

Dear Reader,

Thanks for your interest in this beta version of chapter 3 and 4. While reading, please keep in mind the following:

- This is a [beta release](#): it is meant to test the ideas and improve its readability. English grammar hasn't been corrected either.
- It is written as a book chapter, with many references to material in Chapter 2, which can be [downloaded online](#). The reader who wants a shorter introduction to the kind of models we have in mind can look [at Section 3.1 in this paper](#).
- We intend to make a paper version of these two chapters. Comments and suggestion regarding the required background material are most welcome.

Given the early stage of the chapter, please do not quote it without our permission.

Enjoy reading!

Olivier Roy and Eric Pacuit, Munich and Tilburg, April 2011.

Chapter 3

Interactive Rationality and the Dynamics of Reasons

Part 1: Reasons and Normative Facts

Release beta 1

In this chapter and the next we turn to the notion of *interactive* rationality. Rationality has been extensively scrutinized in philosophy of action and meta-ethics in recent years.¹ Most of these discussion, however, have not considered in details the implications of social interaction for the conception of rational agency. Does interaction bring in rational requirements of their own? Is *interactive* rationality something more than individual rationality? In this chapter and the coming ones we argue that it is.

The foundation of our argument is a reformulation of contemporary epistemic game theory in terms that are familiar to philosophy of action and meta-ethics: normativity and reasons. This is the main task we set ourselves to in this chapter. We will consider *choice rules* and *solution concepts* put forward in the decision- and game-theoretical literature, and show that they can be seen as potential “normative sources”, issuing ought statements and singling out what facts count as reasons for action in interactive situations. The next chapter we analyze *responsiveness* to such reasons.

We do not propose a general theory of rationality, normativity and reasons, though. Our goal is to lay the ground, within contemporary epistemic game theory, for the development of such a theory. This chapter thus contains no lengthy philosophical discussion of what are rationality, reasons and ought statements. Our methodology is rather to pick some key elements in the extensive philosophical literature on the topic, bring them in contact with epistemic game theory, and extracts the specific, interactive elements that arise from the encounter. This is a first, modest step, although in our view a conceptually important one, towards

¹See [22, 53] for such surveys.

the development of a full-fledged theory of interactive rationality.

We will be focusing on two choice rules, namely dominance and admissibility, and on the related notions of best response and Nash Equilibrium. We will see that admissibility is of particular interest, both from a normative and from an analytical point of view. It thus deserves a closer analysis, which we provide in Chapter 5. Team- or group-oriented notions of interactive rationality can also be analyzed using the tools we provide in this chapter. They raise questions of their own, though, and so we postpone their discussion to later in this book. For now our goal is to set the stage, to provide a broad outlook on notions of interactive rationality in terms of ought statement and reasons.

3.1 Meta-ethical Background

We start with a brief introduction to key philosophical notions that we use in this chapter: *normative facts*, *ought statements* and *reasons*. We do not survey the extensive philosophical discussion of these notions, nor do we pretend to do justice to the full, substantive meta-ethical theories we draw from. We simply pick the minimal set of conceptual tools that we need to lay the ground for a general, epistemic analysis of interactive rationality. Many substantial theories of rationality are compatible with this minimal set, and a full-fledged normative theory of interactive rationality will of course require filling in the philosophical details. But this is not the task we set ourselves to in this chapter.

3.1.1 Normative facts and ought statements

The basis ingredients of our analysis are the notions of *normative facts* and *ought statements*. By ought statements we mean statements of the following form:

- Agent i ought to play strategy s_i , written $O_i s_i$;
- Agent i ought not to play strategy s_i , written $O_i \neg s_i$.

Ought statements thus prescribe to agents actions to take, or not. Here we are concerned with what J. Broome [17] calls *owned* oughts: statements that prescribe what *agents* ought to do, not what ought to be done or ought to be the case abstractly. We take normative facts to consist in the obtaining of ought statements. When an agent ought (not) to take a certain action in a given context we say that the corresponding normative fact obtains.

Our use of the word “fact” suggests some form of moral realism [43], but the present approach is not committed to this. All we need to proceed is the minimal assumption that in certain circumstances it is the case that agents ought, or not, to take certain action.² We do not make any assumptions as to how they obtain,

²The present approach is thus *not* compatible with radical forms of error theories about the normative. C.f. references in [50], for the moral case.

that is, what makes a given ought statement “hold”, “valid” or “true”. We simply take ought statements and normative facts are primitive, leaving room for many substantial account of how they come about.

We make ought statements and normative facts relative to the state of information disclosure in which they are issued. We thus distinguish *ex interim* from *ex post* normative facts. *Ex interim* normative facts are relative to the agent’s partial information state in a given context. *Ex post* normative facts refer to what agents ought to do [or have done], given what the others *in fact* are doing [or did].³ Consider the game in Figure 3.2 (page 6), and suppose that Ann mistakenly believes that Bob plays *R*. Given her partial information about Bob, there is a sense—to be made explicit below—in which Ann ought to play *B*: this is her best response given her beliefs. We call this ought statement an *ex interim* ought. But there is another sense in which Ann rather ought to play *T*: it is Ann’s best response to what Bob is *in fact* doing. One might call this second sense of ought “objective”, because it refers to how things in fact are, independently for Ann’s beliefs. We prefer to call it *ex post* because, as we shall see, such ought statements can be given a natural reading in terms of what ought to be done when all strategic uncertainty is resolved.⁴

Borrowing again a terminology from [17], we call a *normative source* something that issues or generates normative facts in a given context.⁵ Morality is the paradigmatic example of a normative source. Other sources are conceivable, though: rationality is an obvious, although rather contested candidate.⁶

Here we are concerned with *potential* normative sources; potential “normative system that generates” [46, p.51] ought statements. We say “potential” because we leave it open whether the sources we identify are genuinely normative. All that matter for the present is that one can generate a list of ought statements from these sources. Whether these ought truly put normative pressure on the agents is a question for the full-fledged inquiry that should be taken up on the basis of the present work.

³Thanks to Wlodek Rabinowicz for pointing us the importance of distinguishing these two senses of “oughts”.

⁴What count as reasons for action given *ex interim* and *ex post* reminds of, but seems to cut across a number of important distinctions made in philosophical literature. More on this in the next section.

⁵Broome rather talks about sources of requirement, because for him not all sources are normative - not all sources of requirement issue normative oughts. Here we are concerned with potential normative sources, and so we do not need this generality.

⁶The canonical reference is [31], but see also [53] for a broader survey.

3.1.2 Reasons

We take reasons in to be *facts*.⁷ In the mathematical models we introduce later, we simply identify facts with propositions, i.e. sets of states, ignoring the metaphysical complications that this might create.⁸

We are interested in *normative* reasons, facts that explain the obtaining of ought statements [17]. We do not go into details about what this explaining relation is. We content ourselves with the following intuitive illustration: Faced with, say, the fact that Ann ought to play B , it is legitimate to ask “why is that so?”. Facts that provide a satisfactory answers to such a question count as normative reasons. An other way to look at normative reasons is to say that they *justify* actions. In the words of M. Smith [46, p.95]: “To say that someone has a normative reason for [a certain action] is to say that [...] her [acting] is justified from the perspective of the normative system that generates [a specific normative] requirement.” In what follow choice rules and the ought statements they generate play the role of these “normative systems”, and the reasons they pinpoint justify the taking, or not, of certain actions.

The distinction we made above between *ex interim* and *ex post* normative facts echoes in the kind of reasons we will consider. Reasons explaining *ex interim* normative facts will refer to the agent’s (state of partial) information—here together with her preferences.⁹ To come back to our example, the fact that (Ann believes that) Bob will play R is Ann’s reason why she ought, *ex interim*, to play B , it’s Ann’s reason *for her* to play R . Reasons explaining *ex post* normative facts rather refer to what the others are actually playing—the fact that Bob is playing L is the reason why Ann ought, in fact, to play T .

This distinction between reasons explaining *ex interim* and *ex post* normative facts come close, but seems different than Willams’ [54] famous distinction between internal and external reasons. The particular cases of *ex post* reasons that we will consider in Section 3.4 are not independent of the agent’s “motivational set” [37]: they depends on the agent’s preferences. A closer distinction is Schroeder’s [44] one between subjective and objective reasons for action. He defines the former as believed facts which, were they true, would count as objective reasons for or against playing certain action. This is certainly the case in our running example. If Ann’s beliefs were true then Bob’s playing R would be an *ex post* reason for her to play L . In the general case, however, Schroeder’s distinction and ours seem to diverge. Beliefs, even true ones, might not be informative enough to zoom in precisely on the a *unique* combination choices of the others,

⁷This is not an uncontroversial view, as many authors rather argue that reasons are properties, cf. [37]. We leave this controversy aside here.

⁸See [32] for details.

⁹To anticipate a point we will make later, this reference to preferences is only accidental to the choice rules we consider. It is not an essential features of what we call *ex interim* reasons, nor of *ex post* reasons, for that matter.

the later being necessary to determine what agents ought to have done, *ex post*. More on this in the next chapter, Section 4.4.1.

Whether *ex interim* or *ex post*, the reasons we consider here are *owned* reasons, by virtue of the fact that they explain owned oughts. They are reasons *for an agent* to play or not a certain strategy. Again, one might investigate whether the further distinction between owned, *ex interim* and owned, *ex post* reasons matches Nagel's/Parfit's [37] famous distinction between agent-relative and agent-neutral reasons. This is not clear to us right now, and not crucial for the present investigation, and so we leave this open.

In what follows we work with *conclusive*, also called *all-out* reasons for action, as opposed to mere *pro tanto* reasons. This is a simplifying assumption. The models we use are not geared towards the process of *weighing* reasons.¹⁰ The reasons that choice rules will pinpoint are binary, so to speak. Either the agent has a reason for (or against) taking certain action, or not. To model the weighing of reasons would require a more graded approach to reasons in games, which we leave for future work.¹¹ Taking the reasons pinpointed by choice rules to be conclusive also means that we leave aside cases where *different* choice rules generate conflicting recommendations for an agent. To anticipate on a point we made earlier, we assume that agents recognize one, and only one choice rule as valid normative source. Normative conflicts are of great interest, but we leave them for follow-up work.

Before moving further, it is worthwhile to reiterate a point we made about normative facts: the present approach is not committed to a specific view on how the normative character of ought statements and their associated reasons is to be derived from the apparently non-normative notion of a satisfaction of a choice rules. This is an important issue, much debated in recent years [22, 53], but for the present purpose we only need to acknowledge that choice rules are potential normative sources, and so we do not need to take side in the debate. As we shall see, some choice rules, e.g. admissibility, lend themselves quite naturally to Humean interpretations, which in turn can be given a reductionist rider [44]. But this needs not be so, and our account is compatible with both options.

3.2 Classical and Epistemic Game Theory

Contemporary epistemic game theory [3, 10] provides a natural testing ground for bringing the philosophical concepts that we just presented to contexts of social

¹⁰For work in that direction, see [21].

