Small steps in dynamics of information
Velazquez Quesada, F.R.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
The non-omniscient and dynamic setting presented in this dissertation adds an extra dimension to standard Dynamic Epistemic Logic: besides the studied acts of observation (update) and revision (upgrade), our agents can get information by means of finer informational actions, like changes in awareness and diverse forms of inference. Thus, our framework can give a different perspective and therefore shed some light in areas that deal, in some form or another, with an agent’s information and its dynamics.

Though in this dissertation we will not pursue particular applications, the present chapter introduces some proposals for connections. We focus on areas in Linguistics and Cognitive Science as well as Game Theory. Our purpose is not to provide formal and deep proposals, but simply to show how the main ideas behind our setting have a wide range of applications.

7.1 Linguistics

Our work can be related with Linguistics, the formal study of natural language. Among all the linguistic areas, there are interesting connections with the study of the notions of attention, questions and pragmatics.

7.1.1 Attention

The interest on the study of a notion of “attention” arises from the observation that, though every day of our life we face an large amount (and probably an infinite number) of possibilities, we, as agents with limited resources, do not have the ability to work properly with all, and at any given moment we only deal with a small subset of them.

There are several ways of defining what an agent is paying attention to, and our approach can represent some of them.
Attention as awareness  Attention can be understood as a language-related notion: we pay attention to the possibilities our current language allows us to express. For example, Natalia is looking for her car’s keys, and she is just paying attention to the possibility for them to be either in the bedroom, or in the kitchen, or in the dinning room. From her point of view, there exists only three possibilities, and she is not considering the possibility for the keys to be in the bathroom because “bathroom” is not in her current language.

More generally, if an agent is only aware of two atomic propositions $p$ and $q$, then she will identify at most four possibilities: the four different combinations of $p$ and $q$’s truth-values. But what she identifies as “the $p$ and $q$ possibility” may correspond to a number of them that differ from each other in the truth-value of atomic propositions the agent is not currently entertaining, like $r$, $s$ and so on. In other words, some possibilities are not considered because the agent’s language is not fine enough to identify them in the first place.

This notion of attention corresponds directly to the notion of awareness we have dealt with in Chapter 4, in which the set of formulas the agent is aware of is generated by the set of atomic propositions she has available in all the worlds she considers possible (Definition 4.6). This gives us the following definition:

$$\text{Att } \varphi := \text{Aw } \varphi$$

Attention given by beliefs  There are others understandings of what it means to be “paying attention” to a given possibility. Following ideas about conscious belief and investigation presented in Stalnaker (1984), the dissertation de Jager (2009) relates the notions of attention and inattention not only to syntactic sources, like the agent’s language, but also to semantic ones, like the agent’s beliefs. For example, Natalia is still looking for her keys, but now she is also not paying attention to the possibility for them to be in the kitchen because she considers that situation very unlikely to be the case.

More generally, even though an agent might be aware of the atomic propositions $p$ and $q$, she might be paying attention just to two possibilities (e.g, the one with $p$ and $q$ true, and the one with $p$ true and $q$ false). Though the other two possibilities are expressible, she is not paying attention to them because, according to her beliefs, they are very implausible.

When beliefs are involved, there are several possibilities for defining what the agent is paying attention to. One option is given by an agent that pays attention to a given $\varphi$ if and only if $\varphi$ is true in at least one epistemically possible situation: \(\text{Att } \varphi := (\sim) \varphi\). But we can also have a more radical agent that pays attention only to those possibilities that occur in at least one world that is more plausible than the current one: \(\text{Att } \varphi := (\leq) \varphi\). We can even have a very radical agent that pays attention only to those possibilities that hold in at least one of the most plausible situations: \(\text{Att } \varphi := [\leq] (\leq) \varphi\).
7.1.2 Questions

There are close relations between our framework and the logical analysis of questions. In particular, we see two important connections.

