Synthese (2012) 185:27–41
Knowledge, Rationality & Action 357–371
DOI 10.1007/s11229-011-9957-x

The psychology of reasoning about preferences and unconsequential decisions

Jean-François Bonnefon · Vittorio Girotto · Paolo Legrenzi

Received: 14 April 2011 / Accepted: 27 April 2011 / Published online: 17 May 2011

Abstract People can reason about the preferences of other agents, and predict their behavior based on these preferences. Surprisingly, the psychology of reasoning has long neglected this fact, and focused instead on disinterested inferences, of which preferences are neither an input nor an output. This exclusive focus is untenable, though, as there is mounting evidence that reasoners take into account the preferences of others, at the expense of logic when logic and preferences point to different conclusions. This article summarizes the most recent account of how reasoners predict the behavior and attitude of other agents based on conditional rules describing actions and their consequences, and reports new experimental data about which assumptions reasoners retract when their predictions based on preferences turn out to be false.

Keywords Reasoning · Preferences · Belief revision · Experiment

J.-F. Bonnefon (✉)
Cognition, Langues, Langage, Ergonomie, Centre National de la Recherche Scientifique, Université de Toulouse, Toulouse, France
e-mail: bonnefon@univ-tlse2.fr

V. Girotto
Laboratoire de Psychologie Cognitive, Centre National de la Recherche Scientifique, Université de Provence, Marseille 1, France

V. Girotto
Universitá IUAV di Venezia, Venezia, Italy
e-mail: vittorio.girotto@iuav.it

P. Legrenzi
Department of Philosophy and Cultural Heritage, University Ca' Foscari of Venice, Dorsoduro, 3484/D, 30123 Venice, Italy
e-mail: paolo.legrenzi@gmail.com

1 Introduction

Alice has hired Bob to do a job for her, which Bob is due to complete in 2 days. Alice knows that she could ask Bob to work faster and complete the job in one day, but that would cost her 300 euros extra, and she is not in a hurry. Is Alice going to ask Bob to work faster? Most of us would immediately reason that she will not, as this would be a poor decision for her to make. In this article, we address the question of the psychological processes underlying this inference: How do people predict what other agents will do, based on their perceived beliefs and preferences?

Although this topic is not new in the Artificial Intelligence and multiagent systems literature (e.g., Bex et al. 2009; Ficici and Pfeer 2008), we explain in Sect. 2 that inferences of that sort are still a recent concern for the psychology of reasoning, which has long focused on 'disinterested' situations, that is, situations where preferences are either unknown or ignored. There is mounting evidence, though, that reasoners take into account the preferences of others, at the expense of logic when logic and preferences point to different conclusions.

In Sect. 3, we introduce the most recent account of how people reason about the decisions of others, based on conditional rules that describe actions and their valued consequences (Bonnefon 2009). In Sect. 4, we extend this framework to the problem of belief revision following unconsequential behavior. We consider situations where reasoners reached a prediction about what another agent would do (based on her perceived beliefs and preferences), only to discover that this agent did not make the decision they predicted. We focus on the kind of belief revision these reasoners are going to engage in. This is an important issue at the intersection of interested reasoning and belief revision, on which the psychology of reasoning has been silent so far. When reasoners discover that another agent did not take the action that they anticipated, are they going to revise what they thought the other agent knew, or what they thought the other agent wanted, or maybe both? To answer these questions, we set up an experiment whose method and results we report in Sect. 5.

2 Farewell to disinterested reasoning

For most of its existence, the psychology of reasoning aimed at assessing the deductive competence of lay reasoners (Evans 2002). By presenting people with a vast array of deduction problems, it was possible to identify the kind of logical arguments that people could endorse easily, and to contrast them with the kind that they found very hard to endorse. Theories of human reasoning were constructed on this empirical basis, with the goal of explaining why some problems were hard and others were easy in terms of the cognitive processing that they required from human reasoners (Braine 1978; Johnson-Laird and Byrne 1991; Rips 1983).

This focus on deductive competence resulted in what is called a 'disinterested' approach to reasoning,[1] which assumed goals and preferences to be mostly irrelevant

[1] The distinction between 'interested' and 'disinterested' approaches to reasoning was made in Chater et al. (1998), and it was discussed extensively in Oaksford et al. (1999) and Oaksford and Chater (2003).

to the scientific objectives pursued by the psychology of reasoning. The historical objective of the psychology of reasoning was to assess how good people were at making logical deductions: Accordingly, the only aspects of a problem that warranted attention were those that changed its deductive conclusions. As long as goals and preferences did not affect the deductive conclusions of a problem, it seemed that they could be ignored. As a consequence, and even though psychologists recognized that logical deductions could be the exception rather than the rule in everyday life (Wason and Johnson-Laird 1972, p. 66) the psychology of reasoning was long oblivious to the fact that reasoners could take preferences into account when calculating conclusions.

Empirical data eventually showed that this attitude was not tenable. Consider for example the following problem:

(1) a. If Ally is invited to the party, she'll buy a new dress;
b. Ally is invited to the party.

