Chapter 1

Predicting Behavior on the Basis of Arguments From Consequences

Jean-François Bonnefon

Centre for Research in Management, Toulouse School of Economics, Toulouse, France, jfbonnefon@gmail.com

Abstract. People reason about the preferences of others when they seek to win them to a course of action, or when they seek to predict their behavior. In order to win others to a course of action, one has to design arguments from consequences, showing others how their preferences will be satisfied. Since this persuasive use of preferences is already well described in the argumentation literature, this chapter offers a survey of the other kind of reasoning about preferences, which occurs when reasoners use arguments from consequences to predict the behavior of others. I will describe what inferences people derive from arguments from consequences, how they fit ambiguous arguments into templates, how they resolve ambiguity in the causal structure of these arguments, how they react to behavior that violates their predictions, and how they use expected affect to predict behavior based on complex arguments from consequences which impact several agents.

1. Introduction

Arguments from consequences attempt to persuade an agent to engage in or refrain from a behavior, by stating desirable and/or desirable consequences of this behavior for the agent — and assuming that the agent will make a consequentialist decision to engage in (resp., refrain from) the behavior if its consequences are, on the whole, positive (resp., negative).

For example, (1a) states the consequence of eating fresh fruits everyday, (1b) makes it explicit that that this consequence is desirable to the recipient, and (1c) concludes the argument by an injunction to eat fresh fruits everyday. Accordingly, Example (1ac) is a fully explicit argument from consequences, which states the consequences of a behavior, the desirability of these consequences, and the conclusion that the recipients ought to engage in the behavior.

(1)

a. Eating fresh fruits everyday will improve your health;
b. Improving your health is desirable;
c. Therefore, you should eat fresh fruits everyday.

Arguments from consequences are ubiquitous in persuasion attempts. For example, Schellens and De Jong (2004) observed that arguments from consequences were present in each and every government brochure they analyzed, that aimed at behavioral change. Given
the prevalence and importance of arguments from consequences, it is perhaps unsurprising that they have been a prime target for investigation in both the argumentation and the reasoning literatures (e.g., Corner, Hahn, & Oaksford, 2011; Evans, Neilens, Handley, & Over, 2008; Feteris, 2002; Hoeken, Timmers, & Schellens, 2012; Thompson, Evans, & Handley, 2005).

Importantly for the present chapter, there are two sides to arguments from consequences. They can be used to persuade people to undertake a course of action, but they can also be used to predict individual behavior. Consider first that, in many instances, only the first part of the argument is made explicitly (Schellens & De Jong, 2004), leaving it to the agent to realize that the consequences are (un)desirable, and that they should (not) engage in the behavior as a result. For example, one may simply state (2a), and leave it to the recipient to fill in (2b) in order to come to the conclusion (2c).

(2)
   a. Maxing out your credit cards hurts your credit score;
   b. It is not desirable to hurt your credit score;
   c. Therefore, you should not max out your credit cards.

Sentence (2a) is the compact form of the argument from consequences (2ac). An interesting characteristic of compact arguments from consequences is that they can be expressed as conditional statements, as in (3a) or (3b).

(3)
   a. If you max out your credit cards, you will hurt your credit score.
   b. If Andrew sends his invoice tomorrow, he will receive $3,000.

This is an interesting characteristic because it makes it easier to connect argumentation to the psychology of reasoning, which always had a special interest for conditional statements. Note in that respect that Sentence (3b) can be used as an argument to convince Andrew to send his invoice the day after, but it also provides a reason to infer that Andrew, as a rational agent, will indeed send his invoice the day after. More generally, the compact form of an argument from consequences (i.e., action a will lead to a consequence valued by an agent A) can be used as a basis for persuasion or prediction. When used as a basis for persuasion, they attempt to convince agent A to take or not to take action a; when used as a basis for prediction, they lead to the inference that agent A will or will not take action a.

