptions of a consequence or option should be valued equally. Instead, it asked whether the value of a reason should be constant across different decision situations, and if equally valued consequences ought to play the same evaluative role across contexts. In lieu of an invariance principle, the contentious view associated to framing effects was invariabilism, the view that the weight of normative reasons should be constant across decision situations. Along with other extreme views about practical reasons, I rejected invariabilism on the ground that, just as goods are not always separable, normative reasons are not always independent.

As an alternative, I defended a principle that regiments illegitimate framing effects by specifying conditions of equivalence between decision situations. The principle required that (a) each of the decision situations can be fully characterized by general differences between options, and these differences are the same in both situations, and (b) the two decision situations do not differ in any particular way the agent cares about, so that these general differences equally apply to both situations. Then, the two choice situations ought to be treated equivalently, and only frames yielding equivalent choices across them will be rationally admissible.

It thus appears that our normative intuitions about framing effects do not require an ultra-intensional conception of representation. However, this does not prove that such a conception is not necessary to account for our other normative intuitions about practical rationality. Are there any instances of rational decision-making that can only be accommodated by the ultra-intensional conception of representation?

In what follows, I examine Bermúdez's account of representations as frames (Bermúdez 2020) and of framing effects. Like Schick, Bermúdez rejects the invariance principle and claims that standard decision theory is insufficient to capture our intuitions about complex decision situations involving inconsistent evaluations of the same outcome or option under different frames. The general intuition that guides Bermúdez's demonstration is that the rigid consistency

standards of traditional decision theory are inadequate when real-world situations demand a richer or deeper understanding.

After presenting his account of frames, I examine how it differs from the account of representation defended by this dissertation. Among the key differences, Bermúdez imposes externalist constraints on beliefs when the present Humean account does not.

I then examine Bermúdez's claims about rational representations and look into two issues. I first argue that his canonical example (Agamemnon) is not sufficient to reject the invariance principle nor to justify the introduction of ultra-intensional frames, consistently with chapter 2. Irrational framing effects can be defined independently from factual neutrality as inconsistent choices across evaluatively equivalent decision situations, as defended in chapter 4. Second, I contend that the specification of factual propositions in a decision problem is not evaluatively neutral, which in turn violates the Due Diligence requirement. If this is correct, this pleads in favor of accounts that acknowledge that all specifications of decision problems and their consequences are value-driven.

5.2 Bermúdez's account of framing: non extensional frames and the rational value of frame sensitive-reasoning

5.2.1 Framing and extensionality

What are frames? Without initially committing to a specific definition, Bermúdez reviews the various uses of the terms in recent academic thought. In psychology and economics, frames often refer to the various modes of presentation and connotations of an outcome. In cognitive linguistics (Lakoff 2004), framings are "*cognitive metaphors that structure how we think and speak about thorny moral and political issues.*"[1] Frames are not "synonymous descriptions" as they do not have identical meaning: they belong to a special category of homonymic descriptions, those that share a common core, its focal meaning. Goffman (1974) defines frames as "*schemata of interpretations*" that allow us to "perceive, identify and label".[2] With this broad definition in mind, Bermúdez investigates the various ways frames relate rationality.

After examining the major arguments raised against frames as fundamentally illogical, Bermúdez shows that framing can help address problems left unsolved by normative theories, particularly regarding the rationality of self control and cooperation. It offers conceptual tools to describe, assess and guide decision-making. For the purpose of this dissertation, I will mainly focus on the author's views regarding the normative status of frame sensitivity.

The starting point of Bermúdez's inquiry is the inappropriate normative treatment the notion of frame has received in decision-theoretic disciplines, which he refers to as the litany of irrationality. This litany of irrationality stems from the "dominant way of thinking about rationality and

[1] Bermúdez 2020, p. 11.
[2] Goffman 1974, p. 21.

rational decision-making in social sciences" that leaves no room to rational framing effects. Indeed, even in descriptive, experimental disciplines, these phenomena are explored, analyzed, and interpreted through the lens of standard normative theories of rationality that all share the common principle of extensionality.

To situate where extensionality comes into play, Bermúdez offers a very clear picture of the distinct steps involved in decision-making. Decision-making consists in deciding what to do, by selecting an action among a set of possible options, based on how each act is valued. To that end, decision-makers "calculate the possible outcomes of each action", and then assess and compare them in accordance with their goals and values. However, Bermúdez rightly insists that "all these calculations do not operate in a vacuum", but apply to representations of concrete decision problems. He thus calls for some schemes to identify available acts and outcomes, and principles to understand how acts and outcomes are affected by factors outside of the agent's control: "*Decision-makers need to situate what they can do relative to constraints imposed by the environment in which they are operating.*"[3] Consequently, he distinguishes four phases of decision-making (see figure 5.1 on p. 110). The first step of the process consists in *representing* a decision problem via possible actions, constraints, and relevant values and goals, after which possible outcomes and their likelihood are *projected*; thirdly, options are valued and compared, via cost benefit analysis under ecological constraints, before one is selected based on the maximization principle.

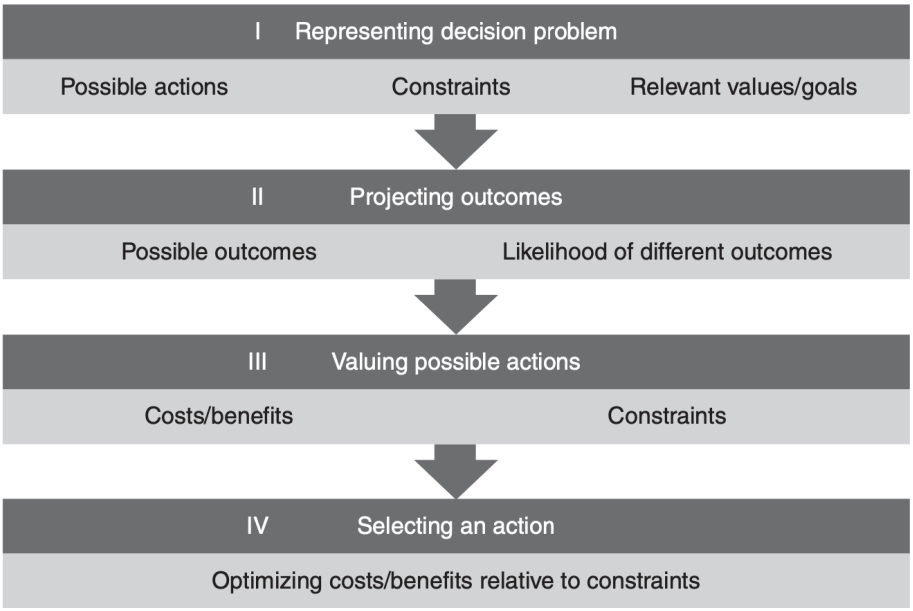

Figure 5.1. Bermúdez's four stages in decision-making[4]

Though decision theory applies to the third and fourth stages only, the litany of irrationality adds an auxiliary requirement to the theory to account for irrational framing effects, which

[3] Bermúdez 2020, p. 69.
[4] Bermúdez 2020, figure 4.1 p. 69.

Bermúdez names Juliet's Principle.

5.2.2 Juliet's Principle

Juliet's principle requires that the value of an outcome (defined Savage-like consequences or events) shall be constant across frames, and more generally that "what we call things should not affect how we value them."[5] This pre-theoretic tenet captures the common intuition shared by all theories of decision-making that frames are evaluatively neutral. In its initial formulation by Arrow (1982), the principle states that when choosing from a choice set, rational choices should be independent from how the set is described.[6] Arrow's motivation was quite understandable: concerned with measuring purchasing power for economics purposes, he defended the hypothesis that the rational homo economicus only cares about maximizing her purchasing power over a bundle of available goods, conditional on her limited income. More generally, if decision theory is formulated in set-theoretic terms (as Savage's), it presupposes a version of the extensionality axiom: the identity of a set is determined by and exclusively by its members, so that sets are description invariant.

Bermúdez is understandingly suspicious of these background auxiliary requirements of evaluative neutrality, and particularly so in the case of frames. The disciples of the litany of rationality overlook one fundamental representational fact, namely that the decision theoretic intuition of frame neutrality is an illusion. Indeed, framing is unavoidable as all intentional objects come under the guises of a frame or perspective. The only theory that investigates this possibility is Kahneman and Tversky's Prospect Theory.[7]

Insightful as it may be on the causes of the empirical invalidity of Juliet's principle, Prospect Theory's ambition is purely descriptive, and Kahneman explicitly resigned himself to the impossibility of reconciling the descriptive and normative dimensions of decision-making and reasoning. Worse, Bermúdez locates the origin of this impossibility in the discrepancies between the normative invariance (Juliet's) principle on the one hand, and observed systematic failures to comply with it on the other. In the light of this normative-descriptive tension, two strategies are available to us.

The first strategy of the extensionality skeptic relies on the notion of informational equivalence. The literature on informational equivalence[8] has attempted to refute the experimental

[5] Bermúdez 2020, p. 67.

[6] "*The chosen element depends on the opportunity set from which the choice is to be made, independently of how that set is described.*" (Arrow 1982, p. 12)

[7] According to Prospect Theory, the decision-making stage, which aims at selecting a favored action, intervenes after an earlier stage of edition and interpretation of the possible outcomes of a decision problem. At that stage, decision-makers first code quantities of goods in terms of gains and losses with respect to a neutral reference point, rather than in absolute terms. By determining the neutral reference point, she "sets the frame relative to which the value of different outcomes is fixed." (Bermúdez 2020, p. 84) Then gains and losses may be aggregated (considered as a whole) or segregated; they may be canceled out when identical or dominated. After this phase of edition and interpretation, decision-makers select a particular action based on a prospective value function, which quantitatively captures how and why people actually choose among prospects.

[8] See McKenzie and Nelson (2003) and Sher and McKenzie (2006) for an explanation of framing effects in choices

conclusions of invariance violation, when apparently coextensional frames bring in implicit relevant information to the decision maker: by making certain aspects more salient, certain descriptions introduce a reference point, norm or expectation, relative to which quantities are actually compared. Consequently some of what are taken to be paradigmatic violations of evaluative extensionality may actually be the result of informational asymmetry between so-called neutral frames. Here, Bermúdez sees informational equivalence as an argumentative springboard to specify his own normative conception of frame sensitivity. As the informational equivalence literature, it may question the descriptive experimental conclusions of extensionality violation by showing that the auxiliary hypothesis of informational equivalence between frames (here contexts) does not hold.