It follows deductively from (1) that Ally will buy a new dress, and the thousands people who were tested on problems like (1) almost unanimously agreed (Evans et al. 1993). This deductive conclusion does not depend on any assumption about Ally's goals and preferences. That Ally might like to be invited to the party, or that she might enjoy buying a new dress, is irrelevant to the fact that Ally buying a new dress is a deductive conclusion of (1). As a consequence, theories of reasoning dispensed from introducing such preferences in their models of how reasoners handled problems such as (1). But then compare (1) to (2):

(2) a. If Ally is invited to the party, she'll buy a new dress;
b. If Ally buys a new dress, she can't pay the rent next week;
c. Ally is invited to the party.

It still follows deductively that Ally will buy a new dress. This time, however, people tend to reject that conclusion (Bonnefon and Hilton 2004). They do so because they assume that Ally does wish to pay her rent, and would not take an action that would make it impossible to pay the rent. In other terms, people abandon the conclusion that deductive logic would sanction, because this conclusion is inconsistent with their assumption that Ally is a rational agent who would not act against her preferences.[2] This and other results (Evans et al. 2008; Ohm and Thompson 2006) show that reasoners take into account goals and preferences when they calculate conclusions, and that they do so to the expense of deductive logic, when logic and preferences point to different conclusions. Exactly how this effect works is the topic of the next section.

[2] People could also have other reasons not to endorse the deductively valid conclusion, which would not depend on the presence of (2-b). For example, they could consider that (2-a) is a default rule rather than a strict implication, and then think of situations in which Ally would not be in the capacity to buy a new dress, even though she is invited to the party. This phenomenon has been well documented in reasoning experiments (e.g. Byrne 1989; De Neys et al. 2003; Politzer and Bonnefon 2006), and has encouraged some psychologists to adopt nonmonotonic logics as a theoretical framework for the study human reasoning (e.g., Benferhat et al. 2005; Stenning and van Lambalgen 2005).

3 Utility conditionals

The critical feature of the Ally example is an *if…then* sentence: 'if Ally buys a new dress then she can't pay the rent next week'. This sentence is a *utility conditional* (Bonnefon 2009), which triggers special inferences based on the presumed preferences of Ally. In general terms, utility conditionals are 'if p then q' sentences such that the occurrence of p, or the occurrence or q (or both) have significant utility for some agents. That is, some agents care about p happening or not, or about q happening or not. Typical utility conditionals describe actions and their consequences:

(3) a. If she eat oysters then she will be very sick;
b. If you testify against me then you will have an accident;
c. If I stop by his office he'll be happy.

In all these examples, an action is described whose consequences matter for some agent. Presumably, 'she' in (3-a) prefers not being sick, the hearer of (3-b) prefers not having an accident, and 'he' in (3-c) would like being happy. According to the disinterested psychology of reasoning, these preferences do not bear on the inferences that reasoners are willing to make from the conditional sentences in (3). These preferences, however, do appear to trigger a variety of inferences, such as:

(4) a. She is not going to eat oysters;
b. The speaker thinks the hearer should not testify, and in fact the hearer is not going to testify;
c. The speaker is going to stop by 'his' office.

Bonnefon (2009) offered a theory of utility conditionals such as (3), and of the inferences such as (4) that people draw from them. The theory consists of a representational tool (the utility grid) that summarizes in compact form the preferences involved in the conditional, and of a set of folk axioms of decision that captures reasoners' likely beliefs about the way most agents make their decisions. To predict the preference-based inferences triggered by a conditional, one applies the folk axioms of decision to the utility grid of the conditional. In the rest of this section, we summarize the key points of the theory, by explaining the language of utility grids, and showing how folk axioms of decision apply to these utility grids.

3.1 Utility grids

Utility grids are a tool to represent the decision-theoretic features of conditional statements, that is, the aspects of a conditional that relate to actions, consequences, and preferences over these actions and consequences. They subsume a number of prior characterizations of conditional contracts or inducements (Amgoud et al. 2007; Beller et al. 2005; Evans 2005; Legrenzi et al. 1996; López-Rousseau and Ketelaar 2004, 2006). The utility grid of an 'if p, then q' conditional statement has the following general form:

$$\left\{ \begin{array}{ccc} x & u & y \\ x' & u' & y' \end{array} \right\}.$$

The first row of the grid contains the information related to the *if*-clause of the conditional. That is, it displays the agent x (left column) who can potentially take action p, and the utility u (central column) that this action would have for a given agent y (right column). The second row of the grid contains the corresponding information with respect to the *then*-clause of the conditional. That is, it displays the agent x' (left column) who can potentially take action q, and the utility u' (central column) that this action would have for a given agent y' (right column).