¹¹Schroeder [44] has an interesting proposal of how to think weighing without explicitly introducing weights. Instead, he rely on the idea that agents have higher-order reasons, i.e. reasons to rank some reasons higher than other on some qualitative scale, and use this idea to define recursively (but informally!) a weighing order on reasons.

interaction. The epistemic view on game-theoretical rationality differs, however, from the one one finds in the classical theory of games. In this section we explain what this difference is, mostly informally.

The constituents of a *game*, from the classical point of view, are: *agents*, their possible *strategies*, and their *preferences* over possible *outcomes*, often simply taken to be strategy profiles. Table 3.2, a *coordination game*, is a typical example of a game in *strategic form*.¹² There are two players, Ann and Bob, each have two strategies, *T* and *B* for Ann and *L* and *R* for Bob, and their preferences are represented in by assignments of utility values on the strategy profiles.

		Bob	
		<i>L</i>	<i>R</i>
Ann	<i>T</i>	1, 1	0, 0
	<i>B</i>	0, 0	1, 1

Figure 3.1: A coordination game

Classical game theory investigated the notion of interactive rationality by developing so-called *solution concepts*. Abstractly, a solution concept for a given game is a set of strategy profiles, profiles that one would intuitively expect to occur at the upshot of rational interaction. Nash equilibrium and iterated elimination of strictly dominated strategies are two paradigmatic examples of solution concepts. In Table 3.2, for instance, no strategy is strictly dominated, that is all strategies can be seen as “rational” in terms of strict dominance, but only two of the four strategy profiles are Nash equilibria: only (T, L) and (B, R) are “rational” if interactive rationality is taken to be embodied by Nash equilibria.

Epistemic game theory [3, 11, 10] holds that games are not enough to understand interactive rationality; one must also consider the specific *informational contexts* in which the agents make their decision. Informational contexts specify the *strategic* and *higher-order uncertainty* that each agent face: her partial information about, respectively, what the others will do and the information they have.¹³ One simple example of informational context for the game in Table 3.2 is shown in the *epistemic-plausibility model*¹⁴ of Figure 3.2. At the actual state TL , Ann considers it strictly more plausible that Bob plays *R*—this is represented by the dashed arrow. We then say that Ann *believes*, here mistakenly, that Bob

¹²In this chapter we will mostly use example from games in strategic form. Our argument, however, is fully general: it carries over to extensive form rationality.

¹³Informational contexts of games, i.e. situations of strategic and higher-order uncertainty, should not be confused with possible situations of imperfect information in extensive form games. The latter describe structural uncertainty built in the game itself: lack of information about what other players have done in previous move. Strategic and higher-order uncertainty, on the other hand, concerns the *expectations* of the agents about what other *will* or *are* doing and expecting. Strategic and higher-order uncertainty can arise in both games of perfect and imperfect information. Thanks to Martin van Hees for urging us to clarify this point.

¹⁴See precise definition in Chapter 2.

plays R . She also believes, this time correctly, that Bob believes that she plays B . The same holds, *vice versa*, for Bob: he mistakenly believes that Ann plays B , but correctly believes that she believes that he is playing R .¹⁵

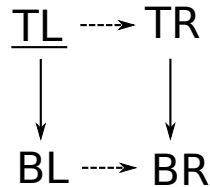


Figure 3.2: A specific informational context for the game in Table 3.2.

The epistemic view on game-theoretic rationality is that to choose rationally in a specific informational context is to choose what is best, or what one ought to choose, *given one's information*. It brings back the theory of rational decision making in games to its decision-theoretic roots. The focus is on rational decision of *individual* decision makers, in specific informational contexts of interaction. These decisions are assessed on the basis of decision-theoretic choice rules, our main object of investigation in Section 3.3. In our example above, at state BR both Ann and Bob have correct believe about each other's strategy choice. Given these beliefs, Ann's and Bob's strategy choice, B and R respectively, are the only strategies that are not dominated. They are *rational* choices, taking dominance as criterion for individual rationality or, in the terminology of the following sections, Ann nor Bob *ought not, ex interim*, to have played otherwise.

The broadly decision-theoretic stance of contemporary epistemic game-theory does not entail that interactive rationality is nothing more than rational decision making in the face of uncertainty.¹⁶ The informational context of an interactive situation is different then when the consequences of one's decision only depend on the state of a non-purposive environment. The former, but not the latter, includes higher-order information and this leads to drastically different outcome in game playing situations.¹⁷

Being decision-theoretic in character does not entail a *Bayesian* perspective either. By Bayesian we mean taking agents to have (graded, conditional) beliefs that satisfy the axioms of probability theory, real-valued preferences and making decisions as maximizers of expected utility. True, most of the current literature in epistemic game theory is taking this route [10]. But is by no mean a condition

¹⁵For probabilistic version of a very similar example see [10, 468-469]. For a detailed introduction to epistemic plausibility models, type spaces, and information contexts in general, together with discussion of the philosophical issues they raise, see Chapter 2 and [refs forthcoming SEP entry].

¹⁶Contrast with Kadane and Larkey [30].

¹⁷See for instance [40] for a dramatic example. We come back to this point in the next chapter, Section 4.5.

sine qua non to the epistemic approach to games and, indeed, in recent years a number of authors have done epistemic analysis with alternative choice rules, for instance admissibility [14] or “knowledge-based rationality” [2]. Here we follow this trend, by considering probabilistic as well as non-probabilistic epistemic attitudes, ordinal preferences, and different choice rules, all of this to be varied according to the substantive normative theory one is willing to endorse.

The epistemic view thus constitutes an important departure from the classical view based on solution concepts. The later are no more the basic objects of investigation. If at all, they show up as results of individual decisions in specific informational contexts. In the example above, at state BR , Ann’s and Bob’s not choosing a dominated strategy given their beliefs leads to Nash equilibrium. But this is not the case in general: state TL provides a clear example.

These two views, classical and epistemic, will be important for what follows. They embody two different types of recommendations or normative assessments in games, what we called *ex interim* and *ex post* above. Looking back at our example, at state TL it seems that neither Ann nor Bob does what her/his reasons require: Ann believes that Bob plays R , and this gives her a [conclusive] reason to play B , and vice-versa for Bob. Neither Ann nor Bob is “rational”, taking dominance, from an epistemic point of view. On the other hand, given what they are in fact playing, and independently of their information, it seems that they actually do what they ought to do: Ann’s plays her the best response given Bob’s choice, and vice versa. TL is indeed a Nash Equilibrium.

The basic ingredients of the theory of interactive rationality, from the perspective of classical game theory, are thus games and solution concepts. Epistemic game theory add informational contexts to the picture, and substitute solution concepts for individual choice rules. Both solution concepts and choice rule determine, in their respective, different sense, what ought, or not, to be done. Or at least this is the view we are taking in the next sections. We start with the epistemic view, studying *ex interim* ought statements (Section 3.3), and then go back to the classical view and its *ex post* perspective (Section 3.4). Our reason for starting with epistemics is that, as we shall see, it can also provide a natural account of the *ex post* perspective, in terms of full information disclosure.

3.3 From Choice rule to *ex interim* oughts

In this section we focus on what we call *ex interim* ought statements: what ought to be done given the agent’s state of partial information. This is the sense in which Ann ought to play B at state TL in the example above, because this what seems best to her given what she expects Bob to do. The view we take here is that choice rules are natural candidates for issuing such ought statements in specific informational contexts. They are potential normative sources. We thus start by introducing these choice rules, then show how to elicit *ex interim* ought

statements from them, and the kind of reasons for action this view pinpoints.

3.3.1 Choice Rules

For illustrative purposes, we focus on two choice rules: dominance and admissibility. These are of course not the only possible choice rules. We chose them because of their importance in contemporary game theory and philosophy of rational decision making. The kind of analysis that we now provide can and, we think, should be extended to other choice rules, for instance maximin, Minmax Regret [24] and some of the so-called fast and frugal heuristics [23].

Dominance

Informally, dominance it prescribe not to choose strategies which leads to strictly less preferred outcomes than another one in all circumstances that one considers possible. Formally, being a dominated strategy defined as follows, in epistemic-plausibility models.

3.3.1. DEFINITION. [Dominance] Let \mathcal{M}, w be a pointed epistemic-plausibility model. Then agent i 's strategy s_i is *dominated* at \mathbb{M}, w , written $\mathcal{M}, w \models \text{Dom}(s_i)$, iff there is a $s'_i \in S_i$ such that:

- for all wR_iw' : $(s'_i, f_{-i}(w')) >_i (s_i, f_{-i}(w'))$.

We simply take the choice *rule* to prescribe not to choose strategies that are dominated.

Let us dwell on this definition for a few moments, for it features important aspects of most decision-theoretic choice rules. The first thing to observe is that it has one placeholder, the relation R_i , which we will call the *epistemic* or *informational* parameter. This parameter, in turn, determines the range of options that are to be assessed according to the constraint specified by the choice rule. For dominance and admissibility, these constraints relate the facts that the agent believes/knows to be true with her preferences. For that reason we call these constraints the *preferential* parameter. This appellation should not be taken too literally, though, as noting precludes the constraints to be based on other features of the game than the agent's preferences over the outcomes.

The epistemic parameter makes rationality relative to the specific informational context of the game or, more specifically, to a specific informational attitude of the agents. In the example above, taking *beliefs* or *maximal plausibility* as the epistemic parameter in assessing whether, say, Ann's strategy choice is dominated at TL would mean checking whether the condition of the preferential parameter is met only at state TR , the state she considers most plausible. At TL Ann's choice is indeed dominated, given her beliefs. Many epistemic characterization results¹⁸ take beliefs as the value of the epistemic parameter, but this is of

¹⁸See for instance the remarks at the end of [10].

course not the only option. It is now common wisdom in epistemic game theory and epistemic logic¹⁹ that informational attitudes come in many forms, from hard attitudes, truthful and non-revisable, to softer ones, more or less revisable, and possibly mistaken. Stronger understandings of dominance, for instance, are also possible, by taking *knowledge* or *hard* information in the epistemic parameter.²⁰ In the example above one would have to check whether the preferential parameter is satisfied at both TR and TL . In this case Ann's choice is *not* dominated, given her hard information.

Fixing the epistemic parameter in one way or another has important consequences, both from an analytic and a normative point of view. First, it determines the sensitivity of the choice rule to changes in the context of the game. Different informational attitudes are more or less stable under information change, and the choice rules will inherit part of this stability. Setting the epistemic parameter is also of normative significance, as we saw in the previous example. Whether a strategy ought to be played or not will depend on the range of possibilities to be considered, which is in turn determined by the epistemic parameter. Stronger epistemic attitudes will make for stronger ought statements, i.e. harder to satisfy, and for stronger, i.e. more stable, reasons. In what follow we do not take stance on whether one choice of the epistemic parameter is more plausible from a normative point of view. Our investigation can rather be seen as an attempt to put into light the consequences, normative as well as analytic, of fixing the epistemic parameter in one way or another.

