Taylor & Francis
Taylor & Francis Group

Why argue? Towards a cost–benefit analysis of argumentation

Fabio Paglieri* and Cristiano Castelfranchi

Istituto di Scienze e Tecnologie della Cognizione – CNR, Roma, Italy

This article proposes a cost–benefit analysis of argumentation, with the aim of highlighting the strategic considerations that govern the agent's decision to argue or not. In spite of its paramount importance, the topic of argumentative decision-making has not received substantial attention in argumentation theories so far. We offer an explanation for this lack of consideration and propose a tripartite taxonomy and detailed description of the strategic reasons considered by arguers in their decision-making: benefits, costs, and dangers. We insist that the implications of acknowledging the strategic dimension of arguing are far-reaching, including promising insights on how to develop better argumentation technologies.

Keywords: argumentation; cognitive models; cost–benefit; dangers; bounded rationality; efficiency

1. Introduction

If you are in a romantic relationship, you should be aware that arguing with your partner is not always the brightest idea – not to mention arguing with your in-laws, your teacher, your boss, or the policeman who just gave you a speeding ticket. Our frequent reluctance to argue is not a form of lamentable timidity, but rather the expression of strategic concerns: we do not engage in argument when doing so is likely to have an overall negative outcome. Arguing is a demanding activity for a variety of reasons, one that we do not undertake without purpose and due consideration – aside from specific and comparatively rare circumstances, e.g. argumentative outbursts. In particular, we estimate the costs, benefits, and dangers of argumentation in order to decide whether it is worth arguing with someone or for something. The same should apply to computational systems: building and using an argument-based technology requires careful assessment of the expected outcome and estimated costs of doing so. And argument-based technologies for open systems are likely to be more effective if they are not 'doomed to argue', but rather allowed to opt for different interactive modalities in different contexts, arguing only when it is expedient doing so.

In spite of the key relevance of a cost–benefit analysis of argumentation, the otherwise prolific field of argument theories did not produce much in this vein, neither with respect to human argumentation nor for argument-based technologies. We try to redress this lack of consideration here. It is important to appreciate that our aim is not to 'humanise' argumentation technologies, but rather to make them smarter according to the objective standards of efficiency, given the purposes and context of application of such technologies. Looking at how people strategically decide when and how to argue is instructive in this respect, i.e. to improve the performance of the computational system and not just for the sake of building Artificial Intelligence (AI) arguers which better approximates the performance of natural arguers. On the other hand, a

---

*Corresponding author. Email: fabio.paglieri@istc.cnr.it

formal and computational model incorporating strategic considerations on costs and benefits will also provide insights into human argumentation, since it will be a model (incomplete, simplified, and normative) of human argumentative planning. Having said this, in the rest of the paper we will focus mostly on the costs, benefits, and dangers of argumentation for natural agents, and only towards the end we will show how these concerns are relevant for and should be incorporated in the design of argumentation technologies.

Even though the cost–benefit character of argumentation has by and large been understudied in the literature,[1] there are interesting approaches that try to address this issue. In particular, Karunatillake and Jennings (2004; for further details, see also Karunatillake 2006) have done substantial work in the domain of negotiation, discussing whether arguing is always the optimal strategy to achieve effective negotiations among artificial agents, or should instead be applied with a modicum of caution. Their analysis supports the latter position, and their overall approach is largely compatible with our own, as witnessed by the following statement of purposes:

> Although argumentation-based negotiation can be effective at resolving conflicts, there are a number of overheads associated with its use. It takes time to persuade and convince an opponent to change its stance and yield to a less favourable agreement. It takes computational effort for both parties of the conflict to carry out the reasoning required to generate and select a set of convincing arguments, and to evaluate the incoming arguments and reason whether to accept or reject them. However, not all conflicts need to be resolved. Thus, for example, when faced with a conflict, an agent could find an alternative means to work around the situation; thereby *evading the conflict* rather than attempting to resolve it. (...) Given the overheads of argumentation, and the alternative methods available for overcoming conflicts (evade and re-plan), we believe it is important for agents to be able to weigh up the relative advantages and disadvantages of arguing, before attempting to resolve conflicts through argumentation (Karunatillake and Jennings 2004, p. 235).

We endorse the basic rationale behind their work and aim at broadening the scope of these considerations to (i) include also dialogical domains different from negotiation and (ii) take into account both costs and dangers of arguing, whereas Karunatillake and Jennings focus solely on the former; (iii) moreover, we explicitly address the comparison between cost–benefit balance in, respectively, natural and artificial arguers, instead of directly referring to the latter. Notably, Karunatillake and colleagues have recently extended their approach beyond negotiation to dialogue games in general, providing a detailed analysis of how the pattern of social influences characteristic of a given multi-agent society might affect the efficacy and effectiveness of argumentation as tools for conflict resolution (Karunatillake, Jennings, Rahwan, and McBurney 2009).

Strategic considerations on *how* to argue are explicitly addressed also by Amgoud and Maudet (2002), focusing at the level of single moves within an ongoing debate: since at each step of the dialogue the agent may be faced with several alternative moves, selecting the best move represents a key strategic concern. Some form of cost–benefit analysis enters the picture here, but the considerations bearing on the agent's decision are quite different from those we will be exploring in this paper. To start with, Amgoud and Maudet assumes that the agent is already engaged in argumentation and has to decide how to best proceed from there, whereas we mainly address the issue of *whether* arguing makes sense in the first place. Second, these authors focus mostly on benefits, understood as the strategic potential of a given move to foster the agent's dialogical goals (given a certain context), whereas costs are largely ignored or at best reduced to the (lost) benefits associated with alternative moves that the agent did not pursue. This implies confining their analysis to ideal agents that do not consume resources to argue, whereas we are interested to explore what happens to arguers who have to operate under constraints of limited resources and bounded rationality, as it happens to be the case for both human and artificial agents (on the importance of studying argumentation under assumptions of bounded rationality, see also Gabbay and Woods 2003; Paglieri and Woods in press).[2]