The set of all agents is $\mathcal{A}$. By convention, the agent who asserts the conditional (if any) is s (for 'speaker'), the agent whom the conditional is asserted to (if any) is h (for 'hearer'), and e (for 'someone else') is an agent who is neither the speaker nor the hearer. When p or q is not an action that can be taken by an intentional agent but is rather an event or a state of the world, it is noted as being undertaken by a special, neutral agent ω. The agent ω can be thought as 'the world' or the body of laws that govern the world.

To simplify matters, utility is usually reduced in the grid to its sign: u and u' take their values from $\{-, 0, +\}$, where $-$ and $+$ respectively stand for any significantly negative and positive values. Note that $u = 0$ means that action p is not known to have any utility for any agent. By convention, such an action has zero utility and the whole set of agents $\mathcal{A}$ as a target.

As an illustration, we can construct the utility grids of the three conditionals in (3):

$$\left\{ \begin{array}{ccc} e & 0 & \mathcal{A} \\ \omega & - & e \end{array} \right\},$$

$$\left\{ \begin{array}{ccc} h & - & s \\ \omega & - & h \end{array} \right\},$$

$$\left\{ \begin{array}{ccc} s & 0 & \mathcal{A} \\ \omega & + & e \end{array} \right\}.$$

Let us consider the first grid, which corresponds to the conditional 'If she eats oysters, then she will be very sick.' The first row represents the *if*-clause of this conditional. Eating oysters is an action of an agent e who is neither the speaker nor the hearer of the statement, and we have no information as to whether this action is preferred or not preferred by anyone: This action is therefore represented by the triple $\langle e, 0, \mathcal{A} \rangle$. The second row of the grid represents the *then*-clause of the conditional. 'She will be very sick' is a state of the world that presumably go against the preferences of the female agent e, it is therefore represented by the triple $\langle \omega, -, e \rangle$.

To illustrate further, let us consider the second grid, which corresponds to the conditional 'If you testify against me then you will have an accident'. Testifying against the speaker is an action of the hearer that presumably goes against the preferences of the speaker. Accordingly, this *if*-clause is represented in the first row of the grid by the triple $\langle h, -, s \rangle$. Having an accident is a state of the world that goes against the

preferences of the hearer. Accordingly, this *then*-clause is represented in the second row of the grid by the triple $\langle \omega, -, h \rangle$.

Finally, the third grid corresponds to the conditional 'If I stop by his office then he'll be happy'. Stopping by 'his' office is an action of the speaker that, by itself, dos not have clear utility to anyone. Accordingly, this *if*-clause is represented in the first row of the grid by the triple $\langle s, 0, \mathcal{A} \rangle$. 'His' being happy is a state of the world that likely fits the preferences of the unidentified male agent. Accordingly, this *then*-clause is represented in the second row of the grid by the triple $\langle \omega, +, e \rangle$.

3.2 Folk axioms of decision

Once a conditional is turned into a utility grid, the theory predicts the inferences it invites by defining folk axioms of decision, which capture the naive understanding that people have of the way other agents make their decisions (Miller 1999; Smedslund 1997). The folk axiom of Self-Interested Behavior, for example, simply states that reasoners believe other agents to be guided by self-interest:

Folk Axiom 1 (Self-Interested Behavior) Reasoners believe that agents take actions that increase their own personal utility, and that they do not take actions that decrease their own personal utility.

As long as reasoners use this naive axiom to predict the actions of other agents, then their inferences can be anticipated by looking for specific configurations of the utility grid. Consider for example this general configuration, where the black dot stands for any legitimate value of the corresponding parameter:

$$\left\{ \begin{array}{ccc} x & \bullet & \bullet \\ \bullet & - & x \end{array} \right\}.$$

The theory predicts that any conditional whose utility grid fits this configuration triggers the inference that agent x will not take action p, by application of the folk axiom of Self-Interested Behavior (Bonnefon 2009; Bonnefon and Hilton 2004; Bonnefon and Sloman 2011; Evans et al. 2008; Ohm and Thompson 2004). Consider again the utility grid of the conditional 'if you testify against me, then you will have an accident':

$$\left\{ \begin{array}{ccc} h & - & s \\ \omega & - & h \end{array} \right\}.$$

This utility grid fits the configuration that triggers the folk axiom of Self-Interested Behavior. Accordingly, the theory predicts that the conditional sentence invites the inference that the speaker is not going to testify against the hearer.

This is not the only inference that the theory predicts from this specific conditional. Indeed, the utility grid above also fits the configuration that triggers another folk axiom, that of Self-interested attitude:

Folk Axiom 2 (Self-Interested Attitude) According to reasoners, an agent thinks that actions which increase his or her own personal utility should be taken by others, when

these other agents can take these actions (and, mutatis mutandis, that an agents thinks that actions which decrease his or her own personal utility should not be taken by others, when these other agents can take these actions).

Various configurations of the utility grid trigger the folk axiom of Self-Interested attitude, among which:

$$\begin{Bmatrix} x & - & y \\ \bullet & \bullet & \bullet \end{Bmatrix},$$

where x and y are two different agents in $\mathcal{A} \setminus \omega$, and the black dot stands for any legitimate value of the parameter. Whenever a conditional has a utility grid that fits this configuration, the theory predicts the inference that agent y thinks that agent x should not take action p (Bonnefon and Sloman 2011; Thompson et al. 2005). In the current example, the theory predicts the inference that the speaker thinks the hearer should not testify against him.