The preferential parameter determines what conditions, here in terms of preferences, should hold by the outcomes considered possible in order for one's choice to be considered (non-)dominated, (non-)admissible or, more generally, to be sanctioned or not by a given choice rule. In the case of dominance the condition is simply that, in each possible situation, the outcome of a given strategy is strictly worst than the outcome of another. This is the bullet point in Definition 3.3.1: for all $w' R_i w$, that is in all situations w' that one considers possible, modulo an instantiation of the epistemic parameter R_i , $(s_i, f_{-i}(w'))$, i.e. outcome of the game in w' if i plays s_i , is strictly worst for agent i than the outcome of the game if i would choose s'_i , *ceteris paribus*, which is precisely what the outcome $(s'_i, f_{-i}(w'))$ is. As mentioned already, by taking maximum plausibility as epistemic parameter strategy T or Ann is dominated at state TL : the condition of the preferential parameter for dominance are met at the state she considers most plausible, TR .

Just like for the epistemic parameter, fixing the preferential parameter has important consequences, both analytic and normative. On the analytic side, the very logical form of the preferential parameter influences the kind of changes in the

¹⁹See Chapter 2 for details.

²⁰Example of epistemic analysis with hard information can be found in Aumann [4], van Benthem [48]

context that a given choice rule will be sensitive to. We will see in Section 4.4 that the single universal quantification of dominance makes it stable under information refining. On the normative side, the preferential parameter can be interpreted in terms of substantive theories of the reasons. We will see in Section 3.3.3, for instance, that admissibility gets a natural Humean interpretation. Here, again, we will not tackle the normative question of whether one of these choice rules gives the correct theory of reasons in interaction. We see our analysis as a necessary preliminary to this normative investigation.

To recap, choice rules are articulated in terms of two parameters, epistemic and preferential. The epistemic parameter restricts the range of possibilities that are to be considered by the agent. To borrow from the epistemological terminology, it fixes the *relevant alternatives*.²¹ The preferential parameter, on the other hand, determines what condition should be met by these alternatives for the choice of an agent to be deemed rational or, from the point of view that we will introduce later, the preferential parameter fixes what *ought to be done* and, indirectly, what counts as reason for or against a given action.

Admissibility

The *admissibility* rule prescribes not to choose a strategy when there is another one which, in all circumstances that one considers possible, yields an outcome at least as good for the agent and, furthermore, when in one such circumstances it does strictly better. Formally:

3.3.2. DEFINITION. [Admissibility] Agent i 's strategy s_i is *not admissible* at \mathcal{M}, w , written $\mathcal{M}, w \models \neg \text{Adm}(s_i)$, iff there is a strategy $s'_i \in S_i$ such that:

1. for all wR_iw' : $(s_i, f_{-i}(w')) \leq_i (s'_i, f_{-i}(w'))$ and;
2. there exists wR_iw' s.t. $(s_i, f_{-i}(w')) <_i (s'_i, f_{-i}(w'))$.

We will consider admissibility in great details in the next chapter, so here only a few remarks are needed. The reader will observe first that the definition also allows the epistemic parameter to vary. Just like before, the dynamics properties of this choice rule will partly depend on how this parameter is fixed. The specificity of admissibility rather lies in the pattern of quantification in the preferential parameter. It quantifies both universally and existentially over what the agent considers possible. It first requires that there is a strategy different from the one actually played such that, *for all* outcomes considered possible, one would have been at least as well off by playing that strategy than playing the actual one, *ceteris paribus*. The second quantification requires that *there exists* at least one outcome in which playing that alternative strategy would have yielded a strictly better payoff, again all other things being equal. Putting these two conditions

²¹See [27] and references therein.

together, a strategy is not admissible given one's information if there is another one that yields a payoff at least as good as the first in all possible situations, without the two being payoff equivalent. As we will see in the coming chapters, it is this particular pattern of quantification that makes admissibility of particular interest, both from a normative and an analytic point of view.

3.3.2 *Ex interim* oughts

In the epistemic game theory literature choice rules are often presented as *definition* of individual rationality. The standard formulation of the epistemic characterization of iterated elimination of strictly dominated strategies (IESDS) [36, 8, 12], for instance, is that *rationality* and common belief in rationality implies IESDS. By contrast, when talking about these results we rather say that dominance and common belief in dominance implies IESDS.

This difference, although mainly terminological, nevertheless points towards a conception of choice rules as normative sources. One natural reading of the idea that such and such strategy is not rational in a given context is that the agents *should* not, or *ought* not to play that strategy in that context.²² This is the reading we explore now.

Since events of the form “such and such choice rule is satisfied at state w ” are well-defined in epistemic-plausibility models, we get a straightforward rendering of the corresponding *ex interim*, normative facts. Each choice rule induces its own set of corresponding ought statements, so we use superscripts to distinguish them:

- *Dominance*: An agent ought ^{D} not to play strategy s_i at state w , written $\mathcal{M}, w \models O_i^D \neg s_i$, iff $\mathcal{M}, w \models \text{Dom}(s_i)$.
- *Admissibility*: An agent ought ^{A} not to play a strategy s_i at state w , written $\mathcal{M}, w \models O_i^A \neg s_i$, iff $\mathcal{M}, w \models \neg \text{Adm}(s_i)$.

The first thing to notice about this definition is that these choice rules issue negative prescriptions. A dominated or a non-admissible strategy is one that one ought not to play. Consequently, the negation of these, i.e. not being dominated or being admissible, translates into *permissions*.²³ Taking dominance as

²²As mentioned earlier, the question whether rationality is normative, i.e. whether one ought to be rational, has been the object of much debate in philosophy. Philosophers and epistemic game theorist do not mean exactly the same things by “rationality”, though. The notion of rationality that is discussed in the philosophical literature is much broader than the (technical) one used in the epistemic game theory—see Section 4.1 below. As we will argue, a natural reading of the later is being responsive to one's believed reasons for action—only one aspect what it means to be rational in a broad sense. From that point of view the present proposal, to see choice rules as potential normative source, is much less controversial. Whether or not rationality in the broad sense is normative, it seems less debatable that being responsive to one's believed reasons is.

²³By taking “ i is permitted to” to be equivalent to $\neg O_i \neg$.

normative source, for instance, gives us that playing a non-dominated strategy is permitted, i.e. that it is not the case that the agent ought not to play it.²⁴

Our reason for thinking of choice rules as issuing negative prescriptions is twofold. First, being dominated or non-admissible are rather strong conditions. Second, there can be many non-dominated or admissible strategies. It seems odd to think that an agent ought to play all of them. In fact, by playing with negations one can check that, following these definitions, positive oughts, e.g. statements of the form $O_i^C s_i$ are true iff s_i is the *unique* dominating or admissible strategy. This is plausible.

One should also keep in mind the *ex interim* character of the ought statements that we extract from choice rules. They depend on specific informational contexts: for the same game it can very well be the case that an agent ought to play a certain strategy in a given context, say given dominance, but that the very same strategy turns out to be permitted in another context. This will come out even more clear in the next section, when we will turn to the reasons that are pinpointed by the above ought statements.

Of course, one could have extracted ought statements from choice rules in a different way. The definitions that follow are not meant as substantive theses regarding what should be done in contexts of interaction, or as the only plausible way of listing what ought (not) to be done given a particular choice rule. All that we need to proceed is that choice rules can be seen, in some way, as normative sources, issuing ought statements in specific contexts. The reason why we go from choice rules to normative facts in a specific way is simply that it appears natural.

3.3.3 Reasons, *ex interim*

The two determinant factors for the satisfaction of a choice rule, i.e. for the obtaining of a normative fact, are, first, the agent's information in a given context and, second, the satisfaction (or not) of the preferential parameter, within the boundaries fixed by the agent's information. In other words, what *explains* the obtaining of a normative fact, in the sense defined above, is a particular alignment of the agent's information and preferences.²⁵ Since we take an owned, conclusive, normative reasons to be a fact that explains why an agent ought (not) to take a certain action, these two factors seems like the natural locus of reasons in the present framework.

3.3.3. DEFINITION. Let w a state in a given epistemic model of a game, and C a choice rule with a given instantiation R_i of its epistemic parameter for agent i .

²⁴We are getting very close to the formulation of a deontic logic here. For the present, however, we do not investigate which deontic principles would satisfy the logics based on the respective choice rules that we study, let alone if these logic would be completely axiomatizable. We leave it for future work, but note the similarity with [28] and [47].

²⁵Again, the role played by preferences here is only due to the specificity of the two choice rules we consider here.

Write $R_i[w]$ for the set of state that are considered possible given R_i . Then $R_i[w]$ is a reason of agent i for [against] playing strategy s_i iff $\mathcal{M}, w \models O_i^C[\neg]s_i$.

This is a general definition, which we instantiate below for each set of normative facts that we defined above. For now a few general remarks are in order.

That we equate the existence of reasons with the obtaining of a normative fact is just an expression of the idea that we are dealing with conclusive reasons. Such reasons imply the existence of strong ought statements, and the other way around: *ex interim* normative facts regarding specific actions imply the existence of conclusive reasons.

This definition identifies the agent's *assumptions*, i.e. her strongest piece of information, with her conclusive reason for or against a certain action.²⁶ Allowing any superset of $R_i[w]$ to count as conclusive reason would induce a form of closure under logical implication. This would make all tautologies conclusive reasons. We do not want to commit to that. On the other hand, allowing reasons to be subsets of $R_i[w]$ would bring us closer to the idea of *pro tanto* reasons and would, furthermore, allow for conflicting reasons for action issuing from a single normative source. The latter seems like an undesirable consequence.²⁷ That leaves us with identifying $R_i[w]$, i.e. assumptions, with the agent's conclusive reason for or against certain action.

One important consequence of this is that if more than one strategy are, say, dominated, then the same fact, the agent's assumption, will count as a reason against playing all these strategies.²⁸ We do not see it as a shortcoming of the analysis. Take for example a two-players coordination game in the style of the one on Figure 3.2, but in which each agent has three strategies, s_1 , s_2 and s_3 . If one agent knows that the other will not play s_1 or s_2 , then this very fact is a conclusive reason, given dominance, for her not to play s_1 , and also for her not to play s_2 . Assumptions, from that point of view, are the strongest reasons that the agents have.

Defined as such, *ex interim* reasons are strongly dependent on the agents' specific "state of mind", so to speak, in a given context. It is worth stressing, however, that these need not to be seen as completely subjective, neither in an intuitive nor in Schroeder's [44] technical sense of the word. For one thing, the reasons mentioned above need not to be subjective in the intuitive sense of not depending on anything else than the agent's informational attitudes. Even by taking beliefs as value of the epistemic parameter, we quite agree with Broome [17]

²⁶See Chapter 2 for the definition of assumption.

²⁷It is clear that different normative source can issue conflicting prescriptions. As we say earlier, we do not cover this here. But taking any subset of $R_i[w]$ to count as a reason would make it possible for a single normative source to pinpoint both reasons for and against playing a certain strategy in a particular interactive situation. Is it not clear to us how plausible are such situations. In any case, they would, again, force us into an analysis of weighing, which we bracket here.