My intention in this chapter is to provide an overview of the work I have conducted over the years on the predictive (rather than persuasive) use of compact arguments from consequences, when they are expressed as conditionals. In the rest of the chapter, we will consider the compact form of arguments from consequences illustrated in Sentence (3b). This compact form consists of a conditional statement ‘if p then q’ where p is an action of an individual, whose consequence q has utility for this or other individuals. First, I will introduce the theoretical framework dealing with these utility conditionals (Bonnefon, 2009). Next, we will see that people have expectations about the structure of such statements, which led them to apply utility templates when an argument from consequences is ambiguous about how agents value some states of the world. We will come to know that the underlying causal structure of these statements impose strong constraints on the inferences they afford, even when the surface structure of the argument does not change. We will consider data showing that expected affect drives the inferences based on arguments from consequences, in complex situations where the actions of an agent impact
other individuals besides herself. Finally, we will examine how people process stories in which characters violate expectations based on arguments from consequences: How do people explain why a character failed to take the action suggested by an argument from consequences, and what can their eye movements tell us about their online processing of such situations?

2. Utility Conditionals

Utility conditionals are statements of the form ‘if p then q’ where p is an action of an individual, whose consequence q has utility for this or other individuals (Bonnefon, 2009). They can be unpacked as: if agent x takes action p which has utility u for agent y, then agent x′ will take action q which has utility u′ for agent y′. This information can be represented in the utility grid of the conditional. The role of the utility grid is to provide a standard representation of who can do what to which consequences for whom, so that formal principles (the folk axioms of decision introduced later in this section) can be applied to the grid in order to systematically predict the inferences afforded by the conditional statement.

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

The set of all agents is denoted by A. By convention, the agent who states the conditional is denoted by s (for ‘speaker’), the agent at whom the conditional is directed is denoted by h (for ‘hearer’), and e (for ‘someone else’) denotes 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. Finally, utility is represented in the grid by 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 the whole set of agents A as a target.

A statement like ‘If Andrew sends his invoice tomorrow, he will receive $3,000’ would typically receive the following utility grid:

\[
\begin{array}{ccc}
  e & 0 & A \\
  \omega & + & e \\
\end{array}
\]

This grid means that if agent e (here, Andrew) takes action p which does not have any obvious intrinsic utility, an event q will occur which has positive utility for the same agent e. To illustrate the power of the utility grid notation, let us consider a diverse set of arguments from consequences, expressed as utility conditionals:
a. If you testify against me, you will have an accident.
b. If I study instead of partying, I will get good grades.
c. If she moves to Paris, I’ll be unhappy.
d. If I let you go away with this, my boss will fire me.

These four examples would receive these four utility grids:

\[
\begin{align*}
\{ h &- s \\
\omega &- h \\
\{ s &- s \\
\omega &+ s \\
\{ e &0 A \\
\omega &- s \\
\{ s &+ h \\
\{ e &- s \\
\end{align*}
\]

Once a utility conditional has been translated to an utility grid, the theory makes use of folk axioms of decision to predict the inferences it affords. These folk axioms correspond to heuristics that people use to predict the behavior of other individuals on the basis of arguments from consequences. For example, the folk axiom of self-interested behavior states that agents tend to take actions that increase their utility, and do not take actions that decrease their utility.

Accordingly, the following grids (where the black dot stands for any legitimate value of the parameter) afford the inference that agent \( x \) will take action \( p \):

\[
\begin{align*}
\{ x &+ x \\
. &. . \\
\{ x &. . \\
. &+ x \\
\end{align*}
\]

whereas the following grids afford the inference that agent \( x \) will not do \( p \):

\[
\begin{align*}
\{ x &- x \\
. &. . \\
\{ x &. . \\
. &- x \\
\end{align*}
\]

To give another example, the folk axiom of limited altruism states that people take actions that increase the utility of other agents, as long as these actions do not decrease their own utility (and \textit{mutatis mutandis}, they do not take actions that would decrease the utility of other agents, unless these actions would increase their own utility). Accordingly, the following grid is one among others which affords the inference that agent \( x \) will take action \( p \):

\[
\begin{align*}
\{ x &0 A \\
\omega &+ y \\
\end{align*}
\]
Looking back at Examples (4-a-d), the folk axioms of self-interested behavior and of limited altruism, applied to the four utility grids of the four statements would predict the following inferences:

(5)
- a. The hearer will not testify.
- b. The speaker will study unless he or she finds positive utility in partying.
- c. She will not move to Paris.
- d. The speaker will not let the hearer go away.