In this article, we limit our presentation to these basic features of the theory, which is fully developed in Bonnefon (2009). The full description of the theory offers more folk axioms of decision, and addresses in particular the issue of utility grids that invite conflicting conclusions, and the issue of augmenting the theory with uncertainty and finer-grained utility. In the rest of this article, we turn to the unexplored issue of what happens when individuals do make the inferences predicted by the theory, only to find them falsified by the facts. What adjustments do reasoners make to their beliefs when other agents make decisions that violate their folk theory of decision?

4 Unconsequential decisions

Let us get back to our introductory example: if Alice asks Bob to work faster, Bob will charge her an additional fee of 300 euros. We can describe this conditional rule by means of its utility grid:

$$\begin{Bmatrix} a & 0 & \mathcal{A} \\ b & - & a \end{Bmatrix},$$

where a is Alice, and b is Bob. The first row of the grid describes the *if*-clause of the conditional. Asking Bob to work faster is an action of Alice that, by itself and on its own, does not have significant utility to anyone. Bob charging an additional fee of 300 euros, however, is an action of Bob that has significant negative utility for Alice, and this is represented in the second row of the grid. According to the theory of utility conditionals, the configuration of this grid is one that triggers the folk axiom of Self-interested behavior. As a consequence, the theory predicts that reasoners infer the following conclusion: Alice will not ask Bob to work faster.

But what if she does, that is, what if reasoners learn that she decided to ask Bob to work faster? This decision would be inconsistent with their expectations, and this inconsistency should trigger cognitive processes aimed at reasoning their way back to consistency (Johnson-Laird et al. 2004). More precisely, they might come to retract or

doubt one of the assumptions they made, which led them to expect that Alice would not ask Bob to work faster (Dieussaert et al. 2000; Elio and Pelletier 1997; Markovits and Schmeltzer 2007; Politzer and Carles 2001). Three assumptions are candidates for such a revision, for all three are necessary in order to expect that Alice would not ask Bob to work faster:

- The *Knowledge* assumption: Reasoners plausibly assumed that Alice knew about the consequences of her request. It was in fact explicitly spelled out in the story that she did. Nevertheless, they might consider that this information was incorrect, and that she did not in fact know what would happen if she asked Bob to work faster. In that case, they have no grounds to expect that Alice will not make the request.
- The *Importance* assumption: Reasoners plausibly assumed that it mattered to Alice that she would pay an additional fee of 300 euros. More precisely, the utility grid assumed that this additional fee had significant negative utility for Alice. If this assumption is retracted, then reasoners have no grounds to expect that Alice will not make the request.
- The *Folk axiom* assumption: Reasoners plausibly assumed that Alice's decisions would reflect the folk axiom of Self-interested behavior (i.e., that she would not take an action that would decrease her utility). If they retract this assumption, they have no ground to expect that Alice will not make the request.

In sum, the reasoners' expectation about what Alice would do arguably rests on three assumptions: That she knew about the consequences of her actions, that she cared about these consequences, and that she made decisions based on her self-interest. Now that the expectation is proven wrong by the facts, reasoners might decide to retract one, two, or all three of these assumptions. In the rest of this article, we will focus on the Knowledge and Importance assumption. The belief that other agents' decisions are guided by self-interest is extremely strong and pervasive (Miller 1999; Ratner and Miller 1998) and we assume that it would only be retracted in special cases. Accordingly, and in the interest of simplicity, we will focus on basic cases wherein the Folk Axiom assumption is taken for granted.

When reasoners expectations in such basic situations are proven wrong by the facts, are they going to retract the Knowledge assumption, the Importance assumption, or both? Retracting or doubting both assumptions would not seem to be an efficient resolution of the inconsistency. Indeed, previous psychological experiments on belief revision suggested that reasoners broadly followed a minimalist strategy, that is, that they did not revise more beliefs than necessary when resolving an inconsistency between facts and expectations (Dieussaert et al. 2000; Elio and Pelletier 1997; Markovits and Schmeltzer 2007; Politzer and Carles 2001). The question, thus, is whether reasoners will apply a priority rule when deciding to revise either the Importance or the Knowledge assumption (that is, whether they will preferentially revise one or the other), and whether this priority rule is the same for all reasoners.

To investigate this issue, we conducted an experiment in which we divided reasoners into two groups. Reasoners in the prediction group read scenarios whose main feature was a conditional rule describing the positive or negative consequences of an action that an agent might decide to take. We recorded whether these reasoners expected

the agent to take the action, and whether they made the Knowledge and Importance assumptions. Reasoners in the revision group read the same scenarios, but were then told that the agent did not take the action that could be expected. We then recorded whether these reasoners retracted the Knowledge assumption, the Importance assumption, or both. The detailed methods and results of this experiment are reported in the next section.