²⁸Thanks to Hannes Leitgeb for pointing this out.

that these beliefs do not necessarily have to be subjective in this sense, that is without any connections to how things actually are. Some relationship with the agent’s evidence for normative facts should also hold.²⁹ What is more, if we allow the epistemic parameter to vary, then the stronger the informational attitude, the less subjective reasons will be. Just think of the extreme case of reasons based on knowledge or hard information. These are definitely owned, but certainly not “subjective” in the intuitive sense. Regarding Schroeder’s definition—recall Section 3.1.2—there is a trivial way in which we depart from his: we do not commit to the view that reasons are, by definitions, facts that are believed by the agent. A deeper difference arises from the relation between *ex interim* and *ex post* reasons, which we discuss in more details in the next chapter (Section 4.4.1).

Reasons from Dominance

Dominance can be seen as issuing strong normative pressure *not to* play a certain strategy. Unpacking the specific preferential parameter of this choice rule, we get:

3.3.4. DEFINITION. [Reasons from Dominance] Agent i ’s $R_i[w]$ is i ’s conclusive reasons for her not to play strategy s_i , given dominance, at state w iff there is another strategy s'_i which she [epistemic – parameter] at w would have yielded outcomes she strictly prefers, *ceteris paribus*.

There is an implicit universal quantification in this formulation. One is asked to check whether, in all situations w' that the agent considers possible given the value of the epistemic parameter, playing s_i against $\sigma_{-i}(w')$ yields a strictly less preferred outcome than playing s'_i against that very combination of strategies of the other players. The *ceteris paribus* clause is intended to remind that for each of these comparisons it is the strategies of agent i that vary. All the rest, and in particular the combination of strategies for the other agents, is to be kept constant.

By instantiating the epistemic parameter we get a more concrete formulation. With plain beliefs, for instance, we get that, in a given context w of a certain game, agent i ’s assumption is a conclusive reason not to play strategy s_i , given dominance, if and only if there is another strategy of her that she believes³⁰ would have given her an outcome she strictly prefers, *ceteris paribus*. To look back at our example above, Ann’s assumption, i.e. that Bob is playing L , is her conclusive reason for her not to play T .

The facts, or sets of outcomes, that count as reason, given dominance, do so because of the particular pattern of preferences concerning these outcomes. This

²⁹This raises the interesting question of objective, in Broome’s sense of the word, expectations in games.

³⁰The correct formulation is “... that she *assumes*, taking belief as R_i ...”. For readability we just say “believes” here.

points towards a Humean reading of this choice rule, although a very weak one. By a Humean theory of reason we mean one that explains reasons in terms of the agent's preferences or desires.³¹ This is clearly the case in reasons induced by dominance. But this is a weak form of Humeanism, or rather what one might call a lower-bound Humeanism. The facts that count as reasons given dominance are such that, if they were true, they would *never* “promote”, “count in favor of” or simply satisfy the agent's preferences. The agent is certain—fill here your favorite value of the epistemic parameter—that choosing such strategy would lead to outcomes that are strictly less preferred. This is a very strong reason *against* choosing a certain strategy, but this says very little about reasons *for* choosing other ones or, in other words, about which action would indeed promote the realization of one's desire, except in case there is a unique non-dominated strategy. In this sense dominance can be seen as a weak form of Humean theory of reasons.³²

Reasons from Admissibility

Admissibility is a weakening of dominance: all dominated strategies are non-admissible, but some non-admissible strategies are not dominated. As we saw, non-admissibility issues negative prescriptions, and *mutandis mutatis* for the reasons it pinpoints:

3.3.5. DEFINITION. [Reasons from Admissibility] Agent i 's $R_i[w]$ is i 's conclusive reasons for her not to play strategy s_i , given admissibility, at state w iff she there is another strategy s'_i which she [epistemic – parameter] at w :

1. would have yielded at least as preferred outcomes and;
2. in at least one situation would have yielded a strictly preferred outcome,

all of this *ceteris paribus*.

The same remarks as for dominance apply here, regarding the implicit universal quantification. It should also be observed that we placed clause (2) within the

³¹See, e.g., the theory developed in Schroeder [44]. Note that he argues that desires are part of that he calls the “background condition”, that they are not reasons themselves but “explain” or “analyze” reasons. We leave these subtleties aside here. Our aim here is rather to highlight the strong relation with desires that reasons have when induced by the dominance choice rule.

³²It is worth observing here that this Humean reading, just like the one in the next section, seems to block the classical decision-theoretical understanding of preferences in terms of consistent choices [42, 29], understanding that is often taking up in game theory. If preferences are to count as reasons for choices, they cannot be themselves based on decisions, at the pain of circularity. Thanks to Julian Reiss and Rohit Parikh for independently pointing out this issue. One way to get the Humean reading to go through here is simply to take preferences as primitive attitudes.

scope of the epistemic parameter. This needs not to worry us, as all attitude that we study here are introspective for their dual.³³

Admissibility allows for more facts to count as reasons against playing a certain strategy than dominance, and thus lend itself to a stronger Humean interpretation. Clause (1) makes it clear that choosing strategy s_i doesn't promote the realization of agent i 's desires anymore than choosing s'_i . Clause (2) gives an additional argument, not to say an additional reason³⁴, for not choosing s_i over s'_i : the former *can* lead to a strictly less preferred outcome. Of course this is not as strong a reason as for dominance. The less preferred outcome will only possibly occur according to the agent's information, while in the case of dominance it is certain. With admissibility, undesirable consequences carry more weight, so to speak. This is why we see dominance as a stronger form of Humeanism. Some facts that would not count as reasons from the perspective of dominance do count as reasons with admissibility.

Yet, just like dominance, admissibility as defined here only set the lower bounds of Humean permissibility. It tells us which actions should not be taken, but not which action should be, except in case there is only one admissible strategy, and do so by relating the possible consequences of one's action to the satisfaction of the agent's preferences. This simple picture of admissibility hides a lot of interesting complications, though. We leave it at that for now, and come back to it in chapters 4 and 5.

This finishes our presentation of *ex interim* normative facts and their associated reasons. So far we said little on the specific features that they get in interactive context. These features appear more clearly once we turn to the dynamic analysis of responsiveness, in the next chapter.

3.4 From Best Response to *ex-post* oughts

In this section we turn to *ex post* ought statements. Their natural game-theoretic counterpart are the notions of best response and equilibrium. We introduce them in the next section, then present a way to view them as normative sources, and finally make a short detour through their epistemic characterization, in order to explain our the "ex post" terminology.

³³If, for instance, one considers it possible that p is the case, viz. do not believe that p is not the case, then one (correctly) believes that one considers this possible. This holds similarly for knowledge and for safe, strong as well as conditional beliefs.

³⁴We take the *combination* of these two clauses to constitute a conclusive reason against playing a certain strategy. Intuitively, neither of these clauses is sufficient to count as a conclusive reason.

3.4.1 Best Response and Equilibrium

Best Response and Nash equilibrium are defined independently to specific informational contexts. They just refer to games, not to the contexts in which the latter are played.

Best response singles out, for each agent and combination of choices of the others, a set of strategies that are considered “rational.” This is being specified in terms of the agent’s preferences:

3.4.1. DEFINITION. Let \mathcal{G} be a game in strategic form and σ_{-i} a profile of strategies for all agents except i . Then the strategy s_i is a *best response* for i to σ_{-i} iff, for all strategies s'_i :

$$(s_i, \sigma_{-i}) \succeq_i (s'_i, \sigma_{-i})$$

A Nash equilibrium is a profile where all agents play their best responses to the choices of others.

3.4.2. DEFINITION. Let \mathcal{G} be a game in strategic form. Then the profile σ is a *Nash Equilibrium* iff, for all agents i , σ_i is a best response to σ_{-i} .

One can think of Nash equilibrium as a profile where no agent has an incentive to deviate unilaterally. No agent would reach an outcome she prefers more by choosing otherwise, everything else, i.e. the choice of the other agents, being kept constant. In the coordination game of page 6, *TL* and *BR* are Nash equilibrium.³⁵

3.4.2 Best Response and Equilibrium Oughts

We want to capture the sense in which, at state *TL* in the model on page 7, Ann and Bob did the best they could, given what the other is doing, despite the fact that, given their beliefs about each others’ action, it seems that they should have done otherwise. Best response gives a natural way to capture this idea.

- Agent i ought^{BR} not to play strategy s_i at state w , written $\mathcal{M}, w \models O_i^{BR} \neg s_i$ iff s_i is not a best response to $\sigma_{-i}(w)$.

Just like for choice rule, we take best-response to issue negative oughts. An agent ought to play otherwise if she doesn’t choose a best response. But best response strategies need not be unique, and in this case the agent will have a range of permitted action, without there being a specific one she is obliged to perform.

³⁵Some well-known facts about equilibria. If one only considers pure strategies, the set of Nash equilibrium of a game might be empty. One of the fundamental result of game theory is, however, that if the set of strategies for all agents is convex, for instance if ones allows for mixing, then Nash equilibrium always exists. Of course, the set of equilibria for a given game can be notoriously large, an observation which, among other considerations, lead to many refinements of Nash equilibrium. In what follows we will not consider these refinements.

		Bob	
		H	T
Ann	h	1, 0	0, 1
	t	0, 1	1, 0

Figure 3.3: Matching Pennies

Ought statements from equilibrium play are defined similarly, except for a proviso taking care of the fact that there might be no way an agent can ensure equilibrium play, *ceteris paribus*.

- An agent ought^N not to play strategy s_i at state w , written $\mathcal{M}, w \models O_i^N \neg s_i$ iff $(s_i, \sigma_{-i}(w))$ is not Nash equilibrium of the underlying game \mathcal{G} , and there is another strategy s'_i for which $(s'_i, \sigma_{-i}(w))$ is a Nash equilibrium of \mathcal{G} .

The second component of the definition is there to ensure that the playing equilibrium is within the agent's power. Look for instance at the game of Matching Pennies (page 19). Here, whatever profile is being played³⁶, there is always one agent which doesn't play a best response, and so there is no profile in which the agents can have insured equilibrium play. The definition above gives an intuitive analysis of that. There is no profiles in which both agents play a best response. Whatever the outcome, there is always one agents for which we can say that she should have played otherwise. Nash equilibrium, however, gives an opposed normative verdict, precisely because of the second clause of the definition. All profiles are permitted.

3.4.3 *Ex post* oughts

Ought statement stemming from best response and Nash equilibrium can be said to be objective in the sense of depending of what, in fact, is being played. But they can be given an more informational rider, referring to the *ex post* stage of information disclosure.

Probably the most well-known epistemic characterization of Nash Equilibrium, from Aumann and Brandenburger [5], state that, for two players, that individual *maximization of expected utility* and *mutual knowledge of strategy choice* are sufficient for Nash equilibrium. The general version of this result, for arbitrary finite number of agents and allowing for mixed strategies, requires *common knowledge of conjectures*, i.e. of each player's probabilistic beliefs in the other's expectations.³⁷ Here it is the converse direction of the result that interests us. In epistemic-plausibility models it boils down to the following:

³⁶In pure strategies, that is.

³⁷What we have to say about this result, however, is more clearly stated by taking its two player, pure strategy form. We stick to it for not.