The theory of utility conditionals thus relies on a representational tool (the utility grid) on which folk axioms of decisions are applied to predict the inferences that people will derive from an argument from consequences. The theory accounts for the results of reasoning experiments featuring arguments from consequences (Bonnefon & Hilton, 2004; Corner et al., 2011; Evans et al., 2008; Ohm & Thompson, 2004, 2006), and it has been further tested on randomly generated statements whose utility grids were not found in previous experiments (Bonnefon, 2012). Consider for example the following statement, which is arguably weird but nevertheless corresponds to a well-formed utility grid:

(6)
If I do this, you will hurt yourself.
\[
\begin{pmatrix}
  s & 0 & A \\
  h & - & h
\end{pmatrix}
\]

The folk axiom of limited altruism applies to this grid and predicts the inference that the speaker will not do ‘this’. And indeed, reasoners gave a −2.7 rating (on a scale from −5 to +5) for the likelihood that the speaker would do ‘this’ (Bonnefon, 2012). Other statements left reasoners puzzled, for example:

(7)
If you hurt me I will help Luis.

Example (7) seems to have some sort of Necker cube quality, flipping between two possible interpretations depending on how utility are assigned to the various agents. Some people seem to interpret it as a threat (You don’t want me to help your enemy Luis, do you? Then don’t hurt me), whereas other people seem to interpret it as a promise (I want you to hurt me so I’m bribing you with a promise to help your pal Luis). This anecdotal observation is interesting because it suggests that people have a tendency to rearrange utilities so that unusual arguments from consequences fall within familiar categories. In the next section, I will consider evidence for this claim, and two examples of the utility templates that seem to guide the interpretation of ambiguous arguments from consequences.

3. Utility Templates

Bonnefon, Haigh, and Stewart (2013) hypothesized that some utility grids would act as templates for the interpretation of ambiguous or incomplete utility conditionals. These templates would correspond to special configurations of the grid that guide and constrain
the interpretation of conditional statements. The first candidate for this template status is the social contract grid:

$$\begin{pmatrix} x & u & y \\ y & u & x \end{pmatrix}$$

This grid denotes that if agent $x$ takes action $p$ with utility $u$ to agent $y$, then $y$ will take action $q$ which has the same utility grid (or more generally utility of the same sign) for agent $x$. For example:

(8)

a. If you vote for me, I will reward you.

b. If he hurts her, she will take revenge on him.

Specialists of reasoning have long observed that individuals are exquisitely sensitive to these social contracts and the inferences the afford (Haigh, Stewart, Wood, & Connell, 2011; Hilton, Kemmelmeier, & Bonnefon, 2005; Legrenzi, Politzer, & Girotto, 1996; Perham & Oaksford, 2005; Politzer & Nguyen-Xuan, 1992). Various authors have attributed this sensitivity either to repeated exposure (Cheng & Holyoak, 1985) or to an innate cheater detection algorithm (Cosmides, Barrett, & Tooby, 2010). As a result of this sensitivity, Bonnefon et al. (2013) suggested that reasoners would give priority to the social contract grid when interpreting or re-interpreting conditionals whose grid is close enough to that of a social contract.

Consider for example the following statement, which features an unknown verb whose valence is ambiguous:

(9)

If Peyton votes for Jesse, then Jesse will yorb Peyton.

Example (9) would correspond to the following utility grid, where the utility of the then-clause is unknown:

$$\begin{pmatrix} x & + & y \\ y & ? & x \end{pmatrix}$$

If reasoners apply the social contract template to the ambiguous statement (9), they should infer that there is positive utility for Peyton to be yorbed by Jesse, or, ore informally, that Peyton likes to be yorbed by Jesse. The same should apply to premises such as (10), featuring two nonverbs of which one is disambiguated in the premises:

(10)

If Peyton tymps Jesse, then Jesse will yorb Peyton. Jesse likes to be tymped.

*Mutatis mutandis*, similar predictions can be derived from statements in which one verb has negative valence, semantically or by disambiguation:

(11)

a. If Peyton harms Jesse, then Jesse will yorb Peyton.

b. If Peyton tymps Jesse, then Jesse will yorb Peyton. Jesse dislikes to be tymped.
In both cases, reasoners who apply the social contract template should infer that Peyton dislikes to be yorbled. And indeed, two experiments offered support for all these predictions.

An additional and stronger prediction is that reasoners might reinterpret the valence of known verbs, if a non-verb is disambiguated so that the resulting grid violates the social contract template. For example:

(12)

a. If Peyton harms Jesse, Jesse will zim Peyton.
b. Peyton likes to be zimmed.