5 Experiment

5.1 Method

Subjects (62 women and 16 men, mean age = 21, standard deviation = 4.7) were recruited on campus at the University of Toulouse (France). They were randomly assigned to the prediction ($n = 35$) and revision ($n = 37$) groups. The experiment was conducted in French.

Subjects in the prediction group read four brief scenarios, and answered to three questions about each. Table 1 displays an example of scenario with its attached questions. All scenarios used the same format: Someone wished to do something (e.g., *Carol wishes to rent a car*), and some action that character could take would save her some amount of money (e.g., *If she uses her fidelity point then she saves 80 euros*). Depending on the scenario, that amount of money could be 40, 80, 160, or 320 euros. Four versions of the questionnaire were constructed so that each amount appeared once in each scenario.

The three questions assessed in turn the subjective confidence that the action would be taken (e.g., *Is she going to use her fidelity points?*), that the character knew about the consequences of this action (e.g., *Does she know that if she uses her fidelity points she saves 80 euros?*), and that the amount was an important one for the character (e.g., *Is 80 euros an important sum for her?*). Responses were given on three identical 11-point scales anchored at *probably not* and *probably.* See Table 1 for an example of exactly what the participants saw.

Table 1 Example of scenario and questions used in the prediction group of the experiment

Carol wishes to rent a car.												
If she uses her fidelity points she saves 80 euros.												
Is she going to use her fidelity points?												
Probably not	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Probably
	−5	−4	−3	−2	−1	0	+1	+2	+3	+4	+5	
Does she know that if she uses her fidelity points she saves 80 euros?												
Probably not	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Probably
	−5	−4	−3	−2	−1	0	+1	+2	+3	+4	+5	
Is 80 euros an important sum for her?												
Probably not	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Probably
	−5	−4	−3	−2	−1	0	+1	+2	+3	+4	+5	

Table 2 Average ratings (standard deviation in between parentheses) of whether the character knew about the consequences of her actions, and whether the amount she could save was important to her, in the prediction and revision groups, as a function of the amount that could be saved

	40 euros	80 euros	160 euros	320 euros
Prediction group				
Expectation	+2.9 (1.8)	+2.9 (2.2)	+3.0 (2.3)	+3.4 (2.1)
Knowledge	+2.6 (2.3)	+1.9 (3.1)	+1.5 (3.1)	+2.0 (2.9)
Importance	+1.4 (2.3)	+2.0 (2.3)	+2.7 (2.4)	+3.2 (2.5)
Revision group				
Knowledge	+1.3 (3.3)	−0.6 (3.8)	−0.4 (3.5)	−0.1 (3.2)
Importance	−0.9 (2.8)	+0.5 (3.0)	+0.2 (2.9)	+0.4 (3.3)

For the prediction group, the first row also shows subjects' average expectation that the money-saving action would be taken

Subjects in the revision group read the same scenarios, to an important exception: They were informed that the character did not in fact take the action that would have saved her money (e.g., *She does not use her fidelity points*). They were then asked about their subjective confidence that the character knew about the consequences of this action (e.g., *Does she know that if she uses her fidelity points she saves 80 euros?*), and about their subjective confidence that the amount featured in the scenario was an important one for the character (e.g., *Is 80 euros an important sum for her?*). They responded to these questions on the same 11-point scale as used in the prediction group.

5.2 Results

Table 2 displays descriptive statistics for all questions in all experimental conditions. Results in the prediction group were straightforward. Whatever the amount involved, subjects expected the character to take the action that could save that amount: Expectation ratings were significantly greater than zero for all four amounts, all $t > 8$, all $p < .001$ (df $= 39$). Similarly, whatever the amount involved, subjects expected the character to know about the consequence of the action: Knowledge ratings were significantly greater than zero for all four amounts, all $t > 3$, all $p < .005$ (df $= 39$). Finally, all amounts were judged important for the character: Importance ratings were significantly greater than zero, all $t > 3.8$, all $p < .001$ (df $= 39$). In addition, and quite expectedly, the perceived importance of the amount increased with its objective value, as shown by an analysis of variance using the importance rating as the dependent variable and the objective value as the predictor, $F(3, 37) = 13.6$, $p < .001$. A follow-up contrast analysis showed that this increase was linear across the four objective values, $F(1, 39) = 39.7$, $p < .001$.[3]

[3] The experiment used the four objective values 40, 80, 160, and 320, without informing the participants of what these value meant in terms of the percentage of the discount that the character could obtain. While this additional information could be relevant to their judgments, the importance ratings do show that our simple manipulation achieved its objectives.

The results obtained from the prediction group thus confirm that subjects provided with a conditional information such as 'if the character takes action p, then she will save x euros' assume that the character knows about this, that saving x euros has positive utility to the character (in proportion of x), and that the character will take action p. What now if the subjects are informed that the character did not in fact take action p?