3.4.3. OBSERVATION. *Let G be a two-players game in strategic form. Then if a profile σ of that game is a Nash equilibrium then there is an epistemic plausibility model \mathcal{M} for it and a state w in that model such that none of the agent chooses a dominated strategy and, for both $i, j \in N$, $\sigma_j(w') = \sigma_j(w)$ for all $w' \sim_i w$.*

This epistemic characterization connects formally the concepts of best response and Nash equilibrium to the *ex post* stage of information disclosure in the two player case. In the previous chapter we indeed defined *ex post* as the stage where all strategic uncertainty, i.e. uncertainty about the strategy choice of others, is resolved. This is precisely the conditions stated in this result, i.e. mutual knowledge of strategy choice. The result says that, from the perspective of each individual, best response can always be thought of in term of knowledge of the choice of the opponent, just like equilibrium play, this time from the perspective of all agents in the game.

Ought statements stemming from best response and equilibrium play can thus always be thought of *ex post* terms, for two players. They describe that what the agents should have done if they would have had full information about the strategy choices of the others or, alternatively, given what that others are, in fact, doing. Interestingly enough, though, for more than two player normative facts stemming from best response and equilibrium play points towards a further stage of information disclosure, where not only the choices are out in the open, but also the agents' expectations about each others' choices. We come back to this in Chapter 4, as this highlights one important, specific character of rationality in social interaction.

3.4.4 Reasons, *ex post*

Best response and equilibrium ought statements or, given what we just said, *ex post* normative facts, get explained by a different kind of reasons than normative facts stemming from choice rule.

(Reasons from Equilibrium)	The fact that the agents different than i play profile $\sigma_{-i}(w)$ at w is a conclusive reason for i not to play strategy s_i , given equilibrium play, iff $(s_i, \sigma_{-i}(w))$ is not Nash equilibrium of the underlying game \mathcal{G} , and there is another strategy s'_i which is a Nash equilibrium of \mathcal{G} .
----------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Reasons from Best Response can be defined a similar manner. Both are owned reasons, but reasons that are not cast in the agent's information at a given state. The facts that count as reason given equilibrium or best response are plain, non-informational. They concern what the others are doing.

The epistemic characterization presented in the previous section allows us to recover the informational, although *ex post* character of such reasons.

(<i>Ex post</i> Reasons)	Let G be a two player game in strategic form, w a state in a given model \mathcal{M} such that $\sigma(w)$ is not a Nash equilibrium, and i an agent for which there is another strategy s_i such that $(s_i, \sigma_{-i}(w))$ is a Nash equilibrium. Let \mathcal{M}', w be a pointed model just like \mathcal{M}, w except that i knows the strategy choice of all agent $j \neq i$. Then $R'_i[w]$ is a conclusive reason not to play s_i , given dominance.
---------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Ex post ought statements can be thought of in terms of an hypothetical state of full information disclosure, and this is where the reasons they generate live, epistemically.³⁸

To go back to the example on page 7, we thus have two readings of the idea that there is a reason explaining the fact that, at state TL , Ann and Bob did the best they could. The first reading is the objective one: the reason explaining this normative fact is the very fact that Ann chooses T and Bob chooses L . The second reading is epistemic: *if* Ann had had full information about Bob's choice, this very information would have constituted a reason for her not to play B , given dominance, and *mutadis mutandis* for Bob.

3.5 Conclusion

This chapter laid the ground for our general investigation into the notion of interactive rationality. We explained how solution concepts and choice rule generates two different kind of normative facts, *ex post* and *ex interim*, facts that get in turn explained by two different kinds of reasons - factual and informational, respectively. Interactive rationality consists, in part, in responding correctly to such reasons. But responsiveness itself is a many-faced phenomena, which we explore in the next chapter.

³⁸The reader should keep in mind, again, that what is meant here by “full information disclosure” is different with two than with more than two players. In the first case this means full knowledge of the opponent's strategy choice, in the other this means that *and* common knowledge of mutual expectations.

Chapter 4

Interactive Rationality and the Dynamics of Reasons (Part 2)

Release beta 1

We are investigating the notion of interactive rationality. The previous chapter laid the foundation of our investigation through a reformulation of contemporary epistemic game theory in terms of normativity and reasons. We now turn to the notion of *responsiveness* to reasons and its relation to interactive rationality. For us responsiveness is a *dynamic, normative* notion, bearing on how agents should react to changes in the context of the game, viz. in terms of changes in the factual and higher-order information. Viewed this way, basic logical observations about the dynamic behavior of choice rules and informational attitudes turns out to be normatively significant. They echoes different strengths of rational responsiveness, and they highlight the specificity of interactive situations.

4.1 Interactive Rationality \neq Responsiveness

We follow Broome [17] and take correct responsiveness to reasons to be only one aspect of Interactive Rationality, broadly conceived. This broad notion of rationality in interaction also encompasses *rational requirements* of epistemic consistency and practical coherence.¹ We take correct response to reasons *and* conformity to rational requirements to be two necessary conditions for interactive

¹By epistemic consistency we mean consistency between the agents' informational attitudes. Practical coherence bears on so-called pro-attitudes [18], such as desires and intentions, and on their relation with informational attitudes. For single-agent situations, such requirements have been extensively studied in the last years. See e.g. [16, 52, 51, 31, 17]. Social interactions, however, do raise requirements of their own. We come back to them in Chapter 5, but for now it is worth noting that requirements of consistency between an agent's first and higher-order information that we presented in chapter 2, i.e. the notion of "consistent belief hierarchies" [13], is a clear example.

rationality in this broader sense.² If an agent doesn't respond correctly to the reasons induced by a given choice rule we will thus say that she is not rational. The reader should keep in mind, however, that in general the implication does not go the other way. Anticipating the discussion in Chapter 5, an agent can be rationally criticized for not reasoning *to* a "proper" context even though she responds correctly to her reasons. We leave this for now, though, and focus on responsiveness.

4.2 Static Responsiveness

The basic, static notion of responsiveness to reasons that we use come from what Broome calls the *enkratic condition* [17]:

(Enkrasia) Rationality implies that if you believe that your reasons require you to do action *A*, then you should form the intention to *A*.

The wider consequent—i.e. what follows "Rationality implies that"—describes the understanding of correct responses to reasons that we adopt here. In the general case, we take static responsiveness to reasons to be a matter of forming (or not) the intentions corresponding to what one believes one ought (or not) to do. Enkrasia states that responsiveness is necessary for rationality, as discussed in the previous section.

This definition makes responding to reasons a matter of forming the right intentions, not necessarily of taking the right actions. The reason for this is that hardware failures, so to speak, should not bear on our assessment of the agents' rationality. As long as the agent forms the right future-directed intentions³, she won't be deemed irrational if, say, a deviant causal chain prevent her from enacting these intentions.

We bracket this distinction between intentions and actions in the present analysis. Epistemic models for games are not designed to study explicitly the

²Depending on how inclusive one thinks of rational requirements, these two conditions can turn out to be jointly sufficient to deem an agent fully rational in situation of interaction, but we will not argue for this claim here. It is also not uncontroversial whether rational requirements are explainable in terms of responsiveness to reasons—see e.g. Broome's [17] argument to the effect that there is no "attitudinal reasons"—or vice-versa. This issue is not crucial either for what follow.

³See Bratman [15] for more on future-directed intentions. As far as responsiveness to reasons is concerned, it seems to us that the formation of future-directed intentions is enough. Whether the agents act on these intentions or not, i.e. whether she forms the corresponding intention in action [45] or makes the appropriate attempt [34], is a matter of rational requirement. As the next paragraph will make clear, however, what we have to say about interactive rationality does not hinge on this point.

relations between intentions and action in interaction.⁴ One is free, however, to interpret them either as models of intention directly, thus interpreting the strategy choices really as the formation of intentions, or to simply make the idealizing assumption that the agents always successfully enact their intentions. In both case the issue of hardware failure is left aside.

The second thing to notice about this definition is that it involves *beliefs* about reasons, instead of reasons themselves. The idea here is that an agent should not be deemed irrational for not responding to facts she doesn't take to be reasons. It might be, for instance, that admissibility proscribes an agent to take certain actions. If, however, the agent doesn't believe in the normative character of admissibility, it is not irrational of her to choose actions that are not admissible. To assess an agent's responsiveness to reasons one needs to take into account what the agent recognizes as normative source.

Just like for the distinction between intentions and actions, in the present context we bracket the issue of recognizing a source as genuinely normative. Part of the reason for this is that the two choice rules we considered in the previous chapter are *introspective*⁵ - and so are the *ex interim* ought statements that they induce. Most other choice rules that we mentioned in passing, maximin for instance, are also introspective. Introspection provides a clear connection between reasons and correct beliefs about them. It does not, however, entail (correct) beliefs that these very facts count as has a reasons for or against certain action. To reach this conclusion one needs the additional premise/assumption that the agent believes in the normative validity of the choice rule, that she recognizes it as a normative source. We leave this last step implicit in that follow. When we talk about believed reasons for or against a certain action, given a choice rule, we will simply assume that the agent recognize this choice rule as a normative source, as providing reasons for action. Similarly, when we will state results involving common belief in certain choice rule, we will simply assume that it is also common beliefs that the agents recognize that choice rule as providing reasons.⁶

⁴Game models with intentions have been studied in recent years, c.f e.g. [38, 39]. It would be interesting to see how the present analysis of interactive rationality should be carried to these richer scenarios.

⁵We say that a choice rule is positively introspective whenever satisfaction of the choice rule implies belief/knowledge/... of this very fact. Positive introspection can of course take many form, depending on whether on the informational attitude one uses in the consequent. But here we are concerned with beliefs, i.e. in believed reasons. It is an easy exercise to check that the two choice rules that we introduced the previous Chapter (Section 3.3.1), are positively introspective for conditional beliefs. This holds in fact for all attitudes presented in Chapter ???. All these attitudes are positively introspective, and that this property carries over to the choice rule. A choice rule is negatively introspective when failure to satisfy the choice rule implies belief/knowledge/... of that fact. Dominance and admissibility are negatively introspective for conditional beliefs. This holds also for knowledge, but not with safe beliefs. Finally, a choice rule is fully introspective when it is both positively and negatively introspective.

⁶Not that we think that it is not important to analyze the state of recognizing and endorsing a given normative source, quite the contrary. This is simply a too large meta-ethical question

We call this view on responsiveness “static” because it relates what the agent believes with what she does in *one* specific context. Formally, this gives us the following, taking the above qualifications into account:

4.2.1. DEFINITION. Let \mathcal{M}, w be a given context for a game \mathcal{G} and $O_i^C[\neg]s_i$ a normative fact, stemming from choice rule or solution concept C , that holds in that context, for reason φ .⁷ Then *i responds correctly to φ* iff *i* [doesn’t] plays s_i at w ($\sigma_i(w) = [\neq]s_i$).

An agent, in a context, responds correctly to her reasons given a specific choice rule when her action is not inconsistent with the normative prescription that this choice rule issues. Static responsiveness to reasons, from that point of view, is just what most epistemic game theorists call the “event”—i.e. proposition or set of state where— that agent *i* is “rational.” Or rather the other way around: the technical notion of “rationality” in the epistemic approach to interaction is, we think, better understood in terms of correct responses to reasons, given a specific choice rule or normative source. As we mentioned already, we reserve the term “Rationality” for the broader notion, encompassing both correct responsiveness to reasons and meeting rational, interactive requirements.