**Questions as aware raising mechanism** In Chapters 3 and 4 we discussed two different understandings of the notion of awareness: as the formulas of an arbitrary set, and as those generated from the set of atoms the agent can use in all the worlds she considers possible, respectively. Our discussion was not only about awareness’ representation, but also about its dynamics, and we provided mechanisms for raising and dropping awareness.

But, though our actions showed the effect that changes in awareness have in the agent’s information, besides situations like the *Twelve Angry Men* example (Section 4.1), we did not provide concrete reasons for these awareness’ modifications. In other words, though we describe how awareness change, we did not justify why these changes happen.

One of the most natural ways to change the current awareness of an agent and focus her attention on a specific issue is by asking a question. Intuitively, when a question is asked, the hearer switches her attention to focus on the just raised issue. This is the approach followed by some of the most prominent logical treatment of questions (Groenendijk 2007; van Benthem and Minică 2009): a question separates the current set of possibilities into several groups according to the possible answers, therefore changing the agent’s attention.

From our fine grained perspective, we can think of a combination of questions and finer representations of information: we can understand a question as an action that increases some current set of ‘relevant propositions’ whose truth value needs to be determined.

A setting in which we can embed the ideas of our framework in a natural way is the DEL approach to questions of van Benthem and Minică (2009). Semantically, their epistemic issue model contains, besides the non-empty set of words, their atomic valuation and an equivalence indistinguishability relation denoted by ∼, an equivalence abstract issue relation, denoted by ≈, that divides the set of possibilities in areas in which the agent would like to be. Syntactically, the epistemic language is extended with a universal modality for the issue relation, [≈], and the universal modality, U.

The important actions in this framework are not only those that announce a fact (an announcement), but also those that raise an issue (a question). While an announcement “ϕ!” differs from that of standard PAL in that it just cut links between worlds that disagree on ϕ without deleting any of them, the effect of asking a question “ϕ?” is a refinement of the issue relation: each issue partition is split into (possibly) two: one with the worlds that satisfy ϕ, and another with those that satisfy ¬ϕ.
Chapter 7. Connections with other fields

This setting allows us to define statements describing the effect of a question. For example, the formula $U ([\approx] \psi \lor [\approx] \neg \psi)$ expresses that $\psi$ is settled as an issue across the whole model (Definition 4 in van Benthem and Minic˘a (2009)). Then, the following formula states that $\psi$ will be settled as an issue across the whole model after asking $\chi$:

$[\chi?] U ([\approx] \psi \lor [\approx] \neg \psi)$

But, from our non-omniscient perspective, the fact that the truth-value of $\psi$ is uniform in all the agent’s $\approx$-partitions is not enough to settle it as an issue explicitly. Consider, for example, our setting of Chapter 3 in which, in order for $\varphi$ to be explicit information, we needed for the agent to be aware of it, defining $Ex \varphi$ as $\Box (\varphi \land A \varphi)$. Following the same methodology, we can interpret $[\approx]$ as implicit issue, and then define its explicit version as

$$\text{Iss } \varphi := [\approx] (\varphi \land A \varphi)$$

Then, according to the mentioned formula, the following one expresses that after asking $\chi$, $\psi$ will be settled as an explicit issue across the whole model:

$[\chi?] U (\text{Iss } \psi \lor \text{Iss } \neg \psi)$

But now we can look for other versions of the act of asking a question. In van Benthem and Minic˘a (2009)’s omniscient setting, “$\varphi$?” raises an issue not only about $\varphi$, but also about all formulas that are logically equivalent to it. But from our non-omniscient perspective, only the issue about $\varphi$ should be raised explicitly, and the rest of the formulas logically equivalent to $\varphi$ should be indeed an issue, but only in an implicit way.