If ‘harming’ is understood as carrying negative utility, as it typically is, then the utility grid of (12-a) should be the following, in line with the information in (12-b):

{\begin{array}{c}
x - y \\
y + x \end{array}}

If reasoners are drawn to the social contract template, they should attempt to reinterpret (12) in a way that would fit a social contract. Given that the interpretation of ‘zimming’ is fixed by (12-b), the only degree of freedom left for shoehorning (12-a) in a social contract template is to reinterpret ‘being harmed’ as having positive utility for Jesse. And again, this is what was observed in an additional study by Bonnefon et al. (2013).

In addition to the social contract template, Bonnefon et al. (2013) identified (and verified) three other templates, suggested by a completion study in which participants had to fill conditional fragments such as:

(13)

If Peyton harms Jesse, then . . .

Participants often completed such fragments in such a way to make them social contracts (11% of completions), but other completions were frequent enough to be considered as potential templates. One of these completions, which was dubbed the Justice template (13% of completions), correspond to the following grid:

{\begin{array}{c}
x u y \\
\omega u x \end{array}}

which would typically be expressed as:

(14)

a. If Peyton harms Jesse, then Peyton will be punished.
b. If Peyton helps Jesse, then Peyton will be rewarded.

The justice template seems to fit with a Just World intuition, according to which good things happen to people who do good, and bad things happen to people who do harm (Hafer & Begue, 2005). The justice template was tested just as the social contract was tested, and these tests revealed comparable interpretation and reinterpretation effects.

To sum up what we have covered so far: arguments from consequences can be framed as utility conditionals, and thus represented as utility grids. Reasoners hold utility templates that guide their interpretation toward some salient utility grid, when the utility of the if-
clause or the then-clause is ambiguous. Once reasoners have settled on a utility grid, they
apply folk axioms of decision (i.e., the folk axiom of self-interested behavior, or the folk
axiom of limited altruism) to predict the behavior of the agents featured in the grid. An
important twist on these results, though, is that utility conditionals can be ambiguous
beyond the utility of their clauses: Another ambiguity has to do with their underlying causal
structure, a problem that I address in the next section.

4. Utility and Causal Structure

Even when utilities can be unambiguously assigned to the clauses of an argument from
consequences, another factor can complicate its interpretation, and affect the inferences it
affords. Specifically, arguments from consequences expressed as utility conditionals can be
ambiguous with respect to their underlying causal structure. Although conditional
statements can support various causal structures (Dancygier, 1998; Declerck & Reed,
2001), we are chiefly interested in the contrast between causal and diagnostic conditionals.
Consider for example:

(15)

a. If there is fire then there is smoke.
b. If there is smoke then there is fire.

Example (15-a) is a causal conditional, in which the if-clause is a cause of the then-clause.
In contrast, Example (15-b) is a diagnostic conditional, in which the if-clause is diagnostic
of the then-clause. Inferences made from causal and diagnostic conditionals generally track
their underlying causal structure (e.g., Ali, Chater, & Oaksford, 2011; Politzer & Bonnefon,
2006), and causal structure can also affect the inferences afforded by an argument from
consequences, when this argument is expressed as a conditional. Consider for example:

(16)

If Corey accepts this deal, then he’s rich.

This conditional would receive the following utility grid:

\[
\begin{array}{cc}
   x & 0 \\
   \omega & A \\
x & \\
\end{array}
\]

This grid expresses the fact that if Corey takes an action which has no apparent intrinsic
utility, a state of affairs will occur that has positive utility for Corey. Applied to that grid,
the folk axiom of self-interested behavior allows the prediction that Corey will accept the
deal. But note that this application of the folk axiom of self-interested behavior assumes
that Example (16) expressed a causal relation between accepting the deal and becoming
rich: presumably, accepting the deal would cause Corey to become rich because it is a very
good deal.

Example (16), though, may very well express a diagnostic relation. Accord- ing to this
diagnostic interpretation, Corey would have to be rich to accept the deal (say, because the
entry cost in the deal is huge), and thus the fact that Corey would accept the deal would
indicate that he’s rich, rather than causing him to become rich. In that case, the folk axiom
of self-interested behavior is silent about whether Corey will accept the deal, because there
is no clear advance in self-interest linked to accepting the deal. Note that a similar analysis
applies to Example (17), which features a negative utility then-clause:

(17)
If Corey accepts this deal, then he’s ruined.