Before moving to dynamic responsiveness, a remark regarding the notion of *correct* response to reasons is in order. Intuitively, the fact that an agent’s intention/action constitute an adequate response to the reasons she believes she has doesn’t necessarily entails that this she is responding *correctly* to her reasons. Correct response seems to entail some kind of causal connection, or at the very least some “reliability” [33] in the relation between an agent’s recognition of owned normative facts and her action. This is yet another issue that we bracket here. We make the clearly unrealistic idealization that the appropriate connection between recognition of reasons and action holds - that plain responsiveness as defined above is just correct responsiveness, and leave the lifting of this idealization for future work.

4.2.1 A Dynamic Perspective on Responsiveness

Static responsiveness, as defined above, is a state-based attribute of the agent but, intuitively, responding correctly to reasons is also a matter of “tracking” reasons through changes in the context.⁸ An agent who respond correctly to her reasons should be disposed to dynamically reconsider and, if necessary, change

for the scope of this book, a question that as, to our knowledge, not received a formal treatment yet.

⁷Recall that reasons, either *ex interim* or *ex post*, are facts/propositions in epistemic plausibility models. As explained in Section 3.3.3, for *ex interim* reason we get $\varphi = R_i[w]$ for a given instantiation of R_i . For *ex post* reasons we get that $\varphi = R'_i[w]$ with R'_i being just like R_i except that *i* knows the strategy that the others are choosing at w (Section 3.4.4).

⁸See [31] for more “state” and “process” rationality.

one’s intention or strategy choice when new information or changes in reasons come in [25]. There is a specific normative dynamics induced by changes in the informational context. This is what we investigate now.

We focus on *informational* changes, also called “epistemic actions” [49]. These are changes in the information the agents have, not changes in the basic, non-informational facts, like strategy choices or the structure of the game. Epistemic actions can nevertheless alter normative facts. Look back, for example, at the context depicted in the Figure on page 7. At state TL , Bob ought to play R , given dominance and his beliefs about Ann’s choice. But suppose now that he learns that Ann might play T after all, resulting in the model of Figure 4.1. In this updated model L becomes permitted for Bob. Epistemic action can thus induce changes in normative facts. In this case we go from a negative assessment to a positive one, but this can also go the other way around. Correct response to believed reasons might become incorrect after some changes in the agent’s informational context.

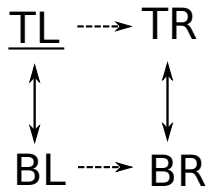


Figure 4.1: A new informational context for the game on page 6.

Our focus on epistemic action entails that we will be mainly concerned with the dynamics of *ex interim* ought statements. Indeed, the strategy profile $\sigma(w)$ that is played at a given state w is kept fixed by epistemic actions, and so the latter do not change *ex post* normative facts. We come back, however, to such facts at the end of the chapter.

We thus analyze responsiveness to reasons in terms of dynamic of normative facts. This is a normative, dynamic perspective on responsiveness to reasons. Since our account of *ex interim* normative facts and conclusive reasons ultimately rests on satisfaction of certain choice rules, studying the dynamic of responsiveness boils down to study the kind of changes in informational contexts that trigger, or not, changes in satisfaction of a given choice rule. This is what we proceed to do in the next sections: trying to understand the dynamic properties of choice rules in order to circumscribe the kind of changes in the context of the game to which normative facts and reasons are sensitive to.

Two remarks are in order beforehand. First, as already mentioned in the previous chapter (Section 3.3.1), the dynamic behavior of choice rules will greatly depends on the value of the epistemic parameter. Making a choice rule dependent, say, on the agent’s hard information will make for extremely stable nor-

mative facts, only variable under truthful information received from a completely trusted source. On the other extreme, taking plain beliefs as the value of the epistemic parameter will make for volatile normative facts, which could be flipped by (almost) any small changes in the agents' informational contexts. In this chapter we will not attempt to cover the dynamic properties that choice rules inherit from there epistemic parameter. We come back to this in the next one, and refer to contemporary work in dynamic-epistemic logic, e.g. [6, 48, 7], for more on the topic. Our focus now is the dynamic behavior that choice rules get from their own logical form.

Second, we should stress that the goal of the remarks in the coming sections is to lay the ground for our methodology. We do not attempt at covering all possible types of information change. Chapter 5 proposes a more in-depth investigation in that direction. For now it is enough to highlight some baseline results.

4.3 Supervenience or the No Mysticism Requirement

Not all changes in a specific context of an interaction should have normative significance. This is an intuitive idea, which has also found echos in the philosophical literature, for instance in the following criterion, paraphrased from [17]:

(Supervenience of Rationality on the mental)	If, in two different contexts of a given game, the agent's attitudes are the same, then she should be rational in one iff she is rational in the other.
----------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------

The basic idea here is that changes that do not affect the agent's "attitudes" should not induce normative changes. The strength of this criterion will of course depend on one's theory of informational and pro-attitudes. For an externalist, different assessment of rationality can be explained by small changes in the agent's environment. A more internalist view of informational attitudes, on the other hand, will be more robust to external changes. Even if the Enkratic condition (Section 4.2) clearly meets the Supervenience constraint⁹, the latter thus needs sharpening.

The present formal approach offers a natural way to sharpen the supervenience requirement, by looking at invariance of epistemic-plausibility models under given (logical) *languages* that describe them. In Chapter 2 we used the language from [6] to talk about epistemic-plausibility models. It contains operators for individual knowledge and conditional beliefs and well as two group or collective attitudes operators, common knowledge and common beliefs. Such languages induce a natural notion of invariance between informational contexts of interaction, which is readily usable to flesh out the supervenience requirement, as follow:

⁹Identity of intentions ensures equivalence in rationality assessment, *ceteris paribus*.

4.3.1. DEFINITION. [No Mysticism] Let \mathcal{L} be a given language interpreted on epistemic-plausibility models, and \mathcal{M}, w and \mathcal{M}', w' be two such pointed models. Then if \mathcal{M}, w and \mathcal{M}', w' satisfy exactly the same formula of \mathcal{L} , for all agents i , then i is rational in \mathcal{M}, w iff she is rational in \mathcal{M}', w' .

We call this principle “no mysticism” because it states that changes in informational contexts that cannot be expressed in a given (formal) language should not bear on our assessment of the agents’ rationality. Mystical changes, so to speak, are ruled out. The connection with the Supervenience Requirement should be clear: insofar as a given logical language can be seen as encoding all the possible attitudes of agents in some classes of interactive situations, then language invariance becomes attitude invariance, and vice-versa.¹⁰

In Section 4.1 we took the view that the interactive rationality implies correct responsiveness to reasons, but not necessarily the other way around. Here we are focusing on responsiveness, but it is worth observing that No Mysticism goes further than that, by requiring that *Rationality* should be invariant under language invariance. We leave this for now, though, and look at whether the satisfaction of a given choice rule is invariant under up to a given language equivalence.

The strength of No Mysticism depends on the expressive power of the language that one uses. Generally speaking, the type of changes in given class of structures up to which a given language is invariant can be characterized in terms of morphisms between structures [26]. In the present case the modal language that we introduced in the last chapter is invariant under *bisimulation* [9], for which we give here a general definition for relational models for games:¹¹

4.3.2. DEFINITION. Let $\mathcal{M}, \mathcal{M}'$ be two models for game \mathcal{G} . Then the mapping $\Leftrightarrow \subseteq W \times W'$ is a *bisimulation* iff, for all w, w' such that $w \Leftrightarrow w'$ and all relations R_i for agent i :

- $f(w) = f'(w')$.
- If wR_iv then there is a $v' \in W'$ such that $w'R'_iv'$ and $v \Leftrightarrow v'$.
- If $w'R'_iv'$ then there is a $v \in W$ such that wR_iv and $v' \Leftrightarrow v$.

¹⁰It is important to observe here that we do not advocate here a deep connection between so-called propositional attitudes and their expressibility in natural language. Our approach is purely methodological, and bears on *formal* languages. We simply observe that such language give a natural way to list the possible attitudes of agents in interaction. How this list is constituted or, alternatively, why a specific language instead of another, is left open.

¹¹Bisimulation is usually defined on arbitrary relation models, given a set of propositional variable in the language. Here we give a special case for models for games, where the strategy function plays the role of the general propositional valuation. See, e.g. [9] for details. Also, the precise definition of bisimulation for the specific language and models mentioned above is more involved, because of the conditional belief operator. Details can be found in [19]. See also [20] for relevant, related remarks.

Bisimilar models satisfy exactly the same formulas of these languages.¹² In the present context: pairs of bisimilar models are “mentally” equivalent, up to a given language that describe the agents’ informational attitudes. Agents know/believe something in one model if and only if she knows/believes the same thing in the other. Observe that we are keeping the underlying game constant throughout, so preferences are also fixed. Also of interest is the fact that no additional clause is required to take care of common knowledge and common belief operators, even though in the definition above only relations for individuals are mentioned.

With this in hand we can re-formulate the No Mysticism principle in terms of bisimulation:

4.3.3. DEFINITION. [No Mysticism - Model-Theoretically] Let \mathcal{G} be a game in strategic form, \mathcal{M}, w and \mathcal{M}', w' two epistemic models for it, and C a choice rule. Then for all strategies s_i of an arbitrary agent i , if $\mathcal{M}, w \leftrightarrow \mathcal{M}', w'$ then $\mathcal{M}, w \models O_i^C[\neg]s_i$ iff $\mathcal{M}', w' \models O_i^C[\neg]s_i$.

To get closer to the idea of static responsiveness to reasons that we gave in Section 4.2, one only have to substitute the strategy $\sigma_i(w)$ that the agent i is actually playing at w ,¹³ for s_i in the this definition. No mysticism states that agents should not be responsive to inexpressible changes in the context.

Note that this requirement is weaker than Broome’s supervenience requirement. Changes incurred under bisimulation *never* echoes in changes of attitudes, given a specific epistemic language. Two non-bisimilar models might happen to satisfy the same epistemic formulas: according to the supervenience requirement they should also satisfy the same normative facts, which according to the second non-mysticism requirement this needs not be the case—although it is the case for normative facts based on admissibility and dominance.

Formulated this way, No Mysticism can be interpreted as fixing an upper bound on how responsive to changes in contexts the agents should be. Changes in reasons that are beyond the expressive power of the language at hand shouldn’t bear on the assessment of the agent’s rationality, or at the very least the agent should not be expected to respond to such changes.¹⁴ A simple inspection of the logical form of dominance, and admissibility reveals that they meet the No Mysticism principle:

¹²The other direction is only true for image-finite models. But, in general if two models are modally equivalent they can be transformed so as to become bisimilar.

¹³And thus w' , because the two models are bisimilar.

¹⁴We use here the two definitions of No Mysticism interchangeably, because of the tight connection between the kind of modal languages we have in mind and bisimulation. But the reader who doesn’t want to commit to this modal perspective and wish to use other, more expressive languages, can resort to the language-based definition of No Mysticism. The results that we mention shortly still hold for more expressive languages, but of course not for arbitrary decrease of expressive power.