Finally, considerations on the computational costs of argumentation-based technologies are abundant in the literature on algorithmic and complexity issues related to such technologies (just to mention few key contributions, see Dung 1995; Dimopoulos, Nebel, and Toni 2002; Dunne and Bench-Capon 2002; Cayrol, Doutre, and Mengin 2003; Dung, Mancarella, and Toni 2007; Dunne 2007). Indeed, this research sub-topic has recently emerged as a major trend in studies on computational argumentation (Bench-Capon and Dunne 2007). However, in this case, the preoccupation is with the tractability of a given argumentation framework, whereas our concerns here are on a completely different scale: even assuming that a given argumentation framework is tractable, the individual agent still has to make a strategic decision on whether arguing is rational or not, given current goals, available resources and relevant context. Obviously, the tractability of the underlying argumentation framework impacts on the strategic considerations of the agent. If the framework makes argumentation intractable or typically very costly for the agent, the likelihood of considering such practice worthwhile is either non-existent or comparatively low. But the crucial point is that, even when argumentation is tractable and its costs are in principle affordable, it does not immediately follow that arguing is the best option for the agent. It is this crucial decision problem that so far has been largely overlooked in argumentation theories, both within and outside AI: our aim now is to move some preliminary steps to explore the issue.

We begin in Section 2 by stigmatising two biases that have plagued argumentation theories so far, somehow preventing a cost–benefit analysis of arguing. Then we provide working definitions of the main concepts used in this article, that is, benefits, costs, and dangers and contextualise such notions within the general framework of expected utility theory (Section 3). The following three sections (Sections 4–6) are dedicated to explore each one of these factors, while their implications for argumentation technologies are assessed in Section 7. We conclude by summarising our main results and outlining future directions for research on argumentative decision-making (Section 8).

2. Two biases in argumentation theories

Argumentation theories often make two crucial assumptions: that a failure of argumentation leaves the arguers in the same situation from where they started, and that dialogical goals are the paramount concerns for the arguers in deciding whether and how to engage in argument. We believe both assumptions to be substantially mistaken and highly misleading, for the reasons highlighted in what follows. This is relevant for our current purposes, because subscribing to these assumptions is what has impeded so far appreciating the importance of a cost–benefit analysis of argumentation.

2.1. *Against argumentative optimism*

Argumentation is typically aimed at improving a given dialogical situation in certain respects, either from the standpoint of one of the parties or according to some common frame of reference. Standard features that argumentation is intended to improve include, among others, the *credibility* of a given conclusion, the level of *agreement* between the parties, the *emotional state* of the arguers and of their social relationship, the *strategic* advantage of each arguer in the context of a broader dispute, and the *social reputation* of the arguer as a reliable source of information and/or counsel. Let us consider an arbitrary desirable property that argumentation aims at improving, be it the credibility of the conclusion, agreement between the parties, level of social inclusion, or something else, and assume that at t^0, before starting to argue, such a property has a certain value N. In argumentation theories it is often implicitly assumed that at t^1, after

argumentation occurred and *failed* to produce its intended outcome, the value of the desired property is still N. According to this view, argumentation can only make things better – even if it utterly fails, the parties are left with whatever they had when they started arguing.[3] This assumption we label as 'argumentative optimism', to stress that it offers too rosy a picture of the possible outcomes of argumentation.

Indeed, it takes no effort to see that argumentative optimism is patently false as far as everyday argumentation between real people is concerned. Argumentation can (and often does) produce *pejorative results* with respect to the initial state of the parties both because it precludes them from pursuing more rewarding courses of action, and because it ends in disaster, i.e. the situation that argumentation was supposed to remedy is in fact worsened by it. As we will explore in greater details in Section 6, there are five main aspects in which argumentation may result in a worsening of the original situation: poor arguments may 'backfire', i.e. end up undermining the credibility of the conclusion they were intended to support (Cohen 2005); argumentation may be conducive of disagreement, escalating the dispute between the parties rather than quenching it (Paglieri 2009); arguments on sensitive issues may create an emotional slippery slope in which both parties suffer affective losses with little or no epistemic gains (Gilbert 1999); ill-advised arguments may draw the attention of the counterpart to issues that serve to strengthen her position, thus undermining the proponent's own case (Paglieri and Castelfranchi 2006); finally, an argumentative failure may well have negative effects on the arguer's reputation as a good thinker, a reliable source, a skilled orator, or any combination of the above, in turn jeopardising his future argumentative ventures (Castelfranchi and Falcone 2010).

Aside from being false, argumentative optimism is also obstructive to the cost–benefit analysis of argumentation, since it leads to focus solely on benefits, overlooking both costs and dangers of the argumentative process. Therefore, in this article, we will distance ourselves from argumentative optimism and embrace a more open-ended view of argumentation. Given a desirable property p that argumentation targets, the value of p after arguing can be either *greater*, *equal*, or *smaller* than the value of p before argumentation occurred – in other words, there is no *a priori* restriction on the value of p at t^1, given the value of p at t^0. Usually, an increase in p is considered an indication of success, even if there may be exceptions: e.g. if the arguer's goal was to persuade the counterpart of a given claim and, after arguing, the credibility of such a claim is indeed increased, but not enough to lead the counterpart to embrace it, this still counts as a failure from the standpoint of the arguer, even if p has grown. Similarly, no change in p typically suggests a (non-catastrophic) failure, with the exception of cases where the arguer had strictly defensive goals, e.g. she wanted to succeed in keeping her position unchanged during the dispute, and indeed she managed to do so precisely by keeping p constant. A decrease in p, on the other hand, is always a sign of argumentative failure with various degrees of severity from moderate to apocalyptic: as an example of the latter, consider the case of a young couple who start discussing whether to get married soon, and instead find themselves breaking up few weeks later, after interminable quarrels on the matter.