The second and more direct strategy available to the extensionality skeptic is to question the legitimacy and scope of Juliet's principle. Bermúdez chooses that latter path, and to that end, introduces Leibniz's principle of logical substitivity (the truth value of sentences is preserved by subject substitution and predicate substitution respectively, as long as the subject's names and the predicates' description refer to the same thing). These notions lay the groundwork for the notions of extensional, intensional and ultra-intensional contexts. "*Sentences for which the two substitutivity principles hold are typically called extensional contexts.*"[9] Sentences failing the substitutivity principle without making the agent irrational to the extent that he is "*unaware of the relevant identities*"[10] are called intensional contexts. Finally, ultra-intensional contexts are intensional contexts where the agent is aware of the relevant identities and yet is still deemed rational. If Bermúdez is right about frames, then sentences stating rational preferences or evaluations may be intensional if they are frame sensitive.

With these notions in mind, Bermúdez can introduce the main thinker of rational ultra-intensionality, Frederich Schick. Through three examples (Orwell, Sartre, and McBeth) of conflicted decision-makers, Bermúdez promotes Schick's intuitions and takes away from his work that (i) conflicts often consists in clashes between framings of the same situation, leading the conflicted agent to displaying quasi-cyclical preferences; (ii) these quasi-cyclical preferences can be rational if they are what Bermúdez calls ultra intensional contexts. If so, what is a rational frame?

Bermúdez makes a new case for ultra-intensional attitudes by refining Schick's claim of rational strong intensionality. Revisiting Orwell's case (cf. section 2.2), Bermúdez distinguishes two possible scenarios. In the former, one understanding (or frame) dominates the other, so that seeing the soldier as a human being prevented Orwell from seeing him as a fascist any longer. Schick's interpretation of the case seems to correspond to this first scenario. In that case, Orwell would not regret his choice ex-post, since his preferences have changed along with the change of frame. Under a second interpretation, one frame does not make the other disappear by

that are not informationally equivalent.

[9] Bermúdez 2020, p. 96.

[10] Bermúdez 2020, p. 96.

dominating it: it simply *downplays the significance of the other*. Orwell may still regret not shooting the fascist though he has no regret when it comes to sparing the human being. If this is the case, the clash of frames rationally persists after settling, and the agent remains conflicted after acting.

Bermúdez thus enriches Schick's framework of understandings by allowing agents to hold several understandings or frames simultaneously. One may come to dominate the other, but unlike Schick's understandings which can only get "switched on and off" binarily, Bermúdez's frames may be more or less recessive. This explains why Orwell may still experience ex-post regret on Bermúdez's interpretation, and not on Schick's. Bermúdez's strategy is not to make rationality constraints more permissible than Schick's account of understanding, but to pave the way for a more subtle account of framing, which rationally allows agents to enrich their representation. Then, where does Bermúdez draw the line between rational and irrational frames?

5.2.3 Rational frame sensitivity

Bermúdez makes a series of arguments in favor of multiple framing. His first argument (let's call it the Due diligence and complex reality argument) claims that normative theories of decision-making fail to take into account two major facts about rational agency. First, decision problems are not found but constructed, and should thus be "properly constructed". One way to capture this requirement of adequacy is the "due diligence requirement", Bermúdez version of Savage's "Look before you leap" requirement that states should leave "*no relevant aspect undescribed.*"[11] In Bermúdez's words, this rational obligation of care and diligence entails that the rational decision-maker should be "*appropriately sensitive to as many potential consequences of the different courses of action available to them.*"[12] Once the agent can be said to be facing a Savagian small world meeting that requirement, decision problems can reasonably be represented according to the traditional options/states/consequences partition. Yet, trying to meet the due diligence obligation often leads us to experience inner conflict and with it to quasi cyclical preferences, as Agamemnon's story illustrates.

Through a prophecy, the King of Mycenae learns that if he wishes to remain loyal to his ships and people, he must appease Artemis by sacrificing his daughter Iphigenia. Agamemnon thus frames that eventuality both as the "murdering his own daughter", but also as "following Artemis' will". It is both true of him that

(i) he prefers "following Artemis' will" to "Failing his ships and people", and

(ii) he prefers "failing his ships and people" to "Murdering his own daughter".

According to Bermúdez, the quasi cyclical preferences induced by these two frames only stem from Agamemnon's striving to attend to the complexity of the situation, that is to say to satisfy the obligation of due diligence. To those who would object that Agamemnon is irrational because

[11] Bermúdez 2020, p. 77.
[12] Bermúdez 2020, p. 121.

he is inconsistent, Bermúdez offers an interesting counter-argument. Consider another version of Agamemnon, Agamemnon minus, who only sees Iphigenia's death as obedience to divine command. That individual has a unique, consistent frame and exhibits transitive preferences. It could thus be argued that Agamemnon minus is more rational than Agamemnon on the ground that he is more consistent than him. Yet Bermúdez maintains that Agamemnon minus is not any more rational than Agamemnon, since "*an important part of rational decision-making is properly understanding the decision problem one faces. [. . .] [For a consistency-based theory of rationality like Hume's,] failing to recognize the complexity of a decision problem seems to me to be a strike against that theory of rationality.*"[13] In complex situations, the violation of invariance may be the symptom of our ability to take multiple complementary perspectives on the associated decision problems.

What I call the "rational emotional response" argument relies on the relationship between frames, emotions, and valuations. In the various examples mentioned so far, different frames give rise to different emotional responses. And there is nothing wrong with having frame-dependent emotional responses. "*If there is irrationality with this phenomenon, it is not due to the fact that emotions are involved, but rather that a perturbing fact, such as a false belief or a failure of theoretical rationality.*"[14] Moreover, it is reasonable to grant that different emotions lead to different valuations of an outcome. So all in all, if it can be rational to frame an outcome in different ways (as Agamemnon), and the associated emotional responses and subsequent valuations are rationally permissible, then so are the preferences resulting from it, even if quasi cyclical.

Empirical studies suggest that cases of diligent, quasi cyclical preferences arise in much more ordinary circumstances than Homeric tragedies, through the relationship between frames and emotions. Contrary to an implicit decision theoretic conception, emotions are not reducible to mere distractions interfering with rational choice; emotional engagement pervasively impacts framing and evaluation. Importantly, the converse mechanism may occur as frames also impact our evaluations through emotions.

In his paradigmatic example, Bermúdez introduces public discussions on climate change, whose complexity may be explained by the fact that it may be viewed either as a natural or a man-caused phenomenon. Suppose I want to contribute to a charity, and hesitate between Climate Change Relief and the International Rescue Committee for war victims. I feel that victims of natural disasters deserve more help than war victims, but when it comes to man-caused disasters, war victims are the priority. Under the natural disaster frame, I would thus prefer giving to CCR than to the IRC, but I would reverse my preferences under the man-caused frame. Were I to frame it both as a natural and a human phenomenon, I would display quasi cyclical preferences. And Bermúdez makes an interesting case that both frames happen to be perfectly legitimate on this particular issue to the extent that "*each of them reflect and highlights a genuine aspect of a complex phenomenon*,"[15] as long as they do not involve false or explicitly inconsistent beliefs.

[13] Bermúdez 2020, p. 117.
[14] Bermúdez 2020, p. 135.
[15] Bermúdez 2020, p. 132.

What is more, even when we violate extensionality because of a failure to update our beliefs in face of new evidence, this epistemic violation should be weighed against "*the need for a theory of rationality to respect the constraints of the informational situation in which people find themselves.*"[16]

The general upshot of Bermúdez's claims is that frames are "*partial perspectives*"[17] that may be rationally combined as long as they are complementary, epistemically sound, and consistent with one another. However, unlike in Schick's theory of understandings, the fact that some frames prevail over others does not mean that the dominated frames become negated: as Agamemnon eventually adopts the murder frame and renounces to sacrifice his daughter, the aspects or dimensions induced by the religious frame only recede in the background, their significance being downplayed. He may still believe that killing Iphigenia would satisfy Artemis will, yet Agamemnon does not stop endorsing that dominated perspective for all that: "*he is not denying that Iphigenia's death would satisfy Artemis [. . .], he is [rather] downplaying [its] significance.*"[18] In Bermúdez's view, one may be logically consistent while holding different valuations of an outcome under different frames, if one "*downplays the significance*" of a dimension of the outcome or option considered.

Let's take stock on Bermúdez's argumentative strategy so far. Theories of decision making widespreadly agree about the evaluative neutrality of frames, where frames may be broadly understood as descriptions of outcomes. Except for too few notable exceptions, the pre-theoretic principle of evaluative extensionality (valuing/desiring equally two *extensionally equivalent* outcomes) is one of the most uncontroversial auxiliary requirements within decision theory. The domain of application of these extensionality requirements is to be found at the representational phase of decision making, concerned with the identification and computation of available options and their associated outcomes. And the success of this first stage depends on the agent's ability to take into account how her environment constrains these features of a decision problem.

As Schick, Bermúdez supports a form of strong intensionality which seems to shed light on cases of conflictual situations like Orwell's. Contrary to him, Bermúdez's chief concern lies in determining the normative foundations of ultra intensionality, and delineating its scope: and in this respect, the fact that Orwell could have "*viewed the situation under two frames is not a sign of irrationality; deliberative rationality does not require him to eliminate one of the two frames.*"[19] Indeed, if Savage's "Look before you leap" assumption is to be taken seriously, then rationally complying with it should encourage us to have multiple frames of the same options. Secondly, the intimate interactions between emotions, and frames (via evaluation and motivation) are such that rationally endorsing frame-sensitive emotions requires accepting cases of frame sensitivity *tout court*. However, since frame sensitivity often arises from situations of inner conflict, making it rationally acceptable implies accepting quasi cyclical preferences such as Orwell's. Yet Agamemnon's case suggests that if anything, seeing situations through multiple

[16] Bermúdez 2020, p. 133.
[17] Bermúdez 2020, p. 134.
[18] Bermúdez 2020, p. 134.
[19] Bermúdez 2020, p. 104.

frames is more rational than following the Humean prescription to see the situation through a consistent but unique lens. As the climate change example illustrates, multiple framing is characteristic of the complex nature of the decision situations we face. Strong support in favor of this conclusion can be found in the role of frame-sensitive emotions for successful rational self-control.

5.2.4 From self-control to non-Archimedean reasoning

In standard temptation cases, agents exert self-control in a sequential choice problem, and may frame the delayed gratification resulting from resisting temptation as either a "delayed larger reward later" or as "having successfully resisted temptation". By framing it in the second way, the agent is exposed to a source of motivational engagement induced by the emotional response to the frame, which actually enables her to exert successful self control. If this is correct, Bermúdez just offered a powerful argument in favor of rational ultra-intensionality. Resisting temptation requires accessing different evaluations of the same outcomes (the tempted and the resolute evaluations), which is not condoned by standard decision theory. But how could reframing be wrong when it allows agents to achieve self-control?