Once again, accepting the deal may cause the ruin of Corey (because the deal is very bad), or it may simply be diagnostic of the ruin of Corey (say, because the deal only appeals to desperate persons). And once more, the folk axiom of self-interested behavior applies under the first interpretation (affording the prediction that Corey will not accept the deal) – but it is silent under the second interpretation.

In sum, some utility conditionals are ambiguous with respect to their underlying causal structure: They can be interpreted causally, making them genuine arguments from consequences, or they can be interpreted diagnostically, in which case it is not clear what reasoners will do. Will reasoners track the causal structure of a utility conditional, and refrain from drawing conclusion when it is diagnostic? Or will they rely on the surface structure of the conditional, and use it as any other argument from consequences, regardless of its underlying causal structure?

In several experiments, Bonnefon and Sloman (2013) investigated this question and provided evidence that reasoners were sensitive to the underlying causal structure of utility conditionals. These experiments featured ambiguous conditionals such as (16) or (17), and used different disambiguating manipulations. For example:

(18)
  a. If he accepts this deal, then he’s ruined. (Because it’s a very bad deal.) Is he going to accept the deal?
  b. If he accepts this deal, then he’s ruined. (Because the deal only appeal to desperate persons.) Is he going to accept the deal?

Or:

(19)
  a. The fact that he buys this house would make him rich as a result. In other words: If he buys this house, then he is rich. Is he going to buy the house?
  b. The fact that he buys this house would indicate that he is rich. In other words: If he buys this house, then he is rich. Is he going to buy the house?

Responses to these questions tracked the underlying causal structure of the causal conditional, and not its surface structure. That is, reasoners were highly likely to agree with the conclusion for (18-a) and (19-a), and much less likely to do so for (18-b) and (19-b). Accordingly, inferences derived from utility conditionals cannot be predicted on the sole basis of the utilities of the if-clause and the then-clause: Causal structure needs be incorporated in the utility grid. Bonnefon and Sloman (2013) suggested one such modification to the utility grid format, which would make the top and bottom rows refer to the cause and the effect, rather than to the if-clause and the then clause. Consider again the example:

(20)
If Corey accepts this deal, then he’s ruined.
Jean-François Bonnefon

The original version of the utility grid theory (Bonnefon, 2009) was insensitive to causal structure, and would have assigned the following grid to (20), whatever its underlying causal structure:

\[
\begin{align*}
\{ x & \quad 0 \quad A \\
\omega & \quad - \quad x
\end{align*}
\]

The updated version of the theory would assign one of two possible grids to (20), depending on whether it is causal (top) or diagnostic (bottom):

\[
\begin{align*}
\{ x & \quad 0 \quad A \\
\omega & \quad - \quad x
\end{align*}
\]

\[
\begin{align*}
\{ \omega & \quad - \quad x \\
x & \quad 0 \quad A
\end{align*}
\]

Folk axioms of decision can then be applied to the retained grid, just as they were in the original version of the theory. The folk axiom of self-interested behavior applies to the grid on the left (affording the conclusion that Corey will not accept the deal), but not to the grid on the right (thus not affording any conclusion). This modification to the utility grid format can thus adequately address situations in which the meaning of a sentence can be that of a genuine argument from consequences, or that of a merely diagnostic observation.

Now that we have addressed the various problems that can complicate the interpretation of an argument from consequences, we can turn to situations in which the argument is clearly spelled and readily affords an inference. Most of these situations will fall under the basic framework already summarized in the Section Utility Conditionals. Some situations, though, call for closer scrutiny. In the rest of this chapter, I will discuss two such special situations. First, we will consider how people process situations in which a character’s actions have consequences for other individuals besides herself, which can conflict with her own personal utility. Second, we will look at how people process and explain situations in which a character acts against her best-interest.

5. Conflicting Utilities

So far we only considered situations in which a single agent acts on the basis of her personal preferences. A more complicated situation arises when the action of one agent have consequences for other people besides herself, especially when consequences for the self and consequences for others have the same valence.

Let us consider a situation in which the actor can decide to take an action which has either positive or negative consequences for herself, and at the same time positive or negative consequences for a recipient. If reasoners expect people to act on the basis of their material self-interest, then they should predict the actor to take the action that maximizes positive consequences for herself, regardless of the consequences for the recipient. Although this view has had its proponents (Kruger & Gilovich, 1999; Miller, 1999), it is at odds with one of the basic results of behavioral economics, according to which people care about the material interests of others, and expect others to feel the same (Cooper & Kagel, in press).