The data obtained from the revision group address this question. Two steps of analysis are necessary to interpret these data, for the group-level analysis will be usefully complemented by an individual-level analysis. At the group level, it appears that the data obtained when the amount that could be saved was small (40 euros) are qualitatively different from the data obtained with the other amounts. An analysis of variance conducted on the knowledge ratings detected an effect of the amount involved, $F(3, 35) = 4.0$, $p = .01$. This overall effect reflects the fact that the knowledge rating was significantly greater than zero in the 40 euros condition only, $t(37) = 2.4$, $p = .02$. For amounts of 80 euros and up, the knowledge ratings were not significantly greater than zero, all $t < 1$, all $p > .36$. In parallel, an analysis of variance conducted on the importance ratings suggested an effect of the amount involved, $F(3, 35) = 2.2$, $p = .10$. This marginal effect reflects the fact that the importance rating was significantly *below* zero in the 40 euros condition only, $t(37) = -2.0$, $p = .05$. For amounts of 80 euros and up, the knowledge ratings were not significantly different from zero, all $t < 1$, all $p > .31$.

This first step of analysis indicates that at the group level, a clear revision strategy emerges only for the 40 euros condition. That is, when the amount that could be saved was 40 euros, and when told that the character did not take the action that could save the 40 euros, subjects assumed that the character knew about the opportunity but did not worry about saving 40 euros.

For other amounts, though (80 euros and up), the group-level analysis would suggest that subjects expressed total ignorance about whether the character knew about the opportunity, *as well as* about whether the amount was important to the character. This interpretation, however, is misleading, as shown by individual-level analyses. A look at the standard deviations in Table 2 already suggests that the variance in responses was larger in the revision group than in the prediction group, for all ratings. This tendency could be the result of different individuals applying different revision strategies. That is, it could be the case that some subjects explained the character's decision by considering that she did not know about the opportunity (without revising how important the amount was to her), whilst other subjects explained the character's decision by considering that the amount involved was not important enough to her (without revising their assumption that she knew about the opportunity).

Figure 1 suggests that this is indeed the case. Each bar in Fig. 1 corresponds to a subject, or a group of subjects (in case of ties), arranged in increasing order of the average Knowledge rating they gave across the four scenarios. The gray bars display these average knowledge ratings. The critical information contained in Fig. 1 is conveyed by the black bars, which show the average Importance rating given by each subject across the four scenarios. What Fig. 1 strongly suggests is that most subjects in the revision group chose one revision strategy and stuck to it. That is, either they revised their assumption that the character knew about the opportunity, or they revised the

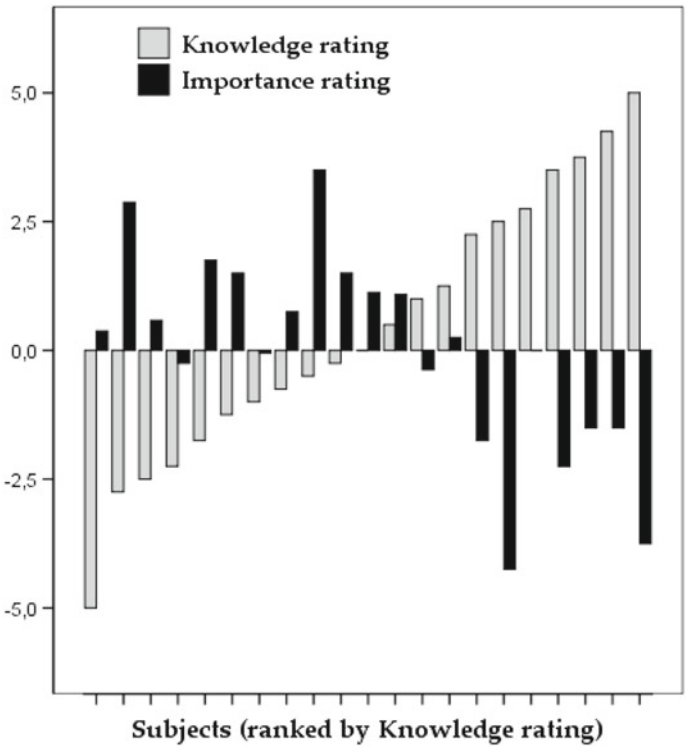

Fig. 1 Individual-level association between Knowledge ratings and Importance ratings in the revision group, across the four scenarios

importance that the character would attach to the amount she could save, but (usually) not both. This is confirmed by the the correlation coefficient between average Knowledge ratings and average Importance ratings, whose value was significantly negative, $r(38) = -.54$, $p = .001$. Across individuals in the revision group, Importance ratings decreased when Knowledge ratings increased, and about 30%of the variation in Importance ratings was explained by the variation in Knowledge ratings.