Collective rationality can also benefit from a theory of rational frames. Bermúdez applies his previous intuitions to game-theoretic issues of cooperation, fairness, and to discursive impasses. In particular, intense disagreements in the public space can often be understood as clashes of frames rather than simple oppositions of beliefs and desires. This is a key moment in Bermúdez exposition, as his normative account of rational frames hinges on his views on discursive deadlocks. His strategy consists in offering a solution to interpersonal conflicts due to clashing frames which requires individual agents to be frame-sensitive enough to reason across frames, and simultaneously and consciously see the situation through several frames at a time. *When a problem is complex enough, clashing frames, with incompatible prescriptions, evaluations and emotions can still be somehow consistent.*

It is useful at this stage to distinguish between factual and non factual propositions, where factual propositions are taken to be "straightforwardly true or false", in the sense that their truth value can be determined by what Bermúdez calls standard techniques. Technical standards vary with context, and may range from personal experience or common sense to uncontroversial empirical investigation. For instance, propositions involving personal identity and personhood are not strictly factual. By contrast, the claim that "a five-week old fetus has a heart" is a factual proposition. Even prophecies may count as factual propositions if the decision-maker in his epistemological context treats them as such. Bermúdez relies on this distinction to both redefine and weaken requirements of non contradiction between beliefs: p and not-p can be simultaneously believed as long as they are not factual propositions. So propositions that are not straightforwardly true or false cannot be required to be non-contradictory. The distinction is motivated by the idea that contradictory beliefs on points of facts are more irrational than contradictory beliefs on non factual issues.

In comparison, Schick did not allow inconsistent beliefs about materially equivalent propositions (strict intensionality of beliefs), though he permitted different understandings and valuations of such propositions. And two propositions are materially equivalent if the physical self-identities they involve are independent or necessary. Bermúdez does not resolve to follow Schick's path as he rightly points out that two propositions describing different outcomes may still be materially equivalent. If so, Orwell's two understandings (of his shooting the fascist soldier and the human being) are only rational if we manage to show that the two propositions are not materially equivalent. Bermúdez thus prefers to build his normative account on a notion of factuality rather than material equivalence and identity, which makes him more permissive on contradictory beliefs than Schick.

Bermúdez does not use the factuality criterion only to loosen the grip of material equivalence over non-factual beliefs, as he also adds *a non-falsity requirement for factual beliefs themselves*. The author justifies this unexpected requirement from Hume's principle: "In short, a passion must be accompanied with some false judgment in order to its being unreasonable; and even then it is not the passion, properly speaking, which is unreasonable, but the judgment."[20] I think this justification relies on a misinterpretation of Hume's principle, which only holds that the irrationality of a passion must come from the irrationality of the judgment. This neither is equivalent to a non-falsity requirement as the irrationality of beliefs may be independent from the truth-value of the corresponding facts; nor equivalent to the fact that irrational beliefs entail irrational passions (the contraposition of Hume's principle).

In fact, most Humean accounts of decision-making are Humean about the a-rationality of desires, but do not require that beliefs be true. After all, many rational decisions are made despite the presence of mistaken beliefs. Why would Bermúdez then make it a prerequisite of rationality? The reason is to be found in his previous argument in favor of the rationality of Agamemenon's frames and preferences. One of his key arguments was that an inconsistent Agamemnon is more rational than Agamemnon minus, who true, is consistent, but fails to capture the complexity of the situation. The price to pay to understand complexity is inconsistency, and complex inconsistent understanding is better than a simple consistent one. **All in all, his account of frame-sensitive beliefs is highly flexible with respect to non-factual propositions and rather demanding for factual ones.** The requirements he sets on beliefs are summarized below.

Consistency requirements:

- Rational agents cannot simultaneously hold contradictory beliefs about *factual* propositions.
- Rational agents cannot believe that p and not-p for any sort of propositions.

No-false factual beliefs requirement:

- Rational agents cannot have false factual beliefs.

[20] Hume 1995, § 2.3.3.6 p. 415.

Moreover, he reformulates Savage's "Look before you leap" principle in terms of appropriate sensitivity:

Due Diligence requirement:

- Rational agents should be appropriately sensitive to as many potential consequences of the different course actions available to them as possible.

Now that he has laid the fundamental rationality requirements of frame-dependent beliefs, Bermúdez can focus on the positive part of his account, which promotes a class of reasonings and evaluative skills which allow agents to reason across frames. Bermúdez's theory goes beyond the kind of "Archimedean reasoning"[21] described by decision theorists and which takes decision-makers to start from an Archimedean point — "a fixed perspective"[22] to understand and assess the world: indeed, Expected Utility Theory takes the decision-maker's utility function to be the Archimedean point, hedonist utilitarianism takes the scale of pain and pleasure as a fixed frame of evaluation. Frame-sensitivity is thus incompatible with Archimedean reasoning, and worse, issues of frame sensitivity occur in cases of quasi-cyclical preferences and clash of values, precisely when Archimedean points are no longer useful.

Bermúdez constructs a more comprehensive account of rational agency embodied by what he calls the frame-sensitive reasoner. The frame-sensitive reasoner goes beyond standard decision theory as she does not limit herself to consistency (which can be achieved by eliminating frames), but attempts to enrich her understanding and evaluation of complex situations. In the light of psychological theories of frame sensitivity, Bermúdez infers a series of techniques of appropriate framing-sensitive reasoning. Four of them hold Bermúdez's attention, as they correspond in his view to the four dimensions of non-Archimedean frame-sensitive reasoning, namely:[23] *reflexive decentering*, *imaginative simulation*, *perspectival flexibility*, *reason construction and juxtaposition* (see figure 5.2 on p. 119).

When *reflexively decentering*,[25] agents step outside of their current frame, and recognize the partiality of understanding a situation via a single frame. It involves meta-awareness about our thoughts and emotions, disidentification from our internal states as separate from ourselves, and reduced reactivity to thoughts and emotions. Combined, they allow the content of representation to shift from the initial frame to a set of frames. The second dimension, imaginative simulation, allows reasoners to imagine "*what it would be like to frame things completely differently, and to simulate actually being in this frame*,"[26] along with the beliefs, desires and emotions which are primed

[21] Bermúdez 2020, p. 240.
[22] Bermúdez 2020, p. 240.
[23] Bermúdez 2020, pp. 243–245.
[24] Bermúdez 2020, figure 11.1 p. 245.
[25] Bernstein et al. 2015.
[26] Bermúdez 2020, p. 243.

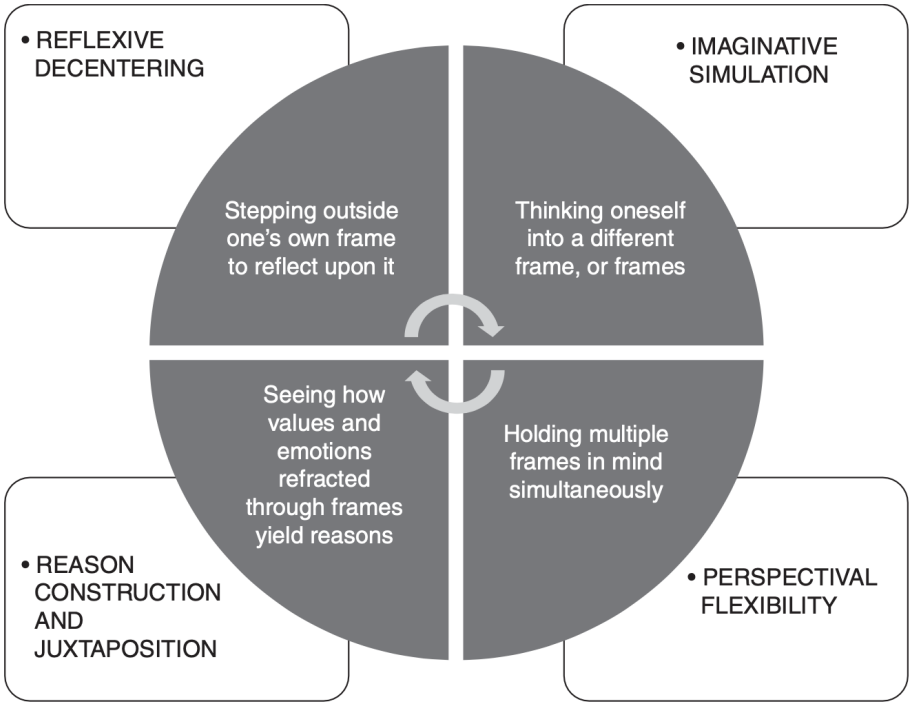

Figure 5.2. Bermúdez's key framing techniques for non-Archimedean, frame-sensitive reasoning[24]

by it. That is because "*frames are often more fundamental than beliefs and desires.*"[27] At the same time emotions influence frames too. Therefore, imaginative simulation also requires simulating emotions giving rise to certain frames typically produced by it.

Perspectival flexibility helps the agent evaluate reasons by enabling her to adopt several frames simultaneously. Bermúdez builds his conception of perspectival flexibility on Selman's theory of social role-taking in children's development[28] and Fisher and Ury's theory of principled negotiation and non-positional bargaining.[29] Both theories presuppose a model of interpersonal framing and perspectival flexibility between self-interested individuals. "*At the ideal limit, the participants would each have internalized the frames and perspectives of other participants, so that each would be approaching the problem with the available conflicting frames in mind. Only then will they be able to start applying problem solving principles such as Fisher and Ury.*"[30]

If reasons are frame-relative and only extensional within a frame, the existence of a complete, frame-sensitive and extensional ranking of outcomes is not guaranteed. Moreover, Bermúdez holds that preferences are rational if they are grounded in reasons which are salient under some frame. In cases of incommensurability or unresolved conflicts, none of them prevail, regardless of the agent's eventual choice: reasons are not orderable but only *juxtaposed*. Frame-relative reasons are accessible by two processes. Frame decomposition or *deconstruction* extracts the

[27] Bermúdez 2020, p. 253.
[28] Selman and Byrne 1974.
[29] Fisher and Ury 1981.
[30] Bermúdez 2020, p. 262.

values expressed by a frame and the emotions driving them. Frames are often embedded in narratives, each offering various reasons in favor or against a certain decision. These reasons can be identified and weighed across frames: instead of opposing narratives and factual claims in a sterile way, one needs "*shift the term of the debate from non factual propositions masquerading as factual ones, to a constructive discussion about the values that underlie those non factual propositions.*"[31]

For instance, apparently factual oppositions about the costs and benefits of GMO on health should be deconstructed if one wishes to understand the values that structure and infuse a particular frame. Anti-GMO activists do not simply hold on to false factual beliefs for Bermúdez, they also have a particular conception of nature (rural and agricultural). Making this conception and its associated values explicit allows us to critically examine them "*on [their] own terms*"[32], by constructively discussing the non-factual propositions of each frame. Reasons and values from each competing frame can now be weighed. If values are commensurable, some will win over others, and the conflict between frames will be dissolved.