Another possibility formulated by De Vito and Bonnefon (2014) is that people expect others to maximize hedonic rather than material utility. That is, they expect others to take actions that maximize their positive affect, rather than their material benefits. If they expect
another individual to feel bad about a selfish action, to the extent that these negative feelings surpass the positive feelings produced by material benefits, then they should expect this actor not to take the selfish action. This analysis extends to all other situations – for example, if a reasoner expect an actor to feel good about being altruist, to the extent that these positive feelings exceed the negative feelings produced by the material cost of altruism, then they should expect the actor to take the altruist action. Note that this is only selfishness under another guise. Indeed, according to this account, actors do not care about the feelings of the recipient: they simply maximize their own positive feelings, regardless of the feelings of the recipient.

To test this hypothesis, De Vito and Bonnefon (2014) presented reasoners with four types of vignettes in which an actor could take an action that had either positive or negative material consequences for herself, and either positive or negative material consequences for the recipient. Just like in the other experiments reviewed so far, reasoners had to predict whether the actor would take the action. Critically, they then responded to an exhaustive inventory of questions about the emotions that the actor would feel after taking or not taking the action, and the emotions that the recipient would feel if the action was taken or not taken. Based on these 32 questions about emotions, it was possible to compute for each situations, the net expected hedonic utility of the action for the actor, and the net expected hedonic utility of the action for the recipient.

Results wholly supported the predictions. The behavior of the actor could not be predicted by simple material utility maximization, but was accurately predicted by the net expected affect of the action for the actor. Furthermore, the behavior of the actor was uniquely predicted by her own affect, independently of the expected affect of the recipient. In other words, when an argument from consequences features both an actor and another individual (the recipient), people expect the actor to selfishly maximize her own positive emotions, which may or may not coincide with the material utility of the action, regardless of the material or affective consequences of the action for the recipient.

6. Utility Violations

The research I reviewed so far firmly established that reasoners draw inferences from arguments from consequences, in particular when they are expressed as utility conditionals. For example, given (21), reasoners infer that Alice, as a self-interested agent, will file her taxes online:

(21)
If Alice files her taxes online, she will save 300 euros.

But how do they react if they discover that Alice did not in fact file her taxes online? That is, how do people process information which contradicts the expectations they formed from an argument from consequences?

There are two ways to address these questions, which I’ll elaborate in this section. Firstly, we can simply ask people for their explanation of Alice’s behavior. This I will call the ‘offline’ method, because it does not track the cognitive processes at work, only their output. Secondly, we can monitor cues to cognitive processing (and specifically, eye movements) at the very moment people discover Alice’s unexpected behavior – this I will call the ‘online’ method.

Bonnefon, Girotto, and Legrenzi (2012) used the offline method to investigate how reasoners would explain an action that contradicted the inferences they had derived from a
statement like (21). They focused on what appeared to be the two most plausible explanations, one based on the beliefs of the agent, and the other based on the preferences of the agent. More specifically, they argued that to infer that ‘Alice will file her taxes online’ from (21) required two assumptions: (a) that Alice knows about the opportunity to save 300 euros, and (b) that Alice cares about saving 300 euros. If one of this assumptions is incorrect, then there is no ground to predict that Alice will file her taxes online. Accordingly, they expected reasoners to revise at least one of these two assumptions when learning that Alice did not file her taxes online.

To test this hypothesis, Bonnefon et al. (2012) assigned one control group of participants to a standard utility conditional reasoning task, and another group to a modified task that included information about the unexpected behavior. That is, participants in the control group read conditionals similar to (21), were asked whether Alice was going to file her taxes online, and rated how probable it was that Alice knew about this opportunity and how probable it was that she cared. The conditionals featured various actions and the money that could be saved ranged from 40 to 320 euros. Participants in the other group read the same conditional statements, but were informed that Alice did not take the action that would have saved her money. They then rated how probable it was that Alice knew about the opportunity, and how probable it was that she cared.