Overall, the picture that emerges from these results is the following. In line with previous theoretical accounts, subjects expected the characters in the stories to take the actions they could benefit from. This expectation went hand in hand with the assumption that the character knew about the potential benefit, and the assumption that the benefit was indeed valued by the character. When informed that the character did not in fact take the action, subjects revised one of these assumptions. When the benefit was small, most subjects decided to revise their assumption that it was of the value to the character. As soon as the benefit grew bigger, though, broadly half the subjects chose to revise the first assumption, and broadly half the subjects chose to revise the second assumption. Although subjects demonstrated some economy of thinking (by not revising both assumptions), it appears that outside the realm of small benefits, neither assumption was consensually more entrenched than the other.

6 Conclusion

Individuals reason their way to outcomes they desire, they assess the motivations of other agents, and they predict behavior based on these motivations. Preferences (or beliefs about preferences) can be both the input and the output of reasoning. It might come as a surprise then that the psychology of reasoning has been mostly concerned with *disinterested* reasoning, of which preferences are neither an input nor an output.

We briefly illustrated how such an exclusive focus was untenable: People take into account the preferences of other agents in their reasoning, and that they do so at the expense of deductive logic, when logic and preferences point to different conclusions. Part of this phenomenon was accounted for by Bonnefon (2009), in the form of a theory of how reasoners predict the behavior and attitudes of other agents based on conditional rules describing actions and their potential consequences. We summarized the main aspects of this theory, and turned to an issue that it had not addressed so far, that is, the kind of belief revision that reasoners considered when their expectations were contradicted by the facts.

The basic situation we considered was one wherein an agent could take some action p that would result in a benefit q. Based on previous work, we hypothesized that reasoners would expect the agent to take the action p, based on three assumptions: That the benefit q was important to the agent (the Importance assumption), that the agent knew about the consequences of action p (the Knowledge assumption), and that the agent would demonstrate self-interested behavior (the Folk axiom assumption). The new question we considered was which of these assumptions would be retracted when learning that the agent did not in fact take action p. We considered basic situations wherein the Folk Axiom assumption was taken for granted, and we only offered reasoners the possibility to retract the Knowledge assumption and/or the Importance assumption.

An experiment confirmed that reasoners did expect the agent to take action p, and did make the knowledge and importance assumptions. Furthermore, this experiment shed light on the kind of belief revision reasoners considered when told that the agent did not in fact take the action. When the benefit was small, reasoners tended to revise the importance assumption rather than the knowledge assumption: They considered that the agent knew about the benefit, but did not care about it. As soon as the benefit grew larger, though, no modal strategy was identified. It appeared that some reasoners focused their revision on the importance assumption, and others on the knowledge assumption, while abiding to some form of minimalist revision.[4] Reasoners tended to consistently revise one assumption but not the other, but they did not agree on which assumption to revise.

[4] According to a common philosophical thesis, changes to beliefs in the face of an inconsistency should be as minimal as possible (e.g. Harman 1986; James 1907). Our results could be taken as descriptive evidence for some version of minimality. Experimental evidence against other forms of minimality are available in the literature, though. For example, reasoners facing an inconsistency can prefer an explanation that indicates both a cause and an effect, to an explanation consisting of the effect alone (Johnson-Laird et al. 2004).

These initial findings naturally call for future refinements. The psychology of reasoning about preferences is young, but it does a reasonably good job at predicting what reasoners think other agents will do. The new challenge that results from our findings is to identify the principles that reasoners apply when these predictions turn out to be incorrect. Our initial findings helped identify some potential structure in this revision process. Future research will have to further clarify this structure, and to work it in existing models of reasoning about preferences. Indeed, formal models have been proposed that specifically look at how to pick the best explanation of abnormal behavior, based on the beliefs and motivations of agents (e.g., Bex et al. 2009, in the legal domain), but these computational models are not, to our knowledge, fueled by experimental data. It is of theoretical and practical importance, though, to establish how people reconcile their discovery of an unexpected behavior to their prior assumptions about the agent's beliefs and preferences. Even if an optimal solution to the problem can be defined in the form of a computational model, it will still be necessary to know about the situations where lay reasoning might deviate from the conclusions of the model—if only to anticipate the biases that might be demonstrated by, say, jurors in a criminal case.

In that regard, the fact that reasoners in our experiment were consistent but split in their revision strategies points to a potential difficulty for formalizing reasoning about unconsequential decisions. It could be the case that people have a stable individual tendency to revise their beliefs about the preferences of other agents, or to revise their beliefs about the beliefs of other agents. If that was true, descriptive formal models would have to incorporate this individual difference parameter, which is not an easy endeavor. Our results, though, are certainly not conclusive in that respect, and further psychometric research is needed before we can make an informed claim about the existence of such a stable individual tendency, as well as about its personality, demographic, or cognitive correlates.

References

Amgoud, L., Bonnefon, J. F., & Prade, H. (2007). The logical handling of threats, rewards, tips, and warnings. *Lecture Notes in Artificial Intelligence, 4724*, 235–246.

Beller, S., Bender, A., & Kuhnmunch, G. (2005). Understanding conditional promises and threats. *Thinking and Reasoning, 11*, 209–238.