Following the spirit of the act of explicit observation (Definition 3.7), an explicit question operation that follows the given intuition, denoted by “$\varphi^{+}\,?$”, can be defined as the former “$\varphi$?” plus the additional effect of making the agent aware of $\varphi$. Given the definition of awareness, the latter requirement boils down to adding $\varphi$ to the $A$-set of all possible worlds. Then we can build formulas like the following, expressing that $\psi$ will be settled as an explicit issue across the whole model after asking $\chi$ explicitly:

$[\chi^{+}\,?] U (\text{Iss } \psi \lor \text{Iss } \neg \psi)$

This new setting, together with the actions defined in Section 3.5, allow us to express combinations of questions and changes in awareness. For example, the following formula expresses that, after asking $\chi$ explicitly, $\psi$ will become an issue settled implicitly across the whole model, and it will be an issue settled explicitly as soon as the agent considers it:

$[\chi^{+}\,?] \left( U ([\approx] \psi \lor [\approx] \neg \psi) \land [+\psi] U (\text{Iss } \psi \lor \text{Iss } \neg \psi) \right)$
These examples show how a question can be understood as a mechanism that raises issues, and therefore creates awareness. Such changes can affect not only the agent’s explicit knowledge (and beliefs), but also other attitudes, like preferences (Guo and Xiong 2010).

Questions and inferences The second strong connection is not with acts of awareness change, but with acts of inference. Most of the scientific inquiry can be described as a combination of questions and inferences, like Hintikka emphasizes in his Interrogative Model of Inquiry (IMI: Hintikka (1999); Hintikka et al. (2002); Hintikka (2007)). In the IMI, inquiry is represented as an information-seeking process in which the inquirer, based on some premises, tries to establish certain conclusion. At each stage, she has a choice between performing a deductive step in which a logical conclusion is derived from the information she has acquired so far, or performing an interrogative move in which she test a fact that she cannot justify or discard with her current information.

In the search for formalizations of the IMI, combinations of frameworks for questions and inference have already produced fruitful results. The master’s dissertation Hamami (2010a) combines a logic for questions with a logic for tableau-based inference, and the inference part shares some similarities with our approach for rule-based inference of Chapter 2, like the definition of explicit knowledge (called local knowledge) and the restriction for explicit information about only propositional formulas. On top of that, the system has the important advantage of providing to the agent a complete reasoning system.

For a simple example of a useful combination of questions and inference, recall the described DEL approach to questions (van Benthem and Minică 2009). Another important action defined there is the action of resolution, “!”, in which the indistinguishability relation ∼ is restricted within the issue partitions (that is, is redefined as ∼ ∩ ≈). As indicated in the mentioned work, this operation is a natural generalization of an announcement that need not have natural language correspondent. With such operation we can build formulas like

\[[\chi]?! U (\sim \psi \lor \sim \neg \psi)\]

expressing that a question about \(\chi\) followed by a resolution will produce knowledge everywhere about whether \(\psi\).

In a non-omniscient setting, a question and a resolution may not be enough to produce explicit knowledge. Even if an implicit question of \(\chi\) followed by resolution produce indeed implicit knowledge about \(\psi\), there is no guarantee that \(\psi\) will be also explicitly known. But then, all the agent needs is a further inference step. So suppose that, indeed, \([\chi]?! U (\sim \psi \lor \sim \neg \psi)\) is the case. Then, we expect the following formula to be true too:

\[[\chi^+]?! [\sim (\chi \Rightarrow \psi)] U (k_{Ex}\psi \lor K_{Ex}\psi)\]
The formula expresses that an explicit question about $\chi$ followed by a resolution and then an inference step will produce explicit knowledge about $\psi$.

The combination of questions and inferences become even more appealing when we look not only at the inquirer’s knowledge and her deductive inferences, but also at her beliefs and her inferences in general.

### 7.1.3 Pragmatics

Suppose your partner tells you truthfully “I’m cooking meat tonight”. What information is conveyed by this announcement?

Besides the plain fact that your partner indeed will be cooking meat tonight, the message usually provides more information. Depending on the specific circumstances, it may also indicate “bring red wine”, “do not be late” or, in some extreme cases, “do not show up at all”. These pieces of information, despite being beyond the proper meaning of the announcement, are usually understood and acknowledged in our conversations.