The key question in this experiment was how participants in the second group would perceive the knowledge and preferences of Alice, as compared to the participants in the first group. The response depends in part on how much Alice could save. For the minimal amount (40 euros), participants typically considered that Alice knew she could save money, but could not be bothered to do so. For any other amount, participants either revised their belief that Alice knew about the opportunity, or their belief that she cared. There was a significant negative correlation ($r = -0.54$) between the two revisions, suggesting that the more participants adopted one of these explanations, the less they adopted the other – as would be predicted by a strategy of minimal belief revision.

In sum, an experiment using the ‘offline’ method (i.e., simply asking participants about their explanation of Alice’s nonconsequential behavior) suggested that people do not have a clear preference for either explanations based on beliefs, or explanations based on desire – although they may have a preference for the latter when the utility involved is small. To achieve a more subtle understanding of how people respond to utility violations, it is necessary to adopt an ‘online’ method, which tracks the cognitive processing of individuals when they first read about the nonconsequential behavior. This is the strategy adopted by Haigh and Bonnefon (2015), who recorded the eye movements of participants reading stories of nonconsequential behavior.

Haigh and Bonnefon (2015) presented participants with vignettes in which a character did not take a beneficial action, systematically manipulating whether the character knew or cared about this opportunity. For example, in the following vignette, Alice knows she can save money by renewing her insurance policy online, and cares about saving that money (emphasis and structure added):

(22)

a. Alice needed to renew her car insurance before it expired.

b. She knew that if she renewed over the Internet she would save £100.

c. Such a saving was important as she was struggling financially and desperately needed to save money.

d. After gathering together the relevant documents she renewed her policy over the phone.

e. The call lasted nearly half an hour.
In other versions of the vignette, Alice did not know about this opportunity: (22-b) was replaced with ‘She didn’t know that if she renewed over the Internet she would save £100.’ Still in other vignettes, Alice did not care about this opportunity: (22-c) was replaced with: ‘However, because she was very wealthy such a saving was not important to her.’ Accordingly, four different versions were constructed for each vignette.

The key idea in this study was to record eye movements in order to capture any disruptions of reading when participants reached the critical, emphasized portion of text in (22-d), which describes the nonconsequential behavior of Alice. The conditions in which reading is disrupted are informative with respect to whether rapid inferences about the behavior of Alice are driven by her desires, her beliefs, or both. For example, if reading was disrupted in conditions where Alice knows she can save money, regardless of her desire to do so, we would be in a position to conclude that online inferences are mostly driven by belief, rather than desire.

Results suggested that inferences were initially driven by desire alone, and only then moderated by belief. When readers reached the critical region in (22-d), the likelihood that they would look back to an earlier region of text (a disruption measure known as First Pass Regression Out) was significantly impacted by whether Alice desired to save money, but not by whether she knew she could. However, the total time readers spent on the critical region before moving on (a disruption measure known as Regression Path Time) was impacted by the interaction of belief and desire. That is, participants took longer to move on when Alice’s behavior contradicted her preferences, but only when she knew about the consequences of her behavior. These two effects suggest that readers made quick online inferences about the behavior of Alice based on her desires alone, but were then able to incorporate the beliefs of Alice in their processing, if they had to make sense of her nonconsequential behavior.

Overall, offline and online studies of reasoners’ reaction to utility violations suggest that the utility component of an argument from consequences (is the action beneficial to an agent?) is more cognitively salient than its knowledge component (does the agent know about the beneficial consequences?). Not only is the utility component processed first (as per the online results), but it can also weigh heavier in explaining nonconsequential behavior (as per the offline results). These results resonate with findings in both the reasoning and the persuasion literature (Corner et al., 2011; Evans et al., 2008; Hoeken et al., 2012) suggesting that when presented with an argument from consequences, people are more sensitive to differences in the desirability of consequences, than to differences in their likelihood.

7. Conclusion

People use arguments from consequences in order to convince others to take a course of action, but also to predict which course of action others will take. In this chapter, I have reviewed the work that my colleagues and I have conducted on this latter, predictive use of arguments from consequences. I believe this work to open new perspectives for argumentation research, by establishing a link between argumentation and theory of mind (i.e., mentalizing, or reasoning about the beliefs, desires and intentions of other individuals). Indeed, the arguments (from consequences) we can use to argue for a course of action, can also be used to predict that an individual will take this course of action, or to explain that an individual took this course of action. This dual nature of arguments from consequences naturally places them at the intersection of argumentation, reasoning, and theory of mind.
References


Predicting behavior on the basis of arguments from consequences