Only then can Archimedean reasoning and extensional decision theory can be applied. A common way of settling for an Archimedean frame is by elimination of unacceptable or repugnant frames. At one extreme of the normative spectrum, extreme Humeans endorse any elimination of frames on the ground of one's values and deeper sentiments, while the other side claims that a full understanding of all possible frames is a requirement of rationality. For Bermúdez, neither extreme is normatively plausible; in particular, what he calls extreme humeanism fails the Due Diligence requirement because it amounts to "*persisting in a discursive deadlock without trying to engage with alternate perspectives.*"[33]

What is the normative status of non-Archimedean reasoning for Bermúdez? We know that it prevails over Archimedean reasoning when the complexity of the situation justifies it. That's what he means when he says that the frame-sensitive reasoner should, as Agamemnon, be "*appropriately sensitive*"[34] to as many potential consequences and available actions as possible, rather than stick to simplistic consistent frames. This Due Diligence requirement of frame sensitivity is specified by the standards of non-Archimedean reasoning and its four dimensions. Whether the associated operations of reason are requirements or permissions, Bermúdez does not say, but the Due Diligence Requirement seems to make them necessary to be qualified a frame-sensitive reasoner and a rational agent. Before assessing this issue further, we give a brief summary of Bermúdez final normative account of rational framing of decision problems in table 5.1 on p. 121.

[31] Bermúdez 2020, p. 269.
[32] Bermúdez 2020, pp. 268–269.
[33] Bermúdez 2020, p. 270.
[34] Bermúdez 2020, p. 121.

Table 5.1. Bermúdez's standards of rational framing

Consistency requirements:
- Rational agents cannot simultaneously hold contradictory beliefs about factual propositions.
- Rational agents cannot believe that p and not-p for any sort of propositions.
- Rational agents cannot have false factual beliefs.

Due Diligence requirement:
- Rational agents should be appropriately sensitive to as many potential consequences of the different course actions available to them as possible.

Rational frame-sensitive reasoning:
- reflexive decentering
- imaginative simulation
- perspectival flexibility
- reason construction and juxtaposition

5.2.5 Reframing and cross-frame evaluation

To be able to adopt a fixed Archimedean valuation of the outcome, one must first step out of her frame to see the outcome from another perspective; then possibly construct an overarching frame of the outcome to weigh reasons across frames if commensurable. Reasons are then the fundamental values extracted from non-factual frames dressed as factual ones. Bermúdez contends that "preferences must be held for a reason", and "*different frames bring different reasons into play.*"[35] A skilled frame-sensitive reasoner should thus be able to reframe an outcome in terms of the core evaluations common to all relevant frames, these newly framed reasons being the adequate justifications for the all-frame considered preference between two outcomes.

The only three attitudes characterizing the frame-sensitive agent in Bermúdez's framework are beliefs, emotions, and preference relations. The agent's evaluation of an outcome under a frame is determined by the frame itself as well as the emotions associated with it. Both frames and emotions impact the evaluation of an outcome, and rationality requirements on frames are formulated in terms of beliefs exclusively. As we will see next, this is made possible by the distinction between factual and non-factual propositions.

5.2.6 Factual and non-factual propositions, methods of cross-frame evaluations and repartitioning

The distinction between factual and non factual proposition is best illustrated in the GMO example. Understanding the anti-GMO side first requires identifying the factual beliefs of its frame, deconstruct them by establishing the particular semantic meaning they attach to

[35] Bermúdez 2020, p. 244.

key syntactic items ("nature" as "a rural and agricultural" conception[36]), and inferring the foundational values underlying this conception (e.g., the value of "tradition"). This way, previously implicit non-factual propositions may be discussed and, in fortunate cases, be weighed against each other.

Let's see how the GMO example may be formalized: p: "GMOs are damaging to human health" and q: "GMOs are damaging to a valuable rural and agricultural lifestyle". How do we infer the non-factual proposition q from the factual one p? Presumably, by relating human health to natural human practices, and by making the relevant rural and agricultural lifestyle part of these natural practices.

In Bermúdez's formal framework, evaluative attitudes such as desirability are not formalized syntactically, and are thus not involved in the operations between propositions. Evaluations are already involved in the semantic content of the proposition which may only be believed or disbelieved. The separation between evaluative and epistemic attitudes instead takes place in the bidimensional space of possible facts and propositions ($F \times P$). It corresponds to a mapping V which associates any fact f to some proposition p when the truth value of p can be determined by pre-accepted techniques of verification. The complement of V's image in P is the set of non-factual propositions such as evaluative and deontic ones.

Then, Bermúdez's point can be understood as the fact that $V(P)$ is only a set of apparently factual propositions, from which should be distinguished all disguised nonfactual propositions, like q: "GMO's are damaging to lifestyles that value rurality and agriculture as natural human practices". Propositions like q can be broken down into q_1: "rural and agricultural practices are natural human practices" and q_2: "rural and agricultural practices are valuable". q_1 is akin to the proposition "a one-week old fetus is a person" in that they both make a definitional or individuational claim: "the set of As is included in the set of Bs"; where B involves non standard controversial claims about metaphysically individuated entities like personhood and nature. By contrast, q_2 is a non-factual evaluative proposition, similar to "the life of a person is valuable". Instead of focusing on factual and "metaphysical" propositions, one should shift the discussion and deliberation to propositions about foundational values like q_2 and its analog on the pro-GMO side q_2'.

In deconstructing frames along the factual/non factual dimension, Bermúdez opens up the possibility of understanding the outcome under a new, consistent, overarching frame within which Archimedean reasonings such as the elimination of dominated alternatives and reason-weighing. If such a frame is successfully designed, it must adequately specify the evaluative relation between q_1 and q_2. For instance, a frame including beliefs such as q_3: "the value of the human lives saved by GMO outweighs the value of the associated damage done to natural human practices" would do the trick.[37] The success of deconstruction depends on the possibility

[36] Bermúdez 2020, p. 268.

[37] Interestingly, note that Utilitarianism can be interpreted as a systematic method of forming an overarching frame of particular decision problems, which atomistically deconstructs non factual propositions in terms of predefined

of untwining evaluative and factual uncertainty. The former is then addressed by reconstruction and reevaluation via non-Archimedean reasoning.

Let's summarize Bermúdez's demonstration. Some decision situations like Agamemnon's instantiate rational violations of Juliet's principle that justify an ultra-intensional conception of frames. In such cases, Humean consistency prescriptions like Juliet's principle and extensionality leave relevant aspects of the outcome undescribed, and thus violate the Due Diligence requirement. In order to know when Juliet's principle applies, the decision-maker must be able to see the world through multiple frames, to correctly disentangle factual from non-factual propositions, and to attempt to construct a comprehensive consistent frame in which he can now weigh reasons extensionally. In what follows, I examine how this account of framing compares to the Humean account of representation defended in this dissertation and argue in favor of the latter.

5.3 Differences between Bermúdez's account of frames and the account defended in chapters 3 and 4

Bermúdez's *Frame It Again* constitutes the first concrete account of representation that attempts to synthesize the conventional normative constraints imposed by theories of rational agency and decision-making, along with the structural constraints of plausibility imposed by folk theory and empirical research on framing effect. As Schick, he commits to a psychologically credible theory of representation, yet unlike him, he successfully escapes most criticisms of normative implausibility that Schick may have been exposed to. Acutely aware of the tension between descriptive and normative decision theory (see Bermúdez 2009), his ambition to challenge the academic doxa on the extensionality principle is in my view partially successful.

In this section, I first examine how our accounts of representation and framing effects differ. I then turn to Bermúdez's claims about rational representations and look into two issues. I first argue that his canonical example (Agamemnon) is not sufficient to reject the invariance principle nor to justify the introduction of ultra-intensional frames, consistently with chapter 2. Irrational framing effects can be defined independently from factual neutrality as inconsistent choices across evaluatively equivalent decision situations, as defended in chapter 4. Second, I contend that the specification of factual propositions in a decision problem is not evaluatively neutral, which in turn violates the Due Diligence requirement. If this is correct, this pleads in favor of accounts that acknowledge that all specifications of decision problems and their consequences are value-driven.

atoms of value such as "human lives" or pleasure and absence of pain". The appeal of such methods lies in the fact that they offer a meaningful, transparent, and applicable way of reevaluating the fundamental values of particular frames, and of weighing them across frames.

5.3.1 Comparing Bermúdez's requirements with those of the present account of representations

Tables 5.2 and 5.3 on pp. 125–126 summarize and compare Bermúdez's account of rational framing with the account of representation defended in the previous chapters.

If we compare Bermúdez's account with the account of representation sketched by the present dissertation, the first important difference regards the status of externalist requirements of rationality. Bermúdez's frames are subject to factual requirements of truth and consistency that are externalist in two ways. As descriptions, frames represent an external world regulated by facts. The domain of representation of frames is thus external. By contrast, I defended an internalist account of representation, where decision problems are formal representations of decision situations constituted of features the agent cares about. The domain of representation is thus defined in reference to the agent's beliefs and values and not in reference to the external world.

Bermúdez's account makes a second externalist requirement via true factual beliefs. Since facts are defined with respect to standard methods of verification, the agent does not need to know these methods in order to rationally frame the associated fact. By contrast the present account only makes requirements of structure and of logical consistency on internal attitudes, in agreement with a moderate Humean conception of practical representations. It therefore rejects the factual truth condition.

While the Due Diligence requirement demands the inclusion of all possible consequences of a course of action, the present account only requires the agent to include consequences that are cared for and screened by the agent's instrumental plans. In that sense, Bermúdez's account is more constraining than mine and leaves the issues of scope and grain of frames untouched. Indeed, Due Diligence can be understood as the requirement to maximize the scope of the decision problem, while instrumental plans set sufficing requirements on the scope and grain. For instance, Due Diligence implies that Agamemnon minus is less rational than Agamemnon because the scope of his decision problem is strictly smaller than his counterpart, when the present account does not.

How do the two accounts compare on framing effects and their rationality? Let's first note that both of them are interested in framing effects that do not stem from mistaken or inconsistent beliefs as in simple illusion cases. Simple cases of perceptual illusions do not threaten the extensionality of beliefs since the agent is not aware of the illusion. For instance, someone inconsistently guessing the length of two segments because of a visual illusion is typically not aware of the fact that the two segments are perfect copies. If he were, both accounts would conclude to the irrationality of the agent, but for different reasons. Bermúdez would argue that since the identity between the segments is a factual identity, the agent is strictly inconsistent and thus irrational. By contrast, my account would blame him for the logical inconsistency of his beliefs.