Where does this extra information (implicatures) comes from? Why is it communicated? What is the role of the proper semantic meaning of the announced sentence in the extra information it provides? These questions are the concern of linguistic *Pragmatics*.

One of the most influential pragmatic theories is the one introduced by Paul Grice (see Grice (1989)). The main idea of his proposal is that, based on the assumption that the speaker obey certain ‘maxims’ about the informative purpose of a conversation, the hearer can extract additional information that is not covered by the semantic meaning of the statement. In other words, Grice proposed that conversational implicatures can be seen as further inferences, and that they can be justified by a reasoning process that takes into account not only the semantic meaning of the announced sentence, but also some aspects of the conversational context.

With this idea in mind, and following the methodology of our approach, implicatures can be seen as the result of further inference steps based on beliefs the hearer has about the speaker’s intentions. Consider the mentioned example: a speaker announces “I’m cooking meat tonight”, and from this the hearer infers that she needs to get red wine. How can this be represented in our setting?

First, note that the assumptions the hearer made about the drinking preferences of the speaker (white wine when cooking fish, red wine when cooking meat) are already encoded in the hearer’s beliefs before the announcement takes place; it is in this sense that the hearer makes assumptions about the speaker’s intentions during a conversation. This situation corresponds to the following model, in which $m$ stands for meat (hence $\neg m$ stands for fish) and $r$ stands for red wine ($\neg r$ stands for white wine). The arrows represent the plausibility relation, with the reflexive arcs omitted.
7.1. Linguistics

The hearer’s plausibility order puts on top the meat-red wine and fish-white wine situations. Nevertheless, this is only implicit; the only explicit information the hearer has is about the food’s choice in each possible situation, and about what drink it would imply in each one of them.

Then the speaker says ‘I’m cooking meat tonight”. The immediate effect of the utterance is that of a public announcement: the hearer will discard those situations she recognizes as ¬m-ones, i.e., the situations on the right column (see Section 4.4.2). This gives us the following model, which we call M:

Here is where the further reasoning takes place. The agent believes explicitly that meat corresponds to red wine, so she can perform an inference that will make her implicit belief about the red wine explicit. In our setting, she has at least two ways of doing it. The first one, a strong local inference with m ⇒ r (Definition 5.18) will only make explicit the red wine belief (BEx r):

The second possibility, an inference with known premises m and believed rule m ⇒ r (Definition 5.15), will additionally acknowledge explicitly a (not plausible but still possible) situation with white wine (BEx r ∧ KEx ¬r):
Pragmatics as iterated best response  The *iterated best response* mechanism proposed in Franke (2009) explains pragmatic phenomena as the result of a sequence of iterated best responses that start from the literal semantic meaning of the announcement and continue for as long as it is reasonable and the agents can do it. We have shown how our setting can capture the small steps that makes explicit the assumptions in each response, and our non-omniscient belief revision (Section 5.3.2) can represent how the agents ‘correct’ their beliefs for future responses. To describe the long-term behaviour, an extension that deal with iterations is needed, as we will discuss in Chapter 8.

7.2 Cognitive Science

Our framework deals with the representation of several notions of information and the way they are affected by diverse actions. Thus, it also has connections with Cognitive Science, the study of mind and how information (perception, language, reasoning, and emotion) is represented and transformed in the brain.