2.2. *In favour of argumentative instrumentalism*

Dialogical goals are often *instrumental* to extra-dialogical goals: the car dealer wants to persuade you of the superiority of a given model in order to sell you that model, not just for the sake of discussion. In the terminology of the theory of goals developed by Castelfranchi and colleagues (Parisi and Castelfranchi 1981; Conte and Castelfranchi 1995; Castelfranchi 1998; Castelfranchi and Paglieri 2007), dialogical goals are rarely terminal, i.e. they tend not to be ends in themselves. There are many ways of skinning a cat and arguing is just one of them. This implies a form of *argumentative instrumentalism*: both the overall decision to argue and more

fine-grained choices of specific moves are ultimately affected by the extra-dialogical goals of the arguer. Argumentative instrumentalism has three facets: first, dialogical sub-goals are instrumental to some dialogical end (e.g. lauding the design of a given car is instrumental to persuade you of its superiority as a purchasing option), which in turn is instrumental to some extra-dialogical goal (e.g. selling you the car); second, the extra-dialogical goal has *priority* over the dialogical one, so if a given move is likely to foster the former but not the latter, the arguer should choose that move (e.g. observing that only few chosen people can afford to buy that model may be instrumental to selling you the car by appeasing your vanity, even if this point has no bearing on the alleged superiority of that particular model over others); third, a given move or set of moves may be instrumental to some other extra-dialogical goal of the arguer, different from the one which motivated to argue in the first place (e.g. the seller may try to impress an attractive female client with his elocution not as an effective mean to sell her the car, but rather hoping to get a date with the client). The prototypical goal structure of an argumentation plan is illustrated in Figure 1, distinguishing between dialogical goals (D-goals) and extra-dialogical ones (ED-goals).

This is in fact a general structure in a cognitive view of planning. The same figure might well represent my plan of cooking a delicious dinner to seduce Mary: some of the internal choices I will make might not be for reaching my culinary goal (i.e. for the dinner to be splendid), but rather aiming directly at the overall extra-culinary goal that motivated me to cook in the first place, that is, making progress in my relationship with Mary. For instance, I may choose to serve a very strong red wine, even if it is not well suited for the food I cooked, in order to get Mary a bit drunk and thus more amenable to concede herself to me tonight. Moreover, some other choices may be motivated by additional but completely independent goals, like not spending too much money for my attempt at seducing Mary or respecting my own dietary restrictions while cooking for her, and so on.

So what is somehow surprising is not that such a scheme applies to argumentation, but rather that so far the import of this (very simple) fact has been largely downplayed in argumentation theories (a notable exception is Gilbert 1997). Possibly due to their logical roots, argumentation studies have tended to treat argumentation as if it was a self-referential activity, with no tie to concerns that extended beyond the boundaries of the dialogue – or, at least, no tie that was relevant to understand the argumentative process under analysis. Goals are frequently discussed
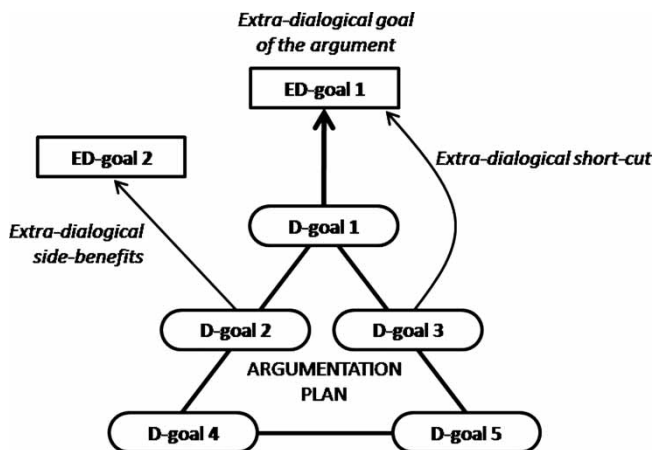

Figure 1. Dialogical and extra-dialogical goals of argumentation.

in argumentation theories, but this interest is typically limited to dialogical goals, as it happens in the well-known taxonomy of dialogue types suggested by Walton and Krabbe (1995; see also Walton 1998) or in pragma-dialectics (van Eemeren and Grootendorst 2004). Even those who advocated a broader and greater attention to goals in argumentation theories, like Gilbert (1997), tended to consider only the question of 'What is the goal of the argument?', whereas a better interrogative would be: 'What are the goals of the *arguer*, such that prompted him/ her to argue and to use certain arguments rather than others?' (see also Hample 2005). So what is needed is a *goal-oriented (motivated) framework for argumentation*, including not only a self-referential representation, but pointing also to other external or higher-level non-argumentative goals that contribute to the origin, the structure, and the outcome of the argumentative process.

We argue *for* something; and we decide to argue or not in view of something, motivated by some expected outcome. Only reference to these goals can explain not only *why* we argue (for what, with what motive), but in part also *how* we do so. In other words, accepting an instrumental view of argumentation is necessary to properly develop its cost–benefit analysis, since benefits and dangers of arguing can only be assessed with respect to the extra-dialogical goals that prompted us to engage in such activity.

3. Different facets of expected utility: benefits, costs, and dangers

The most straightforward (and straightforwardly computable) way of defining benefits, costs, and dangers is in terms of *utility*. Benefits refer to whatever goals (dialogical and/or extra-dialogical) of the agent the act of arguing accomplishes, including both the motivating goal of the argument and any additional positive side effects the agent might obtain. In other words, all the *positive utility* achieved by arguing can be lumped under the heading 'benefits'. Negative utility, on the other hand, is to be spelled out in two different categories: costs and dangers.[4] By costs, we mean the *negative utility produced by the process* of arguing in and by itself, given the agent's energetic balance and goals, and regardless the outcome of the argu-ment. This includes both *direct costs*, i.e. the resources employed to carry on the argumentative process, and *opportunity costs*, i.e. all other potentially beneficial activities the agent had to relinquish in order to argue. Finally, dangers refer to the *negative utility produced by some side effect or consequence* of arguing, such that it jeopardises some goals of the agent. By way of example, imagine a young woman who successfully wins a debate with her fiancé, after a long-winded, highly emotional confrontation that took place in front of their friends: having won the debate and showed to the audience she was right is a benefit for the woman; having committed a lot of time and effort in the process and thus missed the opportunity of going to the movies as planned are both costs; and having slightly humiliated her fiancé in public, therefore fuelling some resentment towards her is a danger she incurred.