Table 5.2. Comparative summary of Bermúdez's and the present account (1/2)

Bermúdez's frames	Care-based instrumental representations
Structural requirements: • Frames are propositions that may be factual, evaluative or metaphysical. • The same option or outcome may be seen under multiple frames simultaneously.	*Structural requirements:* • Decision-theoretic consequences are features making a difference the agent cares about. • These features can be represented as intensional properties of the alternatives. • Care-desire model of agency: agents connect to their cares through deliberation. • These consequences can be treated as weighable properties.
Consistency requirements: • Rational agents cannot believe that p and not-p for any sort of propositions.	*Consistency requirements:* • No contradictory beliefs requirement for all propositions the agent is aware of. • Rationally intended ends and their reconsideration should be determined by cares. • Coherent beliefs, desires and intentional plans as per Bratman's theory. • Coherent specifications of instrumental decision problems. • Consistent evaluations of general and particular properties as weighed reasons.
Externalist requirements: • Rational agents cannot have false beliefs about factual propositions. • Coreportive factual propositions ought to be valued equally. • Rational agents cannot simultaneously hold contradictory beliefs about *factual* propositions.	*All rationality requirements distinguish:* (i) external and internal point of view of assessment (ii) rational criticizability and rational blameworthiness (iii) unreflective, policy-based, and deliberative specifications of the decision problem

Table 5.3. Comparative summary of Bermúdez's and the present account (2/2)

Bermúdez's frames	Care-based instrumental representations
Due Diligence requirement: • Rational agents should be appropriately sensitive to as many potential consequences of the different course actions available to them as possible.	*Instead of Due Diligence requirement:* • Agents are only required to find sufficient means to their ends.
Irrational framing effect: • The agent values the same *factual* outcome or act differently under two frames.	*Irrational framing effect:* Suppose that: (a) Each of the decision situations can be fully characterized by general differences between options, and these differences are the same in both situations. (b) The two decision situations do not differ in any particular way the agent cares about, so that these general differences equally apply to both situations. Then, the two choice situation ought to be treated equivalently, and only frames yielding equivalent choices across situations will be rationally admissible.
Rational frame-sensitive reasoning: (Process rationality that comes in degrees) • reflexive decentering • imaginative simulation • perspectival flexibility • reason construction and juxtaposition	

Evaluative inconsistency can stem from inconsistent beliefs: if one had to make bets about the lengths of the two segments while being aware that they are replicas, inconsistent beliefs about their lengths would lead to inconsistent evaluations of the bets. Yet neither Bermúdez nor I are after this kind of indirect evaluative inconsistency. Remember that if one wants to show that Juliet's principle does not hold, one ought to make a case for decision situations where one can rationally value the same outcome under different frames.

In chapter 4, I argued that we can meaningfully define illegitimate framing effects without resorting to semantic requirements. A decision problem is a selective representation of a decision situation which consists in a set of relevant features. Then, one can define equivalent decision situations via certain identity and difference conditions between the properties of each situation. The decision-maker cannot be subject to illegitimate framing effects unless these conditions are met.

This definition of framing effects contrasts with Bermúdez's in two ways. First, it introduces a distinction between legitimate and illegitimate framing effects independently from an invariance principle requiring equal desirability of coreportive factual propositions. This difference can be illustrated in the Asian disease case: in Bermúdez's account, an agent failing to value equally "not saving n people" and "letting die n people" will be irrational if these propositions are (coreportive) factual ones and believed by the agent. In my account, the same agent will be charged with irrational framing if the agent does not care about the difference between "not saving" and "letting die". Then, the irrationality of the framing effect does not stem from a failure to equally value descriptions of the same factual outcome, but from inconsistent choices in evaluatively equivalent decision situations. Consequently, Bermúdez's requirements on framing effects are more constraining than mine for factual propositions while they are less constraining than mine for non factual ones.

That said, both accounts share the view that orthodox decision theory is over restrictive regarding framing effects, only for different reasons. While Bermúdez deems the invariance principle illegitimate, the present account does not reject it but claims that the principle is not adequate for distinguishing rational from irrational framing effects.

Bearing in mind the key similarities and differences between the two accounts, I now argue in favor of the Humean view of decision-theoretic representation defended by this dissertation.

5.3.2 Assessment of Bermúdez's normative claims

In this section, I first review Bermúdez's key representational claims before focusing on the argument against the invariance principle he develops in his analysis of Agamemnon's dilemma. I argue that the argument does not go through for reasons that are similar to those I gave in Orwell's case in chapter 2. I then look into the distinctive role of factual propositions in Bermúdez's account, and argue that the specification of factual propositions in a decision problem is not evaluatively neutral, which violates the Due Diligence requirement

Agamemnon's dilemma

Let's have a look at Bermúdez's key claims regarding rational representations:

(i) Frames are crucial to our understanding of how actions arise, particularly when the agent faces several clashing frames of a given situation.

(ii) Conflicting *frames* easily lead agents to quasi cyclical preferences.

(iii) While conflicting *decision situations* may be solved by settling on a particular frame, it is not always the case.

(iv) A decision-maker may rationally prefer A to B and B to C though *they know* A and C to be *identical outcomes framed differently*. When this is the case, the underlying preferences are said to be ultra-intensional (strongly intensional for Schick).

(v) The agent's experience of ex-post regret may be a sign of his quali cyclical preferences.

Claims (i) and (ii) rely on arguments similar to Schick's (Orwell's fight against fascism, exposed in chapter 2, is one of Bermúdez's central argumentative examples), into which I will not delve too much. As in the analysis of Orwell's case I made in chapter 3, Agamemnon's conflictual situation can be modeled as a case of incommensurable properties without violation of extensionality. The instantiated differential features Agamemnon cares about, "murdering his daughter" and "following Artemis will" are not the same properties, and may be valued differently. The problem can thus be described as the choice between A: "following Artemis' will" and "murdering his daughter" vs B: "failing his ships". If Artemis' will is only valued as a means to be loyal to ships, then the choice problem can be reduced to "murdering his daughter" vs "failing his ships". In any case, the conflict stems from the fact that he is unable to make an aggregated evaluation of the set of properties instantiated by each option. If this is correct, extensionality is not violated, and Agamemnon's case does not play against the litany of rationality.[38]

Moreover, describing Agamemnon as "in the grip of a framing effect" does not look like a charitable interpretation of his situation: To the best of my understanding, Agamemnon is not consciously valuing the same outcome differently, but faced a difficult choice between two options with incommensurable sets of properties, and rationally went for one of them. Acts typically have a variety of consequences which may be hard to weigh against each other, this does not imply that all of them are cases of extensionality violation, although this may seem so when we describe the option by one of its intentional consequences (here by successively describing

[38]One reason why Bermúdez may refuse to proceed that way is that he is aware of Broome's emptiness problem (Broome 1991): "*one can always get rid of apparent counterexamples to transitivity by redefining the outcomes.*" (Bermúdez 2020, p. 81) If I understand this correctly, what Bermúdez is saying is that his own description of Agamemnon's case cannot be discussed if the emptiness problem is a genuine problem. Broome's argument had traction because the decision problem offered to Maurice was already formulated (going to Rome, staying home, mountaineering) and Maurice refined the initial outcomes. Here, Agamemnon is not in a preformulated small world as Maurice, but in the grand world of Aeschylus' book. What is the most faithful description of Agamemnon's situation? I would not know, and we don't need to know so to conclude that describing it as a set of properties or propositions does not constitute a "reindividuation" of any kind since no formal decision problem was offered to Aeschylus' Agamemnon in the first place.

the first option as the consequence "complying with Artemis' will" and as "murdering his own daughter").

By contrast, the account of representation introduced in chapters 3 and 4 allows Agamemnon to be conflicted within the same frame. His initial plan of leading his ships to war met an obstacle giving rise to a decision situation, between reconsidering his initial plan or following through at the cost of murdering his daughter. If deliberation is successful, Agamemnon may have determined the consequence that mattered most to him. Consistently with Bermúdez, this does not mean that the concerns in favor of the dominated option disappear once Agamemnon's decision is made. Shoemaker (2003) rightly observes that in a choice between two objects of care, the forsaken option does not stop being cared for.[39]

The specification of factual propositions in a decision problem is not evaluatively neutral, which violates the Due Diligence requirement

Although both accounts agree that representations are not evaluatively neutral, they depart on the status of factual propositions. Bermúdez's account stipulates that the truth value of factual propositions can be determined by standard techniques; moreover, the rational agent ought to hold true factual beliefs on Bermúdez's view.[40] By contrast, my account does not treat factual and non factual descriptions differently. One reason for this is that the separation between beliefs and desirability already tracks the distinction between epistemic and evaluative attitudes about features of the domain of representation (i.e. of the decision situation).

A second argument for treating factual and non factual propositions on a par is that as far as practical rationality is concerned, the representation of factual propositions is not evaluatively neutral. Framing a feature as present or absent will depend on pregiven evaluative attitudes in virtue of which the outcome is framed that way. For example, a hill may be treated as an obstacle or a vantage point depending on the agent's pre-established concerns. Informally, one could say that the same object — a hill — may be described differently, or that the same feature can be framed as different properties.[41] Loosely using Bermúdez's vocabulary, it could be said that the presence of an obstacle is a non factual proposition (perhaps disguised as a factual one), while the presence of a hill is a factual proposition. Yet this way of speaking and of distinguishing factual and non factual descriptions is misleading. The fact that a feature is not explicitly described as an evaluative proposition does not imply that the presence or the absence of that feature are evaluatively neutral. This is particularly true in decision situations, where the absence of a certain feature is practically relevant. Suppose I have to relocate to another city for professional

[39] "*Further, because caring is a matter of degree, so is identification. To the extent that I care about something moving me to* Φ *and I also care about something else moving me to* Ψ*, I am partially identified with both potential sources of my action. I will then be led by my ambivalence in this situation to reflect upon my various carings, and the caring(s) that emerges as stronger will motivate my actions. But in doing whatever I do, I do not shunt aside the losing, merely weaker care as alien, as external to me. It — and the psychic elements to which it gives birth — is still a part of who I am, just not as much as the winning care. It will shape my will on other days, but not today.*" (Shoemaker 2003, p. 112)

[40] Bermúdez 2020, p. 233.

[41] See section 2.7. Conflicting consequences can be modeled as properties without violating extensionality.

reasons. The fact that a particular city does not have a high school will be a relevant feature of the decision situation if I care about education. Similarly, I may justify representing a certain landscape as lacking hills by the fact that the presence of hills is not evaluatively neutral for me.

It could be objected that the absence of a hill is not a fact. I don't have any precise views on the matter, but I believe this would only strengthen my initial contention that facts cannot justify the inclusion of the absence of a feature into a decision problem. If facts can't play that justificatory role, may beliefs play that justificatory part? Indeed, it would make sense from a Humean perspective to substitute beliefs to facts as a justification for representing the absence of a feature. In particular, one could appeal to the believed differences between two options: the absence of a feature ought to be included in the representation when the agent believes that the two options differ in some way she cares about. However, we have seen in chapter 4 that some properties of the decision situation play an evaluative role over and above simple atomistic differences between two alternatives. Consequently, it seems that including the absence of a feature into the representation cannot be justified only by appealing to believed differences between options.

A third reason to question the frame neutrality of facts is that some decision situations may require to deliberately omit certain factual descriptions from the decision problem. Examples of such cases were given in chapter 3. A jury in a trial may be asked not to take into account certain pieces of factual evidence for instance. Consequently, agents with different values may disagree on which factual propositions should be included into the decision problem. This jeopardizes the possibility of constructing an overarching frame that is consistent while doing justice to the initial frames when considered in isolation. Applied to the intrapersonal case, seeing the same situation under two frames with distinct (yet consistent) factual descriptions would be more rational than seeing them under one overarching frame. However, the Due Diligence requirement implies that this more exhaustive specification should be preferred.