Benferhat, S., Bonnefon, J. F., & Da Silva Neves, R. M. (2005). An overview of possibilistic handling of default reasoning, with experimental studies. *Synthese, 146*, 53–70.

Bex, F., Bench-Capon, T., & Atkinson, K. (2009). Did he jump or was he pushed? Abductive practical reasoning. *Artificial Intelligence and Law, 17*, 79–99.

Bonnefon, J. F. (2009). A theory of utility conditionals: Paralogical reasoning from decision-theoretic leakage. *Psychological Review, 116*, 888–907.

Bonnefon, J. F., & Hilton, D. J. (2004). Consequential conditionals: Invited and suppressed inferences from valued outcomes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 28–37.

Bonnefon, J. F., & Sloman, S. A. (2011). *The causal structure of utility conditionals*.: Manuscript submitted for publication.

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85*, 1–21.

Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition, 31*, 61–83.

Chater, N., Crocker, M., & Pickering, M. (1998). The rational analysis of inquiry: The case for parsing. In N. Chater & M. Oaksford (Eds.), *Rational models of cognition.* (pp. 441–468). Oxford: Oxford University Press.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory and Cognition, 31*, 581–595.

Dieussaert, K., Schaeken, W., De Neys, W., & d'Ydewalle, G. (2000). Initial belief state as a predictor of belief revision. *Cahiers de Psychologie Cognitive, 19*, 277–288.

Elio, R., & Pelletier, F. J. (1997). Belief change as propositional update. *Cognitive Science, 21*, 419–460.

Evans, J. S. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin, 128*, 978–996.

Evans, J. S. B. T. (2005). The social and communicative function of conditional statements. *Mind & Society, 4*, 97–113.

Evans, J. S. B. T., Neilens, H., Handley, S. J., & Over, D. E. (2008). When can we say 'if'? *Cognition, 108*, 100–116.

Evans, J. S. B. T., Newstead, S., & Byrne, R. (1993). *Human reasoning: The psychology of deduction*. Psychology Press.

Ficici, S. G., & Pfeer, A. (2008). Simultaneously modeling humans' preferences and their beliefs about others' preferences. In *Proceedings of AAMAS'08* (pp. 323–330). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge, MA: Bradford Book.

James, W. (1907). *Pragmatism—A new name for some old ways of thinking*. New York: Longmans Green, & Co.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale NJ: Lawrence Erlbaum Associates.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review, 111*, 640–661.

Legrenzi, P., Politzer, G., & Girotto, V. (1996). Contract proposals: A sketch of a grammar. *Theory & Psychology, 6*, 247–265.

López-Rousseau, A., & Ketelaar, T. (2004). "If...": Satisficing algorithms for mapping conditional statements onto social domains. *European Journal of Cognitive Psychology, 16*, 807–823.

López-Rousseau, A., & Ketelaar, T. (2006). Juliet: If they do see thee, they will murder thee: A satisficing algorithm for pragmatic conditionals. *Mind & Society, 5*, 71–77.

Markovits, H., & Schmeltzer, C. (2007). What makes people revise their beliefs following contradictory anecdotal evidence? The in uence of systemic variability and direct experience. *Cognitive Science, 31*, 535–547.

Miller, D. T. (1999). The norm of self-interest. *American Psychologist, 54*, 1053–1060.

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review and re-evaluation. *Psychonomic Bulletin and Review, 10*, 289–318.

Oaksford, M., Chater, N., & Grainger, B. (1999). Probabilistic effects in data selection. *Thinking and Reasoning, 5*, 193–243.

Ohm, E., & Thompson, V. (2004). Everyday reasoning with inducements and advice. *Thinking and Reasoning, 10*, 241–272.

Ohm, E., & Thompson, V. (2006). Conditional probability and pragmatic conditionals: Dissociating truth and effectiveness. *Thinking and Reasoning, 12*, 257–280.

Politzer, G., & Bonnefon, J. F. (2006). Two varieties of conditionals and two kinds of defeaters help reveal two fundamental types of reasoning. *Mind and Language, 21*, 484–503.

Politzer, G., & Carles, L. (2001). Belief revision and uncertain reasoning. *Thinking and Reasoning, 7*, 217–234.

Ratner, R. K., & Miller, D. T. (1998). The disparity between the actual and assumed power of self-interest. *Journal of Personality and Social Psychology, 74*, 53–62.

Rips, L. J. (1983). Cognitive processes in propositional reasoning. *Psychological Review, 90*, 38–71.

Smedslund, J. (1997). *The structure of psychological common sense*. Mahwah, NJ: Erlbaum.

Stenning, K., & Lambalgen, M.van (2005). Semantic interpretation as computation in nonmonotonic logic: The real meaning of the suppression task. *Cognitive Science, 29*, 919–960.

Thompson, V. A., Evans, J. S. B. T., & Handley, S. J. (2005). Persuading and dissuading by conditional argument. *Journal of Memory and Language, 53*, 238–257.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. London: Batsford.