Some additional remarks are in order. First, all these considerations enter strategic decision-making through the filter of the agent's expectations, insofar as they need to be considered ex ante, before the act of arguing takes place. Thus, here we are talking about the *expected utility* of argumentation, given the context and the agent's goals and beliefs. Second, there is no reason to assume that arguers calculate this expected utility by simultaneously and compre-hensively taking into consideration all factors that might affect it – indeed, decades of studies on decision-making suggest that people are likely to focus only on a limited sub-set of such factors, depending on their cognitive make-up and contextual features (Simon 1955, 1956; Kahneman, Slovic, and Tversky 1982; Rubinstein 1998; Gigerenzer and Selten 2001). Nevertheless, spelling out all the main factors that might (and, in some circumstances, should) affect the decision to argue is a worthy analytic effort, especially if the final aim is to provide suggestions on how

artificial agents may intelligently manage the same decision. Third, and more generally, broadly construing the arguer's strategic considerations in terms of expected utility theory does not commit us to the standard version of the theory. On the contrary, further sophistications (e.g. risk aversion, framing effects, mental accounting, etc.) can and must be integrated in the analysis, but they do not alter the basic rationale behind it. With all the *limitations* and *distortions* that arguers might suffer, they still try to ponder benefits, costs, and dangers of arguing, in order to decide whether to engage in it.

So the crucial challenge is to identify what factors are indeed relevant for making such a decision: what goals are likely to matter and/or should matter for the agent at that juncture and what beliefs determine expectations on satisfaction/frustration of such goals through argumentation. In what follows, we try to offer a principled classification of such factors in everyday argumentation between human agents. The emerging taxonomy will later be confronted with the priorities faced by several types of argumentation technologies.

4.   The benefits of arguing

Insofar as the dialogical goal of argumentation is instrumental to some extra-dialogical goal, there are at least two probabilities that modulate the agent's expected utility on arguing: the likelihood that argumentation will achieve its dialogical goal, e.g. persuading a potential buyer that an item is worth buying (probability of dialogical success, p^D henceforth), and the likelihood that this achievement will produce the satisfaction of the final extra-dialogical goal, e.g. having the person buying the item because he thinks it is worth buying, regardless how this conviction was achieved (probability of extra-dialogical efficacy, p^E from now on). These probabilities may well differ, and they often do. Imagine a determined but careful buyer, that is, a person who is committed to purchasing a certain item but also willing to buy it only when satisfied by the offer of the seller, even if this means taking a long time finalising the purchase. An inexperienced seller faced with such a client may well be sure that, if he manages to present a convincing case, he will sell the item, and yet be very skeptical about his chances of ever persuading the client (that is, $p^E > p^D$ in this case). Now consider a gullible but hesitant buyer, who is easily swayed by the opinion of other people, and yet rarely turns his fancies into concrete actions. An experienced salesman will see immediately that the client is very open to persuasion, but may still despair of ever making a sale with such a chronically undecided type (that is, here $p^D > p^E$).

These two values combine to determine a third probability, which is the one that really matters in the decision to argue: the likelihood that, by arguing, one will satisfy the final extra-dialogical goal (probability of extra-dialogical success, p^S henceforth). In anticipatory considerations of expected utility, p^S can be represented as the expected probability that the arguer's extra-dialogical goal will be achieved, assuming that argumentation is employed to that end. This clearly depends on how likely the arguer is to win the argument (p^D), and how likely is this victory to achieve the desired extra-dialogical end (p^E). How exactly to combine these two values depend on the assumptions made on their relationship: e.g. considering p^S as the conditional probability of p^E given p^D or treating p^D and p^E as referring to independent events and thus determining p^S as their product. The details on how p^S is (or should be) assessed are immaterial to our argument, since all we need to show is that both dialogical and extra-dialogical concerns concur to determine the likelihood of getting what we want through argument.

But again, what is it that we want? In other words, if p^S provides an approximation of likelihood of success, what is the *value* we assign to such an outcome and that we need in order to compute the expected utility function of arguing? This clearly can be answered only

on a case-by-case basis, since the extra-dialogical goals that motivate our arguments and that engender the value of the dialogical goal of winning such arguments vary wildly: from frivolous disputes on trivial matters that we can gladly afford to lose or even dismiss out of hand to matters of life or death, as it often happens for arguments occurring in the legal and medical domains. The stakes of arguing change dramatically across these scenarios and so does its expected utility. Still, there are general constraints that need to be taken into account in all cases, and they do not concern probability alone.

A key issue is the presence and quality of *alternatives*, that is, means other than argument through which the agent's extra-dialogical goal may be achieved. When alternatives are present, they will have their own expected utility, and a rational subject should weigh against each other's argumentative strategies and non-argumentative options that are viable and pick whatever it is that maximises his expected utility function or any other general mechanism used to ensure the agent's satisfaction. As we mentioned before, here we are not committed to any standard version of expected utility theory and have no problem in accommodating all the limitations and correctives that may seem convenient to adopt: e.g. boundedly rational arguers may be capable of considering only a limited subset of alternatives at any given time. The presence or absence of alternatives as well as their quality has a direct effect on the decision to argue, in particular on the assessment of expected benefits associated with argumentation. When the subject has no valid alternative to achieve his extra-dialogical goal other than arguing, renouncing to argue implies forsaking the extra-dialogical goal. In contrast, renouncing to argue in the presence of alternatives leave other options to the subject, such that he may still achieve his extra-dialogical goal after all. The obvious consequence is that in the former case, (no alternatives) the subject will think twice before forsaking argumentation as a viable means, lacking any other option. In other words, when the benefits of arguing are specific of such an activity (*argument-specific extra-dialogical goals*), they are more likely to motivate the agent's to argue.