5.4 Conclusion

This chapter examined Bermúdez's account of frames as decision-theoretic representations, and more particularly the claim that one may value the same option or outcome differently under different frames. Like Schick, Bermúdez rejected the invariance principle and claimed that standard decision theory is insufficient to capture our intuitions about complex decision situations involving inconsistent evaluations of the same outcome or option under different frames.

The present account of representation as well as Bermúdez's account of frames contend that orthodox decision theory is over restrictive regarding framing effects, yet for different reasons. While Bermúdez deems the invariance principle illegitimate, the present account does not reject it but claims that the principle is not adequate for distinguishing rational from irrational framing effects. In addition, Bermúdez sets externalist constraints on beliefs when the present Humean account does not.

I have raised two issues with Bermúdez's account. First, Bermúdez's canonical example (Agamemnon) is not sufficient to reject the invariance principle nor to justify the introduction of ultra-intensional frames, consistently with chapter 2. Irrational framing effects can be defined independently from factual neutrality, as argued in chapter 4 where they are defined as inconsistent choices across evaluatively equivalent decision situations. I then turned to the main rationality constraint imposed by Bermúdez on frames, the evaluative neutrality of facts. I argued that the specification of factual propositions in a decision problem is not evaluatively neutral, which in turn violates the Due Diligence requirement. If this is correct, this pleads in favor of accounts of representation that acknowledge that all specifications of decision problems and their consequences are value-driven.

This dissertation gives an even greater role to evaluative attitudes in representing decision problems than Schick and Bermúdez: indeed, any decision-theoretic consequence represents a feature of the decision situation that has been selected based on some evaluative difference that this feature makes for the agent; either directly through cares, or indirectly through instrumental plans grounded on the agent's cares. This makes the decision-theoretic representations of this account partial, constructed, and value-driven though it does not violate the invariance principle.

References

Arrow, Kenneth J. (1982). "Risk Perception in Psychology and Economics". In: *Economic Inquiry* 20.1, pp. 1–9.

Bermúdez, José Luis (2009). *Decision Theory and Rationality*. Oxford University Press.

Bermúdez, José Luis (2020). *Frame It Again: New Tools for Rational Decision-Making*. Cambridge University Press.

Bernstein, Amit et al. (2015). "Decentering and Related Constructs: A Critical Review and Metacognitive Processes Model". In: *Perspectives on Psychological Science* 10.5, pp. 599–617.

Broome, John (1991). *Weighing Goods: Equality, Uncertainty and Time*. Wiley-Blackwell.

Fisher, Roger and William Ury (1981). *Getting to Yes: Negotiating Agreement Without Giving in*. Houghton Mifflin.

Goffman, Erving (1974). *Frame Analysis: An Essay on the Organization of Experience*. Harvard University Press.

Hume, David (1995). *A Treatise of Human Nature (1739-40)*. Past masters. InteLex Corporation.

Lakoff, George (2004). *Don't Think of an Elephant! : Know Your Values and Frame the Debate : the Essential Guide for Progressives*. Chelsea Green Publishing Company.

McKenzie, Craig R. M. and Jonathan 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.

Selman, Robert L. and Diane F. Byrne (1974). "A Structural-Developmental Analysis of Levels of Role Taking in Middle Childhood". In: *Child Development* 45.3, pp. 803–806.

Sher, Shlomi and Craig R. M. McKenzie (2006). "Information leakage from logically equivalent frames". In: *Cognition* 101.3, pp. 467–494.

Shoemaker, David W. (2003). "Caring, Identification, and Agency". In: *Ethics* 114.1, pp. 88–118.

Chapter 6

Conclusion

Contents

In this chapter, I first summarize the main conclusions of this thesis and provide an overview of its arguments. I then turn to the strengths and weaknesses of the account, and conclude by some open questions for further research.

6.1 Overview of the main conclusions

This dissertation attempted to explore the broad question of the rationality standards ruling decision theoretic representations, and of what constitutes a correct decision-theoretic representation from the perspective of practical rationality. I defined decision-theoretic representations as mappings from a decision situation, described in the vocabulary of philosophy of action, to the kind of formal decision problems discussed in decision theory made out of options, states of the world, and consequence.

Framing effects constituted the starting point of the investigation; I examined several philosophical issues raised by the normative assessment of this phenomenon. Among them, I discussed two principles that have been taken to rule against illegitimate framing effects in the literature, the Invariance or extensionality principle, and Invariabilism. The Invariance principle stipulates that two descriptions of the same decision situation (option, or consequence) should lead to the same evaluation and choice. By contrast, Invariabilism requires that a consequence or option be equally valued *across* decision situations. Invariabilism is one of many views on how sensitive our attitudes should be to the context of the decision situation, where a context is defined as the set of features of the decision situation whose presence may affect the evaluation of a consequence.

Both Invariance and Invariabilism raised the same two questions, which guided the discussion of the dissertation: *How stable should our representations of outcomes be? And how robust should our attitudes be to representational shifts?*

These questions led philosophers to discuss the theoretic neutrality of representations for decision theory. Indeed, since decision theory characterizes rational decisions in terms of beliefs and desires, one may claim that normative decision theoretic standards could be formulated without any reference to representation. In that sense representation would be theoretically neutral for practical rationality. On the contrary, one may claim that representational standards are not reducible to decision theoretic standards over beliefs and desires. To show that representations are not theoretically neutral, one should offer examples of situations where decision theory is either insufficient to determine the rational way of representing the decision situation, or where our intuitions about the correct way of representing the situation cannot be derived from decision theoretic principles. Consequently, settling one the issue of theoretic neutrality implies determining *what aspects of the activity of rational representation cannot or should not be characterized only by beliefs and desires*.

The fourth question raised by the discussion of framing effects is the *evaluative neutrality of rational representations*. Indeed, advocates of the Invariance principle argue that the way we represent a given decision situation should not affect how we evaluate it. Conversely, those rejecting the principle argue that representations have an incidence on practical evaluations: Schick and Bermúdez thus argued that a rational agent may value the same option differently under two different representations (understandings for Schick, frames for Bermúdez) of the decision situation. They thus offer counterexamples to the principle where our intuitions about the rational evaluation of a situation derive from our intuitions about how to frame the problem adequately.

This dissertation was driven by three commitments. Moderate Humeanism, the view that "*rationality only constrains our attitudes indirectly by disallowing certain combinations of beliefs, desires and preferences.*"[1] Second, this work intended to define standards of rationality for bounded creatures like us and thus discuss how bounded agents may rationally represent decision problems in a manner adequate to their goals and constraints. Finally, the account should convincingly explain how decision problems arise for bounded agents, and should account for "normal" as and "anomalous" representations and decisions and contribute to explain "why did the agent phi?" in the spirit of the "ought implies can" adage. Importantly, this dissertation was not intended to address epistemological or linguistic issues of representation, nor to offer action-guiding procedures of representation.

In chapter 2, I examined Schick's theory of understandings as decision theoretic representations and rejected it on two main grounds. First, the existence of rational conflicts does not make a compelling case for understandings. I then questioned the relationship between intensionality

[1] Bradley 2017, pp. 21–30.

and conflict, and contended that situations of conflicts can be modeled without violating weak extensionality by treating consequences as intensional properties of particular options. Schick's account nevertheless offered interesting insights on issues of representation of decision problems, and in particular on the relationship between self-delusion and representation.

I defended Schick's view that representations of decision problems are bound to be selective descriptions. If decision theory accepts that rational agents are not omniscient beings that take into account all possible descriptions of a decision situation, it is reasonable to assume that rational representations are selective. While Schick makes that claim in the case of propositional understandings of options, chapter 3 raised selection problems independently from the Invariance or extensionality principle.

Chapter 3 introduced two major selection issues that arise from what I called the problems of the scope and the granularity of decision problems. Indeed, a tension exists between on the one hand our incentives to isolate certain considerations when deliberating about a particular decision, and on the other the fact that our decisions are often connected through their consequences. As one ponders whether to phi or not, one may take into account additional considerations and options, in such a way that what was initially a simple decision question becomes a decision problem with unclear boundaries. Then, the decision situation may also relate to decision issues that were initially treated separately. While an omniscient decision-maker would probably not face the issue, an account of rationality for the bounded should provide guiding principles to isolate decision problems in a way that is both realistic and permissible.

I then advocated a two-stage view of representation. I first defended what I called the Caring Principle, that consequences are objects that make a differential impact which is relevant for the decision-maker if she cares about it. I offered a number of arguments in favor of adding cares to the standard model of beliefs and desires. First, desires cannot offer a criterion to define self-delusion. Moreover, cares are more persistent than desires; since they expose us to gains and losses and are motivating in normal cases, they are explanatorily more basic than desires. Thirdly, the fact that one satisfied one's desire does not imply that they cared about the object of desire, only that they cared about not having such a desire frustrated. If desires constitute the standard of instrumental rationality, the only way to address frustration is to satisfy the relevant desire; by contrast, cares don't disappear after being satisfied, and even persist when they are frustrated to the benefit of a greater care. One can thus rationally suppress a certain desire that is not endorsed by one's cares. This possibility matters particularly in the cases of deliberate unawareness where some feature of a prospect, or desire should be removed from the description of a decision problem. Finally, introducing cares allows us to simply *model motivation, control and akrasia in order to distinguish irrational representations stemming from impulse and from weakness of will*.

However, this first stage is not sufficient on its own to solve the problems of scope and of granularity. I thus introduced a second principle of representation which claims that the decision

problems we face are shaped by prior intentions and plans as Bratman (1999) defines them. Combined, these principles offer a plausible solution to the problems aforementioned, as well as to Resnik's "specification problem". Finally, I contended that an intention-based account of representation can only address the objections raised at irrational intentions if it is grounded on the agent's cares instead of desires.

This second principle allows the identification of another kind of irrational representation in addition to impulses and weakness of will: instrumental representations (either unreflective, policy-based, or deliberative) failing to satisfy the requirements of specification presented in subsection 3.3.4. Moreover, it allows for a distinction between external and internal standards of assessments of representations. The former are standards of rational criticizability, while only the latter constitute actual standards of rational blameworthiness. These requirements show how one could rationally filter options and set how finely grained states of the world should be under certain structural assumptions (that the space of options is (i) nested for grain (ii) measurable for scope).

How does this instrumental view of representation address the question of how stable our representations and evaluations should be? First, this view values representational stability to the extent that the agent's cares themselves are stable; since plans, policies, and decisions problems are instruments to satisfy these cares, the question of how stable representations should be is answered by the "mechanics" of commitment, deliberation and reconsideration. However, this part of the account does not address the question of how stable our cares as evaluative attitudes should be.