7.2.1 Learning Theory

Leaving aside for a moment the particular frameworks and tools that we have explored so far, the main subject of this dissertation is the study and representation of changes in information. Besides Logic (and, in particular, Epistemic Logic and its dynamic extensions), there are other approaches that deal with epistemic changes. One of these frameworks is *Learning Theory* (LT; see e.g. Jain et al. (1999)), the study of functions that attempt to identify the correct hypothesis from a collection of possibilities based on inductively given streams of data. Though it evolved from the study of language acquisition, learning theory focuses on various properties of the process of conjecture-change over time, and therefore it is applicable also in other fields, like philosophy of science, where it can be interpreted as a theory of empirical inquiry (Kelly 1996).
Using the language learning terminology, the basic premises in Learning Theory are that the agent considers several languages as the possible ones. Then she receives an infinite sequence of data that contains all words of the actual language. Based on this information, the agent tries to identify the actual language, in some cases by making a single conjecture after a finite number of data (finite identification; Mukouchi (1992)), and in some others by making a conjecture every certain time and waiting for the conjecture to stabilize to the correct one (inductive inference; Gold (1967); Angluin and Smith (1983)).

Recent works have looked at connections between Learning Theory and DEL (Gierasimczuk 2009; Ma 2009; Baltag and Smets 2009; Dégremont and Gierasimczuk 2009). The main idea is that, by representing the languages the agent considers possible in a possible worlds style, learning can be seen as a mechanism that, at each stage, decides how the just received data will change the agent’s knowledge/beliefs.

Our fine-grained setting allows us to represent this process from the perspective of non-ideal agents. First, a ‘real’ agent may not have at hand the full language each possibility represents. More precisely, our A-sets that so far have contained the formulas the agent has acknowledged as true in that world, can now contain the words the agent has recognized as part of the language represented by that world. For example, if the agent considers only two possibilities, one standing for the language \(a(a + b)^*\) and another standing for the language \(a(a + b)^*a\), then she knows implicitly that the word \(aba\) is in the language because it is in the two languages she considers possible. But, following our definition of explicit knowledge of Chapters 4 and 5, the knowledge is not explicit if she has not recognized the word in the two possibilities. This can be expressed with the following formula:

\[
K_{\text{Im}}(aba) \land \neg K_{\text{Ex}}(aba)
\]

Second. Though a ‘real’ agent does not need to have the full language each possibility represents, she may as well be able to construct words of it. The agent can have information not only about the words of each language, but also about how to generate more words from current ones. Following the ideas of Formal Grammar, one way to do this is by using production rules: then we can express situations in which, thanks to her current knowledge, the agent can derive another word of the language:

\[
(K_{\text{Ex}}(abX) \land K_{\text{Ex}}(X \Rightarrow a)) \rightarrow \left[(X \Rightarrow a)\right] K_{\text{Ex}}(aba)
\]

Third. Adding beliefs to the picture enriches the setting, allowing us to representing the agent’s implicit and explicit hypothesis about the actual language at each stage and how it changes due to the received piece of information (Baltag and Smets 2009; Dégremont and Gierasimczuk 2009). In our non-omniscient
Chapter 7. Connections with other fields

In this chapter, the following formula expresses that after receiving the explicit information that \( abaa \) is a word in the language, the agent will believe that \( abaaa \) is a word too:

\[
\neg B_{Ex}(abaa) \land [abaa^+ \Box] B_{Ex}(abaaa)
\]

Finally, our setting for inferences involving beliefs allows us to represent small belief changes that are not direct consequence of the received data (just like conversational implicatures). For example, if the agent believes explicitly that \( X \Rightarrow b \) is a production rule of the actual language, then she can use it to generate a new explicit belief. This is expressed by the following formula in which the \( PA \) action model \( (C_{KB}^{X=b}, e) \) is the one of Definition 5.15.

\[
(K_{Ex}(abX) \land B_{Ex}(X \Rightarrow b)) \rightarrow \langle C_{KB}^{X=b}, e \rangle (B_{Ex}(abb) \land \neg K_{Ex}(abb))
\]

### 7.2.2 The notion of surprise

**Surprising observations** Many belief dynamics, like abduction and belief revision, are related to the notion of surprise, and there are some formal approaches to this concept. In particular, the framework of Lorini and Castelfranchi (2007) investigates the role of surprise in triggering the process of belief change, and distinguishes two main forms of surprise: mismatch-based surprise and astonishment. While the first one appears when the agent perceives information that contradicts the beliefs she is currently focusing on, the second one appears when the agent perceives information that is not in the focus of the agent. The latter has two variants: after the agent brings the topic into focus, she recognizes that she did not expect the observation, or even worst, that she expected the opposite of the observation.