Finally, the argumentative activity as a whole or specific argumentative moves made by the subject may well satisfy (or frustrate) other extra-dialogical goals of the agent, different from the one who would motivate arguing in the first place. As far as these additional *side benefits* (or *collateral damages*) are anticipated by the arguer, they might affect his decision to argue.[5] This kind of interplay frequently happens in educational settings: when a father has to make sure his son takes a medicine, mild physical coercion may actually be better than argumentation for that specific purpose – after all, it is faster, more reliable, and in a sense even less painful for the child, since, as soon as the medicine is administered, he is free to return to more pleasant undertakings, instead of having to listen to the long-winded arguments of his concerned parent. Nevertheless, the father may still (rationally) opt for arguing with his son because this will hopefully help realising another important extra-dialogical goal: namely, teaching the child the importance of medications as a remedy against illness.

To sum up: the probability of achieving the benefits that motivated the subject to argue (p^S) is assessed by combining the likelihoods of two distinct events, i.e. successfully carrying out one's argumentative efforts (p^D) and having the desired extra-dialogical effects as a consequence of that success (p^E). Given this probability, the subject's expected utility further depends, and crucially so, on the value of the extra-dialogical goal that would justify arguing, considering also alternative means to achieve that end, as well as alternative goals that arguing might foster or jeopardise. In Section 7, we will discuss the import of these considerations for computational models of argumentation. Now, it is time to move from the 'pros' column to the 'cons' list, in the balance sheet of the arguers: that is, we shall now explore the costs and dangers of arguing, as potential reasons to discourage such a course of action.

5. The costs of arguing

Arguing requires the subject to suffer some *direct costs*, in terms of time, breath, cognitive resources, social exposure, plus what economists call *opportunity costs* – all other things I could have done, if I was not stuck here arguing with you.[6] These costs steadily *increase as a function of argument duration*: the more we argue, the more resources we have to commit to it. The benefits of arguing, however, often do not have the same dynamics. Take persuasion as a case in point: if I stand to gain something from persuading you, whatever benefit I hope to achieve is independent from the time I spend achieving it, whereas the costs are not. The same applies to negotiation dialogues, while other types of argumentative interaction may have different dynamics. Following here Walton's (1998) taxonomy for the sake of clarity, in inquiry and eristic confrontation benefits are likely to increase proportionally with the time spent arguing: the more we argue, the more we improve our understanding of the topic (inquiry), or the more we vent our feelings towards each other and/or manage to hurt each other (eristic confrontation). So in these cases, the extra costs suffered may be more than compensated by the additional benefits of prolonged discussion. As for information-seeking dialogues and deliberation, in these cases benefit dynamics depend mainly on the details of the situation: if the information I seek is quite specific, the sooner I get it the better, whereas if I have broader curiosities, extended discussion may be more beneficial; similarly, if we are deliberating on something specific for which we know all the relevant facts, efficiency favours a relatively short discussion, whereas, if the issue is more complex, prolonged debate may be the best option.[7]

Let us first focus on instances where costs are likely to increase while benefits remain stable over time, e.g. in persuasion and negotiation. Here prolonged argumentation is usually not in the agent's best interest, and this effect is cumulative: the more efforts I devote to convince you, the more I stand to lose if, at the end of the day, you are not convinced. This may be true even when also the benefits are likely to increase over time (e.g. certain instances of inquiry and of eristic confrontation), provided the rate of growth for the benefits is lower than that for costs: when this happens, the two curves are bound to intersect, and when this happens it indicates a drastic shift in the arguer's cost–benefit balance. In fact, at this point, argumentation has turned into a bad deal, even if the arguer is fully successful in it, because its costs have exceeded its benefits. What is worst, the longer the argumentation is further prolonged after this point, the more severe will be the loss suffered by the arguer – again, regardless whether he wins or loses. More generally, what we have in these cases is that the *arguer's cost–benefit balance is deteriorating over time*, even in the event of a final victory (which is the optimistic assumption we are making here). Prolonging the argument is never a good strategy under these circumstances, as it is graphically illustrated in Figure 2.
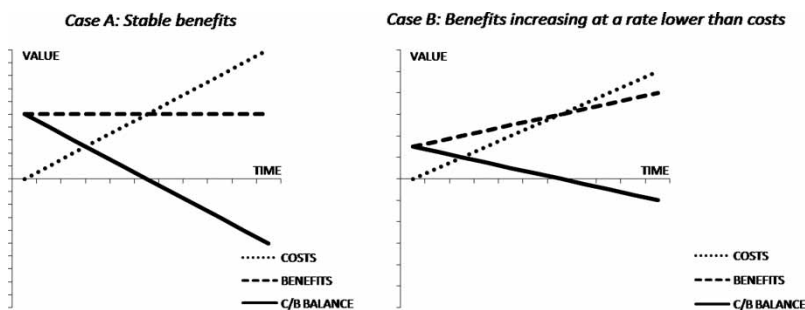

Figure 2. Cost–benefit balance as a function of time.

So far, these considerations indicate that, in general and with very few exceptions, a golden rule of argumentation is 'Keep it short!' – not only for reasons of rhetorical efficacy, but also from a cost–benefit point of view. How does this affect the decision to start arguing in the first place? Insofar as the agent has expectations on how hard it will be to conclude successfully a given argumentative engagement, he is also likely to have a rough estimate of the time necessary to do so: together with an intuition of the cost and benefit functions, this is all the information needed to judge whether arguing is going to be either a bad or a good deal, from the standpoint of its costliness.[8] By the way, we perform this kind of evaluation all the time: when we refuse to enter a debate because 'It would take me too much time to convince you', we are precisely acknowledging that argumentation would be anti-economical in this case; similarly, when we happen to give up during a dispute and walk away from it, this may be because we finally despair to persuade the counterpart, but often is just an indication that now we see the discussion as no longer being productive – that is, capable of yielding benefits higher than its costs.