Chapter 4 discussed the stability of our evaluations in the face of changes in the decision situation. The value of an option can be highly sensitive to uncertain circumstances and unforeseen contingencies. Moreover, the environment (in the sense of the conditions over which the decision-maker has no control), of the decision situation may change, which may prompt the agent to reconsider. While the account of instrumental representation defended in chapter 3 interpreted such cases as cases of reconsideration of the specification of the decision situation, the following chapter explored the issues of context-sensitive instantiation of consequences as well as the sensitivity of their evaluations.

Chapter 4 investigated the idea that a decision situation can be "represented extensionally" by treating its relevant features as properties of the choice situation. Then, these relevant features also constitute the set of weighable reasons of the decision situation. That hypothesis allowed us to distinguish between properties associated with a particular option, with a set of them, and those attached to the situation itself. The context of the situation can then be meaningfully defined as the set of relevant properties of the decision situation, and two kinds of context-sensitivity should be distinguished. First, whether a property is instantiated or not can depend on the context of the decision situation. Second, the evaluation of those properties can itself be context-sensitive (what Dietrich and List (2016) respectively called context-relatedness

and context-variance). In this light, the rationality requirements bearing on the instantiation and the evaluation of options as bundles of represented properties offer grounds for excluding certain ways of representing options. Applied to framing effects, the evaluative requirements raised the questions of (i) when the same consequence should be valued equally across contexts, and (ii) when equally valued consequences should play the same evaluative role across contexts.

Chapter 4 answered these two questions by arguing that if the set of relevant properties are reasons counting in favor of an option of a particular option pair, rational context-sensitivity of properties and their weights ought to satisfy the requirements of the practical rationality of reasons. I thus turned to the literature on reasons to examine how reasons and their weights ought to combine. I introduced five extreme views about reasons (Invariabilism, atomism, holism, generalism and particularism) and rejected them, both on general grounds as well as reasons that are specific to bounded agency. These views thus cannot offer a criteria of demarcation for framing effects. Nevertheless, the discussion about the way we compose reasons and their weights can be used to determine when the instantiation and the evaluation of properties are legitimately context sensitive, via the notion of general and particular differences. I defended a principle ruling out illegitimate framing effects as inconsistent choices in evaluatively equivalent decision situations. I defined equivalent decision situations in terms of general and particular differences the agent cares about, consistently with the definition of decision-theoretic consequences offered in chapter 3. In that light, I pleaded for distinguishing irrationality claims made in three cases: in the lab via a formal description of a decision problem, in the lab via a practical decision situation, and in real-world situations.

I illustrated how this view of framing effects and of rational context-sensitive evaluations can be formalized more precisely with Sher's marginalist model of composition of reasons. Sher's framework[2] shows how one may aggregate and weight reasons and can be interpreted as a theory of context-sensitive evaluations of decision-theoretic consequences. Such a theory would endorse a marginalist view of evaluative context-sensitivity that is not central to my account; yet it has the merit of being a non-trivial theory that rejects the five extreme views about reasons by striking a middle ground between them.

Agents are rationally required to value a reason identically across contexts they deem equivalent with respect to all other reasons; short of that, they may be blamed for inconsistent evaluation or for akratically responding to these evaluations. Since Sher's framework takes as given the set of relevant reasons of a situation, it can be coherently associated with the principle. Combined, they specify how the way we represent decision situations ought to cohere with our evaluative attitudes, across equivalent decision situations (through the principle), as well as across non-equivalent ones (through coherent evaluations, as in cases of double counting). Finally, Sher's model offers insight as to how prudential prescriptions may play a role in ruling out certain kinds of context dependence. If the agent endorses such a prescription, he will accept that the weight of a reason ("stealing is wrong", or S) should be independent of another putative

[2]Sher 2019.

reason ("reggae is being played in the background" or R), so that the conditional weight of S given R is equal the unconditional weight of S. In the Asian disease case, a prescription possibly endorsed by the agent would be that the difference between saving and not letting die should not affect the moral value of options. Again, this kind of prescription is not rationally binding unless the agent herself endorses it.

I conclude that our normative intuitions about framing effects can be meaningfully formulated without resorting to a notion of representation that violates the Invariance principle. However, this does not prove that Invariance violation is not necessary to account for our other normative intuitions about practical rationality. In chapter 5, I examined Bermúdez's account of representations as frames[3] and of framing effects. Like Schick, Bermúdez rejects the Invariance principle and claims that standard decision theory is insufficient to capture our intuitions about complex decision situations involving inconsistent evaluations of the same outcome or option under different frames. Yet his rationality requirements on frames differ from Schick's. The general intuition that guides Bermúdez's demonstration is that the rigid consistency standards of traditional decision theory are inadequate when the complexity of real-world situations demands a richer or deeper understanding.

After presenting Bermúdez's account of frames, I examined how it differs from the account of representation defended by this dissertation. The present account of representation as well as Bermúdez's account of frames contend that orthodox decision theory is over restrictive regarding framing effects, yet for different reasons. While Bermúdez deems the Invariance principle illegitimate, the present account does not reject it but claims that the principle is not adequate for distinguishing rational from irrational framing effects. In addition, Bermúdez sets externalist constraints on beliefs when the present Humean account does not.

I have raised two issues with Bermúdez's account. First, Bermúdez's canonical example (Agamemnon) is not sufficient to reject the Invariance principle nor to justify the introduction of ultra-intensional frames, consistently with chapter 2. Irrational framing effects can be defined independently from factual neutrality, as argued in chapter 4 where they are defined as inconsistent choices across evaluatively equivalent decision situations. I then turned to the main rationality constraint imposed by Bermúdez on frames, the evaluative neutrality of facts. I argued that the specification of factual propositions in a decision problem is not evaluatively neutral, which in turn violates the Due Diligence requirement. If this is correct, this pleads in favor of accounts of representation that acknowledge that all specifications of decision problems and their consequences are value-driven.

This dissertation gives an even greater role to evaluative attitudes in representing decision problems than Schick and Bermúdez: indeed, any decision-theoretic consequence represents a feature of the decision situation that has been selected based on some evaluative difference that this feature makes for the agent; either directly through cares, or indirectly through instrumental

[3] Bermúdez 2020.

plans grounded on the agent's cares. This makes the decision-theoretic representations of this account partial, constructed, and value-driven though it does not violate the Invariance principle. Table 6.1 on p. 140 summarizes the main claims of my account.

Let's take stock. Upon asking what counts as a rational representation of a decision situation, we have seen that framing effects raised several specific philosophical issues. First, how stable should our representations of outcomes be? And then, how robust should our attitudes be to such representational shifts? Furthermore, two kinds of representational neutrality were examined. Can the way we represent a decision situation affect our evaluations of the decision problem? Last, are decision problems theoretically neutral for decision theory?

This dissertation argued that rational representations for the bounded are not theoretically neutral for decision theory: the way we represent a decision situation cannot be reduced to the standard model of beliefs and desires. For different reasons, cares and future-intentions can shape our representations in ways that beliefs and desires alone would not. Decision problems are not evaluatively neutral representations either, as they involve the selection of a set of relevant features. Instrumental representations address issues of selection by making explicit the relationship between our evaluative attitudes (cares and intentions) and the choice of a particular specification of the decision situation.

To the question of the stability of our representations, this thesis answered in several steps. It first concluded that rational conflict is not sufficient to justify a notion of representation accommodating several alternating point of views on the same decision situation. Yet a particular act, state, or outcome can still be included in more than one decision situation: this is what made the issues of scope and grain so problematic. However the account of instrumental representations defended in chapter 3 guarantees that in all three cases of unreflective, policy-based and deliberate representations, the rational agent settles on a single representation of the particular situation, shaped by her plans and cares. In that sense the instrumental part of the account gives a pragmatic rationale for making our representations stable. The part of the account defended in chapter 4 examined the stability of our cares as evaluations that may be sensitive to changes in the decision situation. It addresses the issue in two ways: first, it defines evaluatively equivalent decision situations in terms of general and particular differences that the agent cares about, and determines when a feature should be instantiated in two situations. Two decision situations are evaluatively equivalent if and only they have identical general features and do not differ in any particular way the agent cares about. Last, in non evaluatively equivalent cases, it does not make any specific requirements on evaluations.

In the next section, I discuss the strength and weakness of the account, and conclude by the questions it raises for future investigation.

Table 6.1. Care-based instrumental representations

Structural requirements:

- Caring Principle: decision theoretic consequences are features making a difference the agent cares about.
- These features can be represented as intensional properties of the alternatives.
- Care-desire model of agency: agents connect to their cares through deliberation.
- These consequences can be treated as weighable properties.

Consistency requirements:

- No contradictory beliefs requirement for all propositions the agent is aware of.
- Rationally intended ends and their reconsideration should be determined by cares.
- Coherent beliefs, desires and intentional plans as per Bratman's theory (see tables 3.2 and 3.3 on pp. 61–62).
- Coherent specifications of instrumental decision problems (see tables 3.4 and 3.5 on pp. 65–66).
- Consistent evaluations of general and particular differences as weighed reasons in favor of a an alternative.

All rationality requirements distinguish:

(i) external and internal point of view of assessment
(ii) rational criticizability and rational blameworthiness
(iii) unreflective, policy-based, and deliberative specifications of the decision problem

Irrational framing effects:

(Principle of identification of evaluatively equivalent decision situations)
Suppose that:

(a) Each of the decision situations can be fully characterized by general differences between options, and these differences are the same in both situations.
(b) The two decision situations do not differ in any particular way the agent cares about, so that these general differences equally apply to both situations.

Then, the two choice situation ought to be treated equivalently, and only frames yielding equivalent choices across situations will be rationally admissible.

6.2 Strengths, weaknesses, and open questions

In the Moderate Humean tradition, this account has the appeal of setting minimal, non-trivial requirements of practical rationality on decision theoretic representations. Grounding representations on cares allowed us to separate objective considerations from subjective ones when it comes to assessing agents' representations. What people ought to believe and care about are prudential matters that do not pertain to practical rationality nor representational requirements. In the case of self delusion, it permits to identify internal requirements deriving from consistency between various attitudes, while keeping the issue of their external validity separate.

The view defended by this account can be formalized precisely under further structural assumptions, and accounts for our intuitions in a number of problems of representation for the bounded. By distinguishing representational, epistemic and evaluative issues, it clarifies the rationality requirements ruling framing effects. In complex decision problems like chess, it gives a plausible solution to the problems of scope and grain of the player's representation of the game. In iterative decision situations (like the lunch meeting and the abduction cases, see subsection 3.3.6), it explains how one may sequentially extend the scope and refine the grain of the decision problem under limited agential resources.

In addition, the distinction between external and internal standards of representation inspired by Bratman (see subsection 3.3.3) is key to the justification of an internalist approach of representation that still distinguishes different standpoints of assessment of representations. These standards remain internalist in that they refrain from ruling on the external validity of our representations with respect to some reality that would constitute the domain of representation. Finally, these standards do not involve philosophical assumptions about mental, linguistic, or metaphysical representation.