These two notions of surprise can be represented in our framework. Note that the key difference between the two notions is the focus of the agent, so to get a proper representation, we need to incorporate the awareness of notion (we will use that of Chapter 4) to the beliefs framework of Chapter 5. Assume an extended definition of implicit and explicit beliefs of the following form:

The agent believes implicitly the formula \( \varphi \)  
\[
B_{Im} \varphi := \langle \varepsilon \rangle \Box (I \varphi \land \varphi)
\]

The agent believes explicitly the formula \( \varphi \)  
\[
B_{Ex} \varphi := \langle \varepsilon \rangle \Box (I \varphi \land \varphi \land A \varphi)
\]

We will use the modality \( \langle \varphi^+ \Box \rangle \) for the finer non-omniscient observation (i.e., an announcement with unspecified announcer) sketched in Section 4.4.2.

The first form of surprise, mismatch-based surprise, occurs when the agent faces an observation that contradicts the beliefs she is currently focusing on. Our notion of explicit beliefs already requires the agent’s attention, so in our
setting this situation corresponds roughly to the following formula

\[ B_{\text{Ex}} \neg \chi \land \langle \chi^+ \rangle \top \]

The second form, *astonishment*, occurs when the agent faces an observation that is not in her current focus and, after bringing into focus the related information, she recognizes that either she did not expect the observation, or else she expected exactly the opposite. This situation corresponds to the formula

\[ \neg Aw \chi \land \left( \neg \langle \varepsilon \rangle [\varepsilon] (\chi \land A \chi) \lor \langle \varepsilon \rangle [\varepsilon] (\neg \chi \land A \neg \chi) \right) \land \langle \chi^+ \rangle \top \]

where \( \langle \varepsilon \rangle [\varepsilon] (q \land A q) \) stands for beliefs that just need the agent’s attention (i.e., awareness) to become explicit (cf. Subsection 4.3.2).

But our system can also express situations in which the agent can perceive surprises that happen not only at the explicit level, but also at the implicit one. Such surprises are stronger because what fails is not the agent’s ability to make explicit her implicit beliefs (that is, the surprise does not arise because of lack of reasoning), but rather her plausibility order.

**Other actions producing surprise** We have discussed surprises that occur as a result of observations. But, are there other actions that can produce surprise?

For simplicity, we go back to the awareness-less definitions of implicit and explicit beliefs of Chapter 5. We will say that the formula \( \chi \) is a *weak explicit* surprise if and only if the agent does not believe it explicitly: \( \neg B_{\text{Ex}} \chi \); we will say that \( \chi \) is a *strong explicit* surprise if and only if the agent believes \( \neg \chi \) explicitly: \( B_{\text{Ex}} \neg \chi \). Correspondent notions of implicit weak and strong surprise can be obtained by replacing \( B_{\text{Ex}} \) by \( B_{\text{Im}} \) in the previous definitions. The forms of surprise that can be produced by each one of the actions in our setting depend on what each action needs to take place.

Consider first our knowledge-related actions. In order for the agent to observe some \( \chi \) she does not need any previous information, so an observation can produce not only the forms of surprise we just defined, but also many others. In the case of knowledge-based inference with a rule \( \sigma \), a weak explicit surprise can be produced, since the agent does not need to believe explicitly the rule’s conclusion before applying the rule. But none of the other forms of surprise is possible, and the reason is that in order for the inference to take place, \( \text{cn}(\sigma) \) should be already implicit knowledge, and therefore implicit belief. Then weak implicit surprise is not possible because it asks for \( \text{cn}(\sigma) \) not to be

\footnote{In fact, \( \langle \chi^+ \rangle \top \) does not express that \( \chi \) is actually observed; just that it can be observed. Then, the whole formula does not state that the agent is surprised, but rather than she can be surprised. An alternative definition would take as the evaluation point not the stage before the observation, but the stage after it. Nevertheless, it would need a past-looking modality to express what the agent believed before the observation (cf. Yap 2007).}

\footnote{Again, the formula only states that the agent can be surprised.}
implicitly believed, and the two strong cases are also not possible because in both the agent would need to believe \( \neg \text{on}(o) \) implicitly, and our setting does not allow inconsistent beliefs.