4.3.4. OBSERVATION. Let \mathcal{G} be a game in strategic form, \mathcal{M}, w and \mathcal{M}', w' two epistemic models for it, and s_i be a strategy for an arbitrary agent i . Then if $\mathcal{M}, w \underline{\leftrightarrow} \mathcal{M}', w'$ we get:

- $\mathcal{M}, w \models O_i^D(s_i)$ iff $\mathcal{M}', w' \models O_i^D(s_i)$.
- $\mathcal{M}, w \models O_i^A(s_i)$ iff $\mathcal{M}', w' \models O_i^A(s_i)$.

From a logical point of view, these results are not surprising: it is their normative import that matter here. Responsiveness to *ex interim* reasons from dominance or admissibility meet this upper-bound requirement. From either of these choice rule, agents are not expected to respond to changes in the context that could never translate in changes in their informational attitudes. This is good.

4.4 Upward and Downward Monotonicity

We now consider basic increases and decreases in information, which we treat both model-theoretically.¹⁵ By an increase in information we mean that the agent considers less states possible, and thus knows/believe/... more facts. The other way around, information decreases when the agent is considering more states possible, thus knows/believes/... less. We are interested in transformations of a given epistemic model of a game that lead to increases or decreases in information. Whether we deal with changes in knowledge, beliefs or any other attitudes will depend on the specific value of the epistemic parameter of the choice rules at hand. In most cases we won't have to be specific about this parameter, though, and so will be mostly considering abstract changes in the agent's assumption $R_i[w]$, as we did earlier in this chapter.

The question we want to answer here is whether arbitrary increases or decreases in the agent's information can induce changes in normative facts, which would call for a normative reassessment the agents' strategy choices. In our case this means checking whether the choice rules that we investigated so far are either *upward* or *downward monotonic*:

4.4.1. DEFINITION. A choice rule C is *upward* [*downward*] monotonic whenever, for any state w of a model \mathcal{M} of a given game \mathcal{G} and sets $R_i[w], R'_i[w]$, if $R_i[w] \subseteq [\supseteq] R'_i[w]$ and $\mathcal{M}, w \models O_i^C[\neg]s_i$ given $R_i[w]$, then $\mathcal{M}, w \models O_i^C[\neg]s_i$ given $R'_i[w]$.

An upward monotonic choice rule is thus one that is stable or, in other words, not sensitive to decreases in the agent's information. If agent i 's choice satisfy an upward monotonic choice rule at a given state, then i will still satisfy this choice rule even if she loses information about the current situation. Put in terms of reasons, an upward monotonic rule will not ask for a different response to the new

¹⁵C.f. [35] for a syntactic treatment.

reasons an agent has when these reasons are strictly weaker than the one she has before. A downward monotonic rule works the other way around. It is stable, i.e. doesn't ask for different responses, under increases in information. Monotonicity turns out to be crucial for the epistemic analysis of solution concepts, see e.g. [1]. We come back to it at length in the next chapter. For now we keep to these following basic observations:

4.4.2. OBSERVATION. *Dominance is downward monotonic. Admissibility is neither upward nor downward monotonic.*

These are observations in the very literal sense of the word: one only have to look at the logical form of these choice rules to ascertain them. Dominance, for instance, gets its downward monotonicity from the single universal quantification over the alternatives in the epistemic parameter. If the agent gets a strictly less preferred outcome for all the possibilities in a given domain, here the agent's assumption $R_i[w]$, then she does also for any sub-domains, i.e. non-empty subsets of $R_i[w]$. In normative terms: dominance induces normative facts that are stable under increase of information. If $R_i[w]$ is a conclusive reason for i against playing s_i at w , given dominance, then if the agent receive new, more precise information $R'_i[w]$ then this piece of information is also a conclusive reason for i against playing s_i at w , given dominance. Increase in agent's information do not call for different responses to reasons, given dominance. Observe, however, that dominance is *not* upward monotonic. Indeed, by losing information the agent can take into account new eventualities, viz. possible outcomes, that cancel a the negative recommendation against playing a given strategy.

Admissibility, on the other hand, is not monotonic, neither upward nor downward. Just like for dominance, the failure of upward monotonicity comes from the universal quantification among the outcomes that are considered possible. Moving to a larger domain of quantification means more chance to find a defeater, thus the failure of upward monotonicity. The failure of downward monotonicity, on the other hand, comes from the single existential clause in the definition of admissibility. Moving to a proper sub-domain might leave out the witness to this existential quantification, making admissible a strategy that was not before. In other words, recognizing admissibility as normative source means that both increases and decreases in agent's information might call for different response to the agent's changing reasons. In terms of responsiveness, admissibility is more demanding than dominance. Of course, this doesn't means that *all* such changes will result in changes in conclusive reasons. But we leave it at that for now, as in the next Chapter we come back in great length on the dynamics of reasons based on admissibility.

4.4.1 Objective and subjective reasons revisited

In the previous chapter (Section 3.1.2) we mentioned that the distinction between *ex interim* and *ex post* normative facts and reasons comes close, but is still different from like Schroeder's [44] distinction between subjective and objective reasons. The observations regarding upward and downward monotonicity make this claim more precise.

Schroeder's idea is that a subjective reason for or against i choosing a certain strategy is a fact that i believes to be true and, if this it were true, would count as a objective reason for or against playing a certain strategy. The natural hypothesis is that *ex interim* and *ex post* reasons are, respectively, subjective and objective in this sense. A believed fact is an *ex interim* reason for or against playing a certain strategy, the hypothesis goes, whenever, if that fact were be true, it would count as an *ex post* reason.

This hypothesis holds for both negative and positive prescriptions issuing from dominance, precisely because the normative facts induced by this choice rule are downward monotonic.

4.4.3. OBSERVATION. *Let \mathcal{M}, w be an arbitrary epistemic-plausibility model for a given game \mathcal{G} , and suppose that $\varphi = R_i[w]$ is an *ex interim* reason, given dominance, for [against] playing strategy s_i at \mathcal{M}, w . Then for all $w' \in \varphi$, $\sigma_{-i}(w')$ is an *ex post* reason for [not] playing s_i .*

The argument for this goes as follow: Take any w' in φ . We know by assumption that s_i is either dominated by another strategy s'_i , or is the unique dominating strategy given φ . In the first case, because dominance is downward monotonic, it s_i is also dominated by s'_i given $\{w'\}$. In the second case, it stays that s_i is strictly dominant given $\{w'\}$. This means that $\sigma_{-i}(w')$ is an *ex post* reason [not] to play s_i .

For strategies that are permissible given dominance, however, the correspondence fails. Reasons for not excluding a given strategy might stem of an agent's uncertainty at a given state, like for instance for Bob in the model on page 27. In that specific case, Bob's *ex interim* reason, which can be describe as the fact that Ann either plays T or B , neither exclude R nor L . This reason is trivially true at, say, TL .¹⁶ But in that state T is nor an *ex post* reason permitting R . Here Bob ought, *ex post*, to play L .¹⁷

For similar reasons, the failure of downward monotonicity in the case of admissibility breaks the correspondence *ex interim/ex post* and subjective/objective reasons. Strategies that are permitted, *ex interim*, given admissibility, can be non-admissible *ex post*. More interestingly, the converse is also true: strategies that

¹⁶If Ann plays T at a state then she either plays T or L at that state.

¹⁷Here we slightly abuse our terminology. Strictly speaking, in Chapter 3 we only defined reasons on the basis of positive or negative normative facts, not on the basis of permissions. The latter can be recovered, however, by just considering a reason allowing for a strategy in terms of the absence of reasons for excluding it.

are excluded *ex interim* can become admissible *ex post*. In the next chapter we study this phenomenon in greater details.

4.5 Changes in Higher-Order Information

The dynamic take on responsiveness also highlights the *interactive* character of decision making in social situation. So far we looked at dynamics of reasons very generally, considering abstract increases and decreases in information. Informational contexts of interaction, however, are specific in that they not only include factual information, that is information about who-plays-what, but also higher-order information: information about “who-believes-what”. Sensitivity to changes in higher-order information means that a given choice rule calls for responses to specifically interactive changes in reasons.

When considering changes in higher-order information, we will be focusing on changes in an agent’s knowledge/beliefs/... about the information of others. These are cases where the agent considers possible new configurations of knowledge/beliefs for some others or, in other worlds, new types of others, or maybe rule out some, or maybe both — we won’t be restricting to pure increases and decreases here.

4.5.1. DEFINITION. Let C a given choice rule and \mathcal{M}, w a pointed model for a given game \mathcal{G} . Write $\mathcal{T}_i[w]$ for the set of types for the other players that agent i considers possible at w .¹⁸ Suppose $\mathcal{M}, w \models O_i^C[\neg]s_i$ for a given strategy s_i . Then C is *sensitive to changes in higher-order information* whenever there is a pointed model \mathcal{M}', w' for the same game \mathcal{G} such that $\mathcal{T}_i[w] \cap \mathcal{T}_i[w'] \neq \emptyset$ and $\mathcal{M}', w' \not\models O_i^C[\neg]s_i$.

4.5.2. OBSERVATION. *Dominance and admissibility are sensitive to changes in higher-order information.*

In both cases, basic results in epistemic game theory provide the required witnesses for this observation. The epistemic characterization of iterated elimination of strictly dominated strategies (IESDS) [36, 8, 12] is that dominance and *common belief* in dominance is sufficient for the latter.¹⁹ In this result, the n^{th} step in the hierarchy of common beliefs²⁰ corresponds to the fact that players don’t

¹⁸In epistemic plausibility models types can be identified with hierarchies of conditional beliefs and knowledge. See Chapter 2 for more details.

¹⁹The usual formulation of this result is that “rationality” and common belief in rationality is sufficient for IESDS. Our formulation only marks the conceptual difference, that we stressed in Section 4.1, between rationality broadly conceived and correct responses to what count as reasons given a certain choice rule, here dominance.

²⁰The 0^{th} step in the hierarchy is the big conjunction $\psi^0 = \bigwedge_{i \in N} \varphi_i$ with φ_i being of the form “agent i believes that no one chooses a dominated strategy”. The $n + 1^{\text{th}}$ step is the big conjunction, for all agents, of “agent i believes ψ^n ”. See Chapter 2 for details.

play dominated strategies corresponds to the n^{th} round of elimination of strictly dominated strategies. In other words, as levels of higher-order information about dominance pile up, new strategies might become dominated, given the agents' new information. For our present investigation this means the following: changes in higher-order information may call for different responses to reasons, given dominance. Facts about what others know/believe/... can count as reasons for or against playing a certain strategy, given dominance.

The situation is similar with admissibility, although slightly more complicated. Steps in the hierarchy of common belief in admissibility do not correspond to steps in the elimination of weakly dominated strategies, and the limit of the latter algorithm needs not to correspond to the set of strategies that can be played under admissibility and common belief of admissibility [41, 14]. One can, however, easily construct an example where normative facts induced by admissibility changes with changes in higher-order information. The peculiarity of admissibility is not its sensitivity to changes in information about information, but rather how it behaves under such changes, so to speak—more on that in the next chapter.