Finally, time is relevant also in determining the perceived value of, respectively, costs and benefits, since the former are suffered immediately and with certainty by the arguer, whereas the latter, aside from being more or less uncertain, usually do not occur until the argument is over and often may be delayed even further (especially if we consider extra-dialogical benefits; see Section 4). It is a known fact that the utility of an outcome is discounted over delay, and we also know that temporal discounting in human agents is myopic – that is, we tend to over-value events that are temporally proximal and undervalue those that are temporally remote (Laibson 1997; Ainslie 2001; Berns, Laibson, and Loewenstein 2007). This has two substantial implications on the effect of costs in the decision to argue: (i) the costs of arguing are likely to *loom larger* than benefits and dangers, since the latter are typically both delayed and uncertain, whereas the former are immediate and certain; (ii) in argumentative decisions, the agent's preferences may exhibit the same kind of *dynamic inconsistency* that plague intertemporal choices in general (Strotz 1956; Kirby and Herrnstein 1995), so that the arguer considers arguing the optimal strategy while the moment of doing so is still far in the future, but when the time of engaging in debate comes closer at hand, suddenly its costs appear disproportionally high in consideration of its delayed expected benefit, and thus the agent may balk at it.

To sum up, costs of arguing include both direct consumption of valuable resources (time, cognitive effort, physical energy, etc.) and loss of opportunities to achieve alternative goals relevant to the arguer. Both categories of costs are bound to increase over time, while the benefits of arguing in most cases will either remain stable or increase at a lower rate: this implies that the utility function of arguing is decreasing over time, so that argumentation, if prolonged for more than a certain duration, will become a liability, regardless of the arguer's success in achieving his goal. These facts are (or should be) factored in the decision to argue or not, depending crucially on the expected time horizon of the discussion. Time not only happens to multiply the costs of arguing, but also acts as a filter in assessing the utility of costs *vs*. benefits, since the former are typically immediate and certain, whereas the latter are delayed and uncertain. Due to hyperbolic temporal discounting, this may lead the subject to overestimate costs and downplay benefits of arguing, as well as generating some temporal inconsistencies in his argumentative plan.

6. The dangers of arguing

Here we use the label 'dangers' to refer to the negative utility produced by a side effect or consequence of arguing, such that it jeopardises some goals of the agent. This is a technical and somehow restrictive use of the term, since in loose talk people consider as dangers also things that do not qualify under our definition: for instance, most notably, the 'danger' of losing the argument. However, this restrictive connotation is useful to distinguish dangers

from costs, as discussed in Section 3. Moreover, we are especially interested in the role of dangers in the decision to argue: this means that here we are considering dangers that are deliberately accepted (or rejected) by the agent, not just unforeseen events that happen to create negative utility. As a consequence, these argumentative dangers entail a certain level of *personal responsibility*: e.g. if things go badly, the arguer may regret it or blame himself for making a poor decision, and these negative self-evaluations constitute themselves a liability of arguing.

Some dangers are *specific* of a given instance of argumentation, since they refer to specific goals of the agent that arguing might put in jeopardy: e.g. arguing with one's in-laws may involve the danger of straining their relationship with the arguer's spouse, and this is likely to be an unwanted consequence of the debate and a (defeasible) reason not to get involved in it. Clearly, specific dangers of arguing need to be treated on a case-by-case basis. In contrast, here we are interested to briefly discuss *generic* dangers of arguing: that is, types of risks that are frequently associated with argumentation and thus recur in the arguer's decision-making across different contexts and situations.

First of all, an argument may 'backfire', to use Cohen's (2005) apt terminology, in the sense of *undermining the credibility of the conclusion it aimed to prove*. When an argument fails to prove its conclusions, this should not be taken to indicate that the audience is left with the same views on the matter they had before being exposed to the argument. Quite often, what happens is that a failed argument is taken as *evidence to the contrary*: the very fact that an argument for p failed is easily interpreted as an argument for $\neg p$ – which is clearly a highly unsatisfactory result for the arguer. Even more dramatically, once argumentative failure has occurred, the more effort the arguer put into the original argument, the more likely it is that the audience will consider this failure as good proof of the falsity of the intended claim. After all, if a determined and competent arguer, after putting so much effort into arguing for p, still failed to prove it, the most likely reason for that failure is that p is false, since there is no question as to the competence and motivation of the arguer. Like in a nightmarish subversion of all standard values, backfiring arguments may happen to retort against the arguer his very best efforts. This also implies that the dangers of arguing, in terms of negative effects on the credibility of the conclusion, increases as a function of duration, efforts, and quality of the argumentative process.[9]

A second relevant danger concerns the possibility that *argumentation will be conducive of disagreement* between the parties, rather than helping to solve it (Paglieri 2009). With very few exceptions, in argumentation theories it is taken for granted that agreement, partial or complete, is the desirable end-state of several types of dialogue (Walton and Krabbe 1995; Walton 1998), or even of argumentation in general (Gilbert 1997; Johnson 2000; van Eemeren and Grootendorst 2004). But argumentation, even when explicitly aimed at increasing the level of agreement between the parties, may fail spectacularly to do so. When this happens, it is not true (contra argumentative optimism; see Section 2.1) that people are left where they started: in many circumstances, their predicament will be worsened by having argued, in the sense of heightening their differences of opinion rather than solving them. Moreover, this happens because, and not in spite, of the nature of their argumentation: it is not due to a rational or procedural mistake made by one or both parties, but rather a natural consequence of their (correct) ways of arguing.