Now consider our belief-related actions. Arbitrary acts of revision (i.e., arbitrary rearrangement of beliefs) can definitely produce surprises, again because there is no precondition attached to them. Whether a non-arbitrary rearrangement can produce surprising information or not depends on what the agent needs to perform the action. For example, our inference with known premises and believed rule aims to produce a situation in which the agent believes explicitly the rule’s conclusion (and yet considers explicitly a possibility in which it fails). Then, the requirements imply that the rule’s conclusion is already implicitly believed, and therefore, though the action can produce weak explicit surprise, it cannot produce the weak implicit or any of the strong versions. But this is reasonable, because as we discuss, this kind of inference resembles default reasoning, an inference that works on what is most likely to be the case.

Other forms of inferences involving beliefs can produce surprises. The already discussed abductive reasoning requires from the agent some information about the rule and its conclusion, but since this does not imply any attitude about the premises, surprises are possible. Nevertheless, recall that the goal of abductive reasoning is to find the ‘best’ an explanation of a given \( \chi \). Though an explanation can be definitely surprising, there are usually several candidates, and choosing the ‘best’ is generally understood as choosing the one that will produce the less amount of changes or, in other words, the one that is less surprising (and even not surprising at all). In fact, the notion of ‘best’ explanation could be tied to how surprising this explanation would be.

### 7.3 Game Theory

Game Theory analyzes competitive situations in which several agents have to make a (sequence of) choice(s). The outcome of the situation is then determined by the individual choices of each one of them. A typical example is the so called centipede game, whose diagram is shown below.

![Centipede Game Diagram](image)

This game takes place between two players, \( a \) and \( b \), with 1 the starting point and 2, 4 and 5 the final ones. At points 1 and 3, an agent has to take a decision (\( a \) and \( b \), respectively), and the final payoffs of the game, indicated in the form (player \( a \), player \( b \)), are determined by the players’ choices.
One of the main goals of *Game Theory* is the study of *solution concepts*: formal rules that indicate optimal strategies for the game, and therefore predict the (not necessarily unique) final outcome. For example, *backward induction* predicts that in the centipede game of above, if point 3 were reached, *b* would choose 4 instead of 5, getting 100 instead of 99, and leaving *a* with 0 instead of 99. But if *a* recognizes this, then she will realize that choosing between 2 and 3 actually means to choose between a payoff of 1 and a payoff of 0, respectively. Then she will choose 2, and the final payoff of the game will be (1,1).

In order to indicate what each player will do, a solution concept needs to make assumptions not only about the nature of the game (perfect/imperfect information, strategic/extensive game, etc.), but also about the nature of the involved players. Recent literature (Aumann (1995); Stalnaker (1996); Aumann and Brandenburger (1995); Polak (1999); Chen et al. (2007) among others) has looked at the epistemic conditions that players need to satisfy in order to follow the solution concept’s specification, and some of them have used *Epistemic Logic* and *Dynamic Epistemic Logic* tools to make formal these epistemic requirements. It turns out that the assumption of rationality that some of the most important solution concepts in the literature make (the mentioned *backward induction*, *Nash equilibrium*, *iterated elimination of strictly dominated strategies* among others) implies not only that every player will always make the choice that will give her the best possible outcome, but also that the players are always able to perfectly calculate every single consequence of every action. Not surprisingly, the predictions of solution concepts based on such strong assumptions usually do not coincide with the choices real agents do when facing these situations. Even in approaches that model games with incomplete information (e.g., Feinberg (2004)), it is implicitly assume that the players can derive all logical consequences of the information they have, and this is not necessarily the case.