Just like in the previous sections, these two observations are formally very basic: their main import is on the normative level. They show that normative facts issuing from dominance and from admissibility are sensitive to changes in higher-order information. Agents in social contexts are required to keep track of these changes. Responsiveness to reasons in interaction is different than responsiveness to reasons in pure single-agents contexts.

Dominance and admissibility are very sensitive to the specifics of interaction, though. There are simple games where minimal changes in arbitrary high order of information about dominance propagate all the way down to what is rational to do or, in the present terminology to, normative facts, c.f. [40]. This raise important normative questions: does it make sense to *require* the agents to keep track of what they believe about what others believe about..., up to arbitrary high levels? Can an agent be blamed because he did not take into account the fact that, say, mutual belief in dominance failed at the 100th level? We do not attempt to answer this question here. For now, the simple fact that it can be raised is sufficient for our claim: once again, responsiveness to reasons in interaction is different than responsiveness to reasons in pure single-agents contexts.

4.6 Responsiveness and *ex post* reasons

The static notion of responsiveness (Section 4.2), requires agents to respond to reasons they *believe* they have. But *ex post* reasons, either from equilibrium or from best response, are not necessarily introspective.²¹ State *TL* in the figure on page 7 provides an example. It seems that, in general, agents are thus not

²¹See footnote on page 25 for a definition of introspection.

required to respond to such reasons, even though it might be that, *ex post* they should have done otherwise.

It can, however, very well be that agents *believe* that they have *ex post* reasons. In that case believed *ex post* reasons are “subjective” in the sense that we developed in Section 4.4.1.

4.6.1. OBSERVATION. *Let \mathcal{M}, w be an arbitrary observational context for a given game \mathcal{G} and assume that agent i believes that she has an *ex post* reason for [against] playing strategy s_i . Then if this belief was true i would have an *ex post* reason for [against] playing strategy s_i .*

The argument for this is no more than unpacking the definitions, and observing that having a belief in an *ex post* reason boils down to believing that the other agents are playing a specific profile σ_{-i} .

Of course, in cases where an agent does have full information about the choices of other, *ex interim* and *ex post* reasons come together, and the static responsiveness requirement kicks in. It is a easy check that, in such contexts, *ex post* reasons meet the No Mysticism requirement, and they are (trivially) downward monotonic.²² More interestingly, *ex post* reasons are *invariant* under changes in *higher-order* information, for two players. Indeed, correct or incorrect response to *ex post* reasons, in the two players case, is fully determined by the single level of knowledge of choices of the other that the agents have at that stage. Uncertainty about the information of other doesn’t come does not change the *ex post* normative facts. The situation, of course, is different for more than two players. In these case higher-order knowledge of strategy choice does make a difference. This this is another clear instance of the specificity of interactive rationality.

4.7 Conclusion

In this chapter we have made the first concrete steps in our study of the notion of Interactive Rationality. Our focus was on one aspect of this broad notion: responsiveness to reasons, from a static as well as from a dynamic perspective. By laying down basic facts about the dynamic behavior of normative facts induced by choice rules, we have been able to formulate a upper bound requirement on interactive responsiveness, the No Mysticism requirement. We also saw that basic monotonicity properties of dominance and admissibility shed light on the distinction between objective and subjective reasons. By looking at the dynamics of higher-order information we then saw that social interaction does indeed brings in something new to the theory of reasons.

This is only the starting point of the investigation in the dynamics of reasons, tough. As well will see in the next chapter, bringing together responsiveness

²²Upward monotonicity makes less sense in this case. Loosing information agent can mean leaving the *ex post* stage.

to reasons and rational, interactive requirements unfold the whole complexity of interactive rationality, in a dynamic perspective.

Acknowledgments

This chapter already greatly profited from the comments and suggestions of many. Thanks to Martin van Hees, Peter Timmerman, Frank Hindriks, Lieuwe Zijlstra, Constanze Binder, Jeanne Peijnenburg, Wes Holiday, Hannes Leitgeb, Conrad Heilmann, Elias Tsakas, Julian Reiss, Jack Vromen, Wlodek Rabinowicz and Denis Bonnay.

Bibliography

- [1] K.R. Apt. The many faces of rationalizability. *The B.E. Journal of Theoretical Economics*, 7(1), 2007. Article 18.
- [2] S. Artemov. Knowledge-based rational decisions. Technical report, CUNY Graduate Center, 2010.
- [3] R. Aumann. Interactive epistemology I: Knowledge. *International Journal of Game Theory*, 28:263–300, 1999.
- [4] R.J. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8:121–133, 1994.
- [5] R.J. Aumann and A. Brandenburger. Epistemic conditions for nash equilibrium. *Econometrica*, pages 1161–1180, 1995.
- [6] A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In Giacomo Bonanno, Wiebe van der Hoek, and Michael Wooldridge, editors, *Logic and the Foundation of Game and Decision Theory (LOFT7)*, volume 3 of *Texts in Logic and Games*, pages 13–60. Amsterdam University Press, 2008.
- [7] A. Baltag and S. Smets. Group belief dynamics under interactive revision: fixed points and cycles of joint upgrades. In Aviad Heifetz, editor, *Proceedings of TARK 2009*. ACM, 2009.
- [8] D. Bernheim. Rationalizable strategic behavior. *Econometrica*, 52:1007–1028, 1984.
- [9] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, Cambridge, 2001.

- [10] A. Brandenburger. The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory*, 35:465–492, 2007.
- [11] A. Brandenburger. Epistemic game theory: an overview. In S. Durlauf and L. Blume, editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, Basingstoke, 2008.
- [12] A. Brandenburger and E. Dekel. Rationalizability and correlated equilibria. *Econometrica*, 55:1391–1402, 1987.
- [13] A. Brandenburger and E. Dekel. Hierarchies of beliefs and common knowledge. *Journal of Economic Theory*, 59, 1993.
- [14] A. Brandenburger, A. Friedenberg, and H. J. Keisler. Admissibility in games. *Econometrica*, 76:307–352, 2008.
- [15] M. Bratman. *Intention, Plans and Practical Reason*. Harvard University Press, London, 1987.
- [16] M. Bratman. Intention, belief, practical, theoretical. Unpublished Manuscript, Stanford University, January 2006.
- [17] J. Broome. Rationality through reasoning. Manuscript.
- [18] D. Davidson. *Essays on Actions and Events*. Clarendon Press, Oxford, 1980.
- [19] C. Dégremont and O. Roy. Agreement theorems in dynamic-epistemic logic. In A. Heifetz, editor, *Proceedings of TARK XII, the 12th Conference on Theoretical Aspects of Rationality and Knowledge*.
- [20] L. Demey. Agreeing to disagree in probabilistic dynamic epistemic logic. Master’s thesis, Institute for Logic, Language and Computation, Amsterdam, 2010.
- [21] F. Dietrich and C. List. A reason-based theory of rational choice. *Nous*, Forthcoming.
- [22] S. Finlay. Recent Work on Normativity. *Analysis*, 70(2):331, 2010. ISSN 0003-2638.
- [23] G. Gigerenzer and R. Selten. *Bounded rationality: The adaptive toolbox*. the MIT Press, 2002. ISBN 0262571641.
- [24] J.Y. Halpern and R. Pass. Iterated regret minimization: A new solution concept. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09)*, pages 153–158, 2009.

- [25] G. Harman. *Change in View*. MIT Press, 1986.
- [26] W. Hodges. *A shorter model theory*. Cambridge Univ Pr, 1997. ISBN 0521587131.
- [27] W. Holiday. Epistemic logic and relevant alternatives. In M. Slavkovik, editor, *Proceedings of the 15th Student Session of the European Summer School in Logic, Language, and Information*, pages 4–16, 2010.
- [28] J.F. Horty. *Agency and deontic logic*. Oxford University Press, USA, 2001. ISBN 0195134613.
- [29] R. Jeffrey. *The Logic of Decision*. McGraw-Hill, New-York, 1965.
- [30] J. B. Kadane and P. D. Larkey. Subjective probability and the theory of games. *Management Science*, 28(2):113–120, 1982.
- [31] N. Kolodny. Why be rational? *Mind*, 114(455):509, 2005. ISSN 0026-4423.
- [32] K. Mulligan and F. Correia. Facts. In E. Zalta, editor, *Stanford Encyclopedia of Philosophy (Winter 2011 Edition)*, 2007.
- [33] R. Nozick. *The nature of rationality*. Princeton University Press, 1994. ISBN 0691020965.
- [34] B. O’Shaughnessy. Trying (as the mental "pineal gland"). *The Journal of Philosophy*, 70(13, On Trying and Intending):365–386, Jul. 19 1973.
- [35] R. Parikh. Monotonic and non-monotonic logics of knowledge. *Fundamenta Informaticae*, XV:255–274, 1991.
- [36] D. Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52:1029–1050, 1984.
- [37] Michael Ridge. Reasons for action: Agent-neutral vs. agent-relative. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, CSLI, fall 2008 edition, 2008.
- [38] O. Roy. A dynamic-epistemic hybrid logic for intentions and information changes in strategic games. *Synthese*, 171(2):291–320, 2009. ISSN 0039-7857.
- [39] O. Roy. Interpersonal coordination and epistemic support for intentions with a we-content. *Economics and Philosophy*, 26(3):345–367, 2010.
- [40] A. Rubinstein. The Electronic Mail Game: Strategic Behavior Under "Almost Common Knowledge". *The American Economic Review*, 79(3):385–391, 1989. ISSN 0002-8282.

- [41] Larry Samuelson. Modeling knowledge in economic analysis. *Journal of Economic Literature*, 57:367 – 403, 2005.
- [42] L.J. Savage. *The Foundations of Statistics*. Dover Publications, Inc., New York, 1954.
- [43] Geoff Sayre-McCord. Moral realism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, CSLI, winter 2010 edition, 2010.
- [44] M.A. Schroeder. *Slaves of the Passions*. Oxford University Press, USA, 2007. ISBN 0199299501.
- [45] J.R. Searle. *Intentionality*. Cambridge University Press, 1983.
- [46] M.A. Smith. *The moral problem*. Wiley-Blackwell, 1995. ISBN 0631192468.
- [47] A. Tamminga. Deontic logic for strategic games. Technical report, University of Groningen, 2010.
- [48] Johan van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, 2010.
- [49] H. van Ditmarsch, W. van de Hoek, and B. Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese Library Series*. Springer, 2007.
- [50] Mark van Roojen. Moral cognitivism vs. non-cognitivism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2011 edition, 2011.
- [51] J.D. Velleman. What good is a will? Downloaded from the author’s website on April 5th 2006, April 2003.
- [52] R.J. Wallace. *Normativity and the Will*. Oxford University Press, 2006.
- [53] J. Way. The Normativity of Rationality. *Philosophy Compass*, 5(12):1057–1068, 2010. ISSN 1747-9991.
- [54] B. Williams. *Moral luck: philosophical papers, 1973-1980*. Cambridge Univ Pr, 1981. ISBN 0521286913.