I believe the account is compatible with various philosophical views on practical rationality: it encompasses consequentialist and deontic views on consequence, and accommodates Aristotelian definitions of options as teleological acts valued for some further ends and as procedural activities that are ends in themselves.[4,5] The thesis shares Schick's and Bermúdez's concerns with the litany of rationality in decision theory; although it endorses Invariance, this answer is only tentative: the account holds until new counterexamples can be found against the principle.

One of the main weaknesses of this work is that it does not determine the scope covered by the account? Are there non-instrumental decision-theoretic representations that can still be endorsed by a moderate Humean view of rationality? Second, this thesis did not offer a systematic analysis of cares for decision theory. Instead, it presupposed a minimalistic view of cares, that implies evaluations, weighing and preferences when these cares are commensurable; their consistency is guaranteed by the Invariance principle and the logics of instrumental planning. As with

[4] Aristotle 1989, Book 9 (Theta), [1048b].
[5] Seibt 2023.

desires, the instrumental value of intended objects of cares is transmissible from ends to means. The account does not need more to consistently justify its conclusions, yet I believe that a more thorough conceptual analysis of cares would be very fertile: in particular, the question of how one comes to discover one's cares is left open. Additionally, the relationship between cares and time deserves further scrutiny. I suspect that cares are the attitudes that ground our subjective and internal evaluations of time. If successful, this analysis would give further weight to the notion of cares for decision theory, and would be able to assess the rationality of phenomena like present-bias, or procrastination.

References

Aristotle (1989). *Aristotle's Metaphysics (Vols. 17–18)*. Trans. by Hugh Tredennick. Harvard University Press.

Bermúdez, José Luis (2020). *Frame It Again: New Tools for Rational Decision-Making*. Cambridge University Press.

Bradley, Richard (2017). *Decision Theory with a Human Face*. Cambridge University Press.

Bratman, Michael (1999). *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge University Press.

Dietrich, Franz and Christian List (2016). "Reason-Based Choice and Context-Dependence: An Explanatory Framework". In: *Economics & Philosophy* 32.2, pp. 175–229.

Seibt, Johanna (2023). "Process Philosophy". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2023. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2023/entries/process-philosophy/.

Sher, Itai (2019). "Comparative Value and the Weight of Reasons". In: *Economics & Philosophy* 35.1, pp. 103–158.

Bibliography

Aristotle (1989). *Aristotle's Metaphysics (Vols. 17–18)*. Trans. by Hugh Tredennick. Harvard University Press.

Arrow, Kenneth J. (1982). "Risk Perception in Psychology and Economics". In: *Economic Inquiry* 20.1, pp. 1–9.

Bermúdez, José Luis (2009). *Decision Theory and Rationality*. Oxford University Press.

Bermúdez, José Luis (2020). *Frame It Again: New Tools for Rational Decision-Making*. Cambridge University Press.

Bernstein, Amit et al. (2015). "Decentering and Related Constructs: A Critical Review and Metacognitive Processes Model". In: *Perspectives on Psychological Science* 10.5, pp. 599–617.

Bradley, Richard (2017). *Decision Theory with a Human Face*. Cambridge University Press.

Bratman, Michael (1987). *Intention, Plans, and Practical Reason*. Cambridge: Cambridge, MA: Harvard University Press.

Bratman, Michael (1999). *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge University Press.

Broome, John (1991). *Weighing Goods: Equality, Uncertainty and Time*. Wiley-Blackwell.

Broome, John (1993). "Can a Humean be moderate?" In: *Value, Welfare and Morality*. Ed. by R. G. Frey and Christopher W. Morris. Cambridge University Press, pp. 51–73.

Broome, John (1999). "Can a Humean be moderate?" In: *Ethics out of Economics*. Cambridge University Press, pp. 68–88.

Brown, Campbell (2013). "The Composition of Reasons". In: *Synthese* 191.5, pp. 779–800.

Buchak, Lara (Nov. 2013). *Risk and Rationality*. Oxford University Press.

Comesaña, Juan and Peter Klein (2019). "Skepticism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2019. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2019/entries/skepticism/.

Dancy, Jonathan (2004). *Ethics Without Principles*. New York: Oxford University Press.

Daniels, Norman (2023). "Reflective Equilibrium". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2023. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/fall2023/entries/reflective-equilibrium/.

Dietrich, Franz and Christian List (2013). "A Reason-Based Theory of Rational Choice". In: *Noûs* 47.1, pp. 104–134.

Dietrich, Franz and Christian List (2016). "Reason-Based Choice and Context-Dependence: An Explanatory Framework". In: *Economics & Philosophy* 32.2, pp. 175–229.

Dreier, James (1996). "Rational Preference: Decision Theory as a Theory of Practical Rationality". In: *Theory and Decision* 40.3, pp. 249–276.

Fisher, Roger and William Ury (1981). *Getting to Yes: Negotiating Agreement Without Giving in*. Houghton Mifflin.

Frankfurt, Harry (1982). "The Importance of What We Care About". In: *Synthese* 53.2, pp. 257–272.

Frege, Gottlob (1892a). "On Sinn and Bedeutung". In: *The Frege Reader*. Ed. by Michael Beaney. Wiley-Blackwell, pp. 151–172.

Frege, Gottlob (1892b). "Über Sinn und Bedeutung". In: *Zeitschrift für Philosophie Und Philosophische Kritik* 100.1, pp. 25–50.

Frigg, Roman and James Nguyen (2021). "Scientific Representation". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2021. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2021/entries/scientific-representation/.

Gert, Bernard (1990). "Rationality, Human Nature, and Lists". In: *Ethics* 100.2, pp. 279–300.

Gigerenzer, Gerd (2000). *Adaptive Thinking: Rationality in the Real World*. Oxford University Press, USA.

Goffman, Erving (1974). *Frame Analysis: An Essay on the Organization of Experience*. Harvard University Press.

Goodman, Nelson (1955). *Fact, Fiction, and Forecast*. Harvard University Press.

Hume, David (1995). *A Treatise of Human Nature (1739-40)*. Past masters. InteLex Corporation.

Jacob, Pierre (2023). "Intentionality". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Spring 2023. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/spr2023/entries/intentionality/.

Joyce, James M. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press.

Lakoff, George (2004). *Don't Think of an Elephant! : Know Your Values and Frame the Debate : the Essential Guide for Progressives*. Chelsea Green Publishing Company.

Lewis, David (1983). *Philosophical Papers, Volume I*. Oxford University Press, USA.

McKeever, Sean and Michael Ridge (2006). *Principled Ethics: Generalism as a Regulative Ideal*. Oxford University Press.

McKenzie, Craig R. M. and Jonathan 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.

Merriam-Webster (2024). *Context*. In: *Merriam-Webster.com Dictionary*. URL: https://www.merriam-webster.com/dictionary/context (visited on 01/11/2024).

Neuman, John von and Oskar Morgenstern (1944). *Theory of Games and Economic Behavior*. Princeton University Press.

Orilia, Francesco and Michele Paolini Paoletti (2022). "Properties". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2022. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/spr2022/entries/properties/.

Orwell, George (1954). "Looking Back on the Spanish War". In: *A Collection of Essays*. Written in 1942. Garden City: Doubleday.

Pettit, Philip (1991). "Decision Theory and Folk Psychology". In: *Essays in the Foundations of Decision Theory*. Ed. by Michael Bacharach and Susan Hurley. Blackwell, pp. 147–175.

Pettit, Philip (2010). "Deliberation and Decision". In: *A Companion to the Philosophy of Action*. Ed. by Constantine Sandis and Timothy O'Connor. Blackwell, pp. 252–258.

Pinto, Robert C (2009). "Argumentation and the Force of Reasons". In: *Informal Logic* 29.3, pp. 268–295.

Quinn, Warren S. (1990). "The Puzzle of the Self-Torturer". In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 59.1, pp. 79–90.

Ramsey, Frank (1926). "Truth and Probability". In: *F. P. Ramsey: Philosophical Papers*. Ed. by David H. Mellor. Cambridge University Press, pp. 52–109.

Rawls, John (1951). "Outline of a Decision Procedure for Ethics". In: *The Philosophical Review* 60.2, pp. 177–197.

Rawls, John (1971). *A Theory of Justice*. Harvard University Press.

Resnik, Michael D. (1987). *Choices: An Introduction to Decision Theory*. University of Minnesota Press.

Rosner, Jennifer Amy (1998). "Reflective Evaluation, Autonomy, and Self-knowledge". PhD thesis. Stanford University.

Ross, David (1930). *The Right and the Good. Some Problems in Ethics*. Ed. by Philip Stratton-Lake. Clarendon Press.

Savage, Leonard J. (1954). *The Foundations of Statistics*. Wiley Publications in Statistics.

Schick, Frederic (1991). *Understanding Action: An Essay on Reasons*. Cambridge University Press.

Schick, Frederic (1997). *Making Choices: A Recasting of Decision Theory*. Cambridge University Press.

Schick, Frederic (2003). *Ambiguity and Logic*. Cambridge University Press.

Seibt, Johanna (2023). "Process Philosophy". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2023. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2023/entries/process-philosophy/.

Selman, Robert L. and Diane F. Byrne (1974). "A Structural-Developmental Analysis of Levels of Role Taking in Middle Childhood". In: *Child Development* 45.3, pp. 803–806.

Sen, Amartya (1993). "Internal Consistency of Choice". In: *Econometrica* 61, pp. 495–521.

Sher, Itai (2019). "Comparative Value and the Weight of Reasons". In: *Economics & Philosophy* 35.1, pp. 103–158.

Sher, Shlomi and Craig R. M. McKenzie (2006). "Information leakage from logically equivalent frames". In: *Cognition* 101.3, pp. 467–494.

Shoemaker, David W. (2003). "Caring, Identification, and Agency". In: *Ethics* 114.1, pp. 88–118.

Simons, Daniel J. and Christopher F. Chabris (1999). "Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events". In: *Perception* 28.9, pp. 1059–1074.

Thoma, Johanna (2019). "Decision Theory". In: *The Open Handbook of Formal Epistemology*. Ed. by Richard Pettigrew and Jonathan Weisberg. PhilPapers Foundation, pp. 57–106.

Tversky, Amos and Daniel Kahneman (1981). "The Framing of Decisions and the Psychology of Choice". In: *Science* 211.4481, pp. 1–30.

Tversky, Amos and Daniel Kahneman (1986). "Rational Choice and the Framing of Decisions". In: *The Journal of Business* 59.4, S251–S278. (Visited on 01/11/2024).

Tversky, Amos and Daniel Kahneman (Nov. 1991). "Loss Aversion in Riskless Choice: A Reference-Dependent Model". In: *The Quarterly Journal of Economics* 106.4, pp. 1039–1061.

Watson, Gary (1975). "Free Agency". In: *The Journal of Philosophy* 72.8, pp. 205–220.