Our framework allows us to model situations in which the involved agents are non-omniscient, and therefore allows us to explain why non-ideal players do not necessarily behave in an optimal way. Consider again the presented centipede game, and suppose *a* knows explicitly not only the structure of the game but also her preferences about the final state. Moreover, suppose that she believes explicitly that, if the game reaches point 3, *b* will choose 4. These assumptions can be expressed with the following formulas:

<table>
<thead>
<tr>
<th>Game’s structure</th>
<th>Final state’s preferences</th>
<th>Rationality</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K^a_{Ex} (1 \rightarrow [\text{Next}] (2 \lor 3))$</td>
<td>$K^a_{Ex} (2 \rightarrow \text{Neutral}_a)$</td>
<td>$B^a_{Ex} (3 \rightarrow [\text{Next}] 4)$</td>
</tr>
<tr>
<td>$K^a_{Ex} (4 \rightarrow \text{Worst}_a)$</td>
<td>$K^a_{Ex} (5 \rightarrow \text{Best}_a)$</td>
<td></td>
</tr>
</tbody>
</table>
The game starts, so \( K^a_{\text{Ex}} 1 \) is the case. By deduction, \( a \) knows explicitly that she has a choice between 2 and 3, that is, \( K^a_{\text{Ex}} [\text{Next}](2 \lor 3) \). What will she do?

The important point here is that a non-omniscient agent may have not noticed the connection between her current choice, her explicit belief about \( b \)'s choice at 3, and the payoffs in each case. In other words, she may have not realized that if she chooses 3, the game is very likely to end with 4 as the final state. In order to link those pieces of information, she needs to apply the following truth-preserving rule

\[
\{ [\text{Next}](2 \lor 3), 3 \rightarrow [\text{Next}](\text{Neutral}_a) \} \Rightarrow [\text{Next}](\text{Neutral}_a \lor [\text{Next}]\text{Worst}_a)
\]

While the first premise, \( [\text{Next}](2 \lor 3) \), is something \( a \) knows (explicitly), the second one, \( 3 \rightarrow [\text{Next}]\text{4} \), is something she only believes (explicitly). Then, after the inference step, \( a \) will only believe (explicitly) the conclusion, that is,

\[
B^a_{\text{Ex}}[\text{Next}](2 \lor [\text{Next}]\text{4})
\]

But again, being non-omniscient, she may still need to link the final states with her preference about them, that is, she may need to apply

\[
\{ [\text{Next}](2 \lor [\text{Next}]\text{4}), (2 \rightarrow \text{Neutral}_a), (4 \rightarrow \text{Worst}_a) \} \Rightarrow [\text{Next}](\text{Neutral}_a \lor [\text{Next}]\text{Worst}_a)
\]

Again, the conclusion of this rule will be only believed (explicitly), that is,

\[
B^a_{\text{Ex}}[\text{Next}](\text{Neutral}_a \lor [\text{Next}]\text{Worst}_a)
\]

Only after these two inference steps \( a \) will realize that, according to her beliefs, her choice between 2 and 3 actually boils down to a choice between \( \text{Neutral}_a \) and a future \( \text{Worst}_a \).

Though the example makes some simplifications, it definitely highlights one of the main reasons why real agents might not choose the solution backward induction proposes: even if they have full knowledge about the structure and payoffs of the game and even if they believe they all will pick the highest payoff when having the choice, they might fail in establishing a direct relation between early moves in the game and later outcomes. In these cases, our non-omniscient analysis allows us to model not only the information these non-ideal agents have, but also the reasoning steps they need in order to reach an information state in which the strategy proposed by the solution concept will actually be played.

---

3Other explanation is a common agreement to reach an outcome that is better for both agents (5 in our example).