Four main factors contribute to make *escalation of disagreement* a potential danger of perfectly rational arguments (for further details on this point, see also Paglieri 2009). From an epistemological point of view, divergence of opinion *per se* is considered legitimate and possibly faultless, whereas keeping wrong beliefs in the face of (supposedly) valid arguments is regarded as a stubborn and irrational attitude: as a consequence, any disagreement that survives arguing, especially after long-winded discussion is likely to be less tolerated than any pre-argumentative

disagreement between the parties. Hence arguing may turn a legitimate dissensus into an irrational refusal to listen to reason, from the standpoint of both parties, thus worsening the quality of their disagreement. Second, the decreasing utility of argument over time (see Section 5) implies that the longer we argue, the more I stand to lose – and nobody likes losing more rather than less, especially is the chances of ever convincing the counterpart seem to become smaller and smaller the longer we talk. These economical pressures contribute to make us increasingly short-tempered and impatient with our interlocutors in prolonged debate, and this in turn increases the likelihood of escalation of disagreement. Third, arguing leads to explore in further detail a matter over which we disagreed from the onset, and this may well multiply the potential issues of dispute between us: if we disagree on whether Obama or McCain should have been elected as President and start arguing to settle the matter, it is likely that we will discover that we disagree also on many other things, some of them even more important than the original point of discussion, e.g. abortion rights and their extent. Finally, taking at face value what is said is, in most social contexts, the 'right' thing to do, i.e. what social conventions demand of people: if someone doubts your word on a given matter, the skeptic has better to produce some reason to back up that challenge, otherwise he will be labelled as unfriendly, rude, blunt, uncouth, opinionated, etc. Conversely, arguing, either to persuade the opponent of some claim he does not yet endorse, or to refute the reasons provided by the proponent, challenges this default trust, and thus requires some justification, whereas the same is not true for lack of argument (e.g. directly assenting or keeping silent).[10] This implies that the social framing of argumentation is, in and by itself, liable of being called into question as a breach of social trust, thus becoming the object of a further (meta)dispute on the legitimacy of arguing in that particular context.[11]

The damages associated with argumentation, and thus its dangers, can also concern the *emotional well-being* of the arguers. Arguing is rarely a neutral activity (Gilbert 1997), and the confrontational setting that it often implies may put a significant strain on people's mood: the arguers will typically feel compelled to succeed, while at same time being exposed to the critical cross-examination of the counterpart. If winning an argument is likely to cost you a friendship, or even your own peace of mind, you should (and will) think twice before entering that particular debate. Indeed, considerations of emotional risks enter the decision-making process of arguers: this is why we are very reluctant to argue about topics that are sensitive for ourselves or the counterpart, or both. For instance, only the most foolhardy husband will engage his wife in a debate on what is best to do during childbirth. Since breaching certain topics can be painful, and since this fact is often known to both parties, arguing about such topics appears as a deliberate aggression, cruelly aimed at the counterpart's 'soft spot'. This is likely to make the parties deaf to reason and turn even the most amicable suggestion into a fight. This is especially true with family and friends, and help explaining why some of the bitterest quarrels happen with people who are very dear to us. The feelings of friends and relatives are more easily hurt in a debate because venting open disagreement, as required by arguing, could be construed as a betrayal of legitimate social expectations and not because arguing in itself is especially hurtful for them. I was expecting your approval, and the fact that it is denied to me breaks this expectation, thus producing a certain amount of pain, disquiet, or discomfort. A stranger, lacking any specific expectation on whether you share or care for his views, is much less vulnerable to your argumentative attitude.

A further danger of arguing concerns drawing *unwanted attention* from the counterpart on topics that it would be in the arguer's best interest to keep hidden or out of focus. Argumentation works by raising certain issues and making explicit their connections with facts already proven, or at least provable: the thematic focus of the discussion shifts during the discussion and can be manipulated to try serving the best interest of each party – an operation known in pragma-dialectics

as strategic manoeuvring (van Eemeren and Houtlosser 2002; van Eemeren 2009). The danger inherent in such process is that, if mismanaged, it can end up undermining the arguer's own case: this happens when the arguer's inept manoeuvring makes the opponent notice things that are either damaging for the conclusion the arguer wants to prove (e.g. fuelling the opponent's counterarguments) or counterproductive for some other extra-dialogical goals of the arguer or both. Just to mention a very disastrous example, imagine a discussion where Adam tries to persuade Bob to lend him some money; when Adam argues that 'You should lend me the money, since you already did last year', Bob retorts with 'This reminds me – you never paid me back! Not only I will not lend you any new money, now I also demand immediate payment of your standing debt to me!'. Clearly, Adam should have considered more carefully all the dangers associated with invoking precedents in a discussion on debts.

Finally, argumentation is risky also inasmuch as a failure may well have *negative effects on the arguer's reputation*, not only as a skilled arguer but also as a rational thinker. Albeit reputation (Conte and Paolucci 2002) and trust (Castelfranchi and Falcone 2010) are important research topics in social psychology and distributed AI (for a review, see Sabater and Sierra 2005), they have rarely been focused in the context of argumentation theories – and even when they were, their role was mainly confined to specific argumentation schemes, such as argument from expert opinion or from testimony (Walton, Reed, and Macagno 2008). In contrast, reputation and trust have a much more pervasive impact on argumentation, since we frequently trust (or distrust) the arguments of a speaker mainly or even solely on the ground of his reputation. Moreover, a good reputation as an arguer entails benefits that go beyond argumentation: being right and being able to prove it is considered indicative of many positive qualities, such as lucidity of judgment, strength of character, clarity of reasoning, inventiveness, and fluency. On the contrary, argumentative failure is taken to show lack of these desirable features, and this is something we definitely do not want the general public to witness. Indeed, we would be happy to showcase our argumentative triumphs, if only there was no danger of making our dialectical debacles equally manifest. But there is such a danger, so caution suggests keeping our disagreements as private as possible, to avoid devastating effects on our reputation as rational agents (see also Paglieri 2009).

To sum up, aside from any specific risk that might be associated with a given piece of argument, arguing in general entails certain dangers, of which arguers are aware and that factor in their decision on whether and how to argue. These dangers include undermining one's own case, escalating the disagreement with the counterpart, worsening the emotional state of both parties, attracting the opponent's attention on unwanted aspects of the subject matter, and ruining the arguer's reputation due to failures in proving one's point. Not all these aspects need to matter for computational agents as they do for human arguers, yet many of them are absolutely crucial – as we shall discuss in the next section.

7. On the import for argumentation technologies

It is time to ponder to what extent consideration of benefits, costs, and dangers is crucial for technological applications of argumentation theories. Argumentation technologies constitute a vibrant field of research in AI, as witnessed by the recent inauguration of this journal, as well as by several scientific meetings (the workshop series ArgMAS, CMNA, Persuasive Technologies, and since 2006 the large-scale biennial conference COMMA), many other publications (aside from the proceedings of the conferences just mentioned, see also Reed and Norman 2004; Walton 2005; Rahwan and Simari 2009; and the special issue of *Artificial Intelligence* edited by Bench-Capon and Dunne (2007)), and some large-scale research projects (e.g. ASPIC – Argument Service Platform with Integrated Component, http://www.

argumentation.org/ and the support COST action on Agreement Technologies, http://www.
agreement-technologies.org/).

Broadly speaking, the relevance of argumentation theories for AI is threefold, spanning both
programming issues, human–computer interaction (HCI), and computer-mediated communi-
cation (CMC). Concerning computer programs and protocols, especially in distributed
systems, argumentation techniques are used to guarantee a good balance between flexibility
and stability (Rahwan and Simari 2009): each software agent is endowed with more or less
sophisticated argumentative skills to interact with other agents and solve specific problems,
e.g. information retrieval, automated decision-making, negotiation, etc. As far as interfaces
are concerned, argumentation theories provide inspiration on how to design dialogue systems
that human users will perceive as natural and convincing, and that will be able to engage in
conversations of non-trivial complexity, e.g. providing appropriate replies, asking pertinent
questions, offering suggestions in light of the user's needs and tastes, etc. (Guerini, Stock,
and Zancanaro 2007; Mazzotta, de Rosis, and Carofiglio 2007). As for CMC applications,
argumentation theories offer insights on how to support online interactions between end-
users, both in synchronous and asynchronous modalities and with different levels of tutoring,
in order to favour an orderly and productive development of the debate (Sillince and Saeedi
1999; Andriessen, Baker, and Suthers 2003), and also help defining criteria for qualitative com-
parison between online discussion and face-to-face interaction (Weinberger and Fischer 2006;
Joiner, Jones, and Doherty 2008).

As discussed above, the assumption 'more argument, better results' is mistaken as a general
rule: in addition, it may also be very misleading, when it comes to argumentation technologies.
Let us start from the application of argument theories to communities of artificial agents inter-
acting with each other. These agents typically have no special concern for their personal
emotional welfare, and they do not care for social image and for being in good terms with
their fellows: indeed, they have no feelings that can be hurt, so there is nothing personal at
stake in their interactions. Nonetheless, their decision to engage in argumentation should still
be guided by (i) energetic constraints, so that they should not pursue argumentation if the
expected benefits are lower than the likely costs and (ii) some measure of foresight on the
extra-dialogical consequences of arguing, so as to weigh benefits and dangers of argumentation
in deciding whether to engage in it. In contrast, building artificial arguers that are blind to such
elementary precautions may result in computational inefficiency, either because the agents lose
themselves in the subtleties of prolonged debate, thus slowing or sidetracking the performance of
the system as a whole (something both programmers and users are keen to avoid) or because the
agents fail to consider the risks of arguing and therefore produce a result which is below standard
(e.g. obstinate agents that keep using poor arguments may well undermine the credibility of
the claim they intend to support, due to backfiring effects). The right rule in this case is to incor-
porate strategic considerations in the agent's decision to either argue or let the matter rest, and
these considerations should somehow approximate those discussed above.

A cost–benefit analysis of argumentation is even more important when it comes to design
users' interfaces based on argumentation technologies or software for supporting computer-
mediated arguments. In HCI and CMC, it is paramount that the software interface is capable
of eliciting agreement in its users, both as a baseline to ensure smooth interaction in general
(users should not get mad at the program or at each other), and concerning specific functional-
ities of the system (users should be willing to accept or at least consider the system's
suggestions). The former type of agreement is instrumental to the latter, especially in those
cases where (i) the software is supposed to make decisions or suggestions in the best interest
of the user and (ii) this may not always coincide with what the user would do, left to his
own devices. Expert systems capable of supporting human decision-making constitute very

promising applications in AI, but they need to be designed keeping in mind the users' argumentative attitudes. In particular, users should not be subjected to questioning or correction unless this is strictly necessary – and here 'necessary' means not only that the user may be about to make a mistake, but also that such a mistake is *relevant* for the user's current goals. In these applications, it is of the utmost importance to include an 'agreement strategic manager' as part of the artificial system, to ensure that the user is not bothered all the time with puny arguments or fastidious suggestions, and he is instead best prepared to keep an open-minded attitude towards the system's indications when these are truly relevant. In short, a computer should not be designed to be quarrelsome or pedantic, lest it ends up like the Talking Cricket in *The Adventures of Pinocchio*, with an infuriated user throwing a mallet at it.

As we discussed, a cost–benefit analysis of argumentation involves a rich variety of considerations: which ones are to be significant for a given argumentation technology depend on the specifics of that technology, including its intended context of application. For instance, an argumentation-based protocol for negotiation among autonomous artificial agents needs to be concerned with the reputational effects of arguing only to the extent that it incorporates some kind of agent's reputation (for arguments on why it should, see Conte and Paolucci 2002); similarly, an expert system with the sole purpose of advising end-users on the manifest structure of their arguments need not fear to hurt their feelings – albeit, it would still need to have some inkling on the users' goals, as to prioritise the advice given to them. However, even if different technologies have different needs with respect to the assessment of benefits, costs, and dangers of arguing, some general considerations may be distilled from the analysis of human argumentation, and offered to computer scientists as a sort of (preliminary and incomplete) checklist in designing future argumentation technologies:[12]

(1) The instrumental link between the dialogical goal of the argument and the *extra-dialogical goal that motivates arguing* in the first place should be preserved by the system, to ensure the proper degree of flexibility: in the presence of better alternative means, the system should not be fixated on arguing at all costs.

(2) The system should have information on the connections between argumentative moves and the user's relevant extra-dialogical goals (other than the one prompting him to argue), to assess on the fly potential *side benefits* and/or *collateral damages* of arguing: this is necessary to ensure that the utility function incorporated by the system is comprehensive enough, in view of the user's preferences.

(3) Some knowledge on *alternative means* to achieve the extra-dialogical goal of arguing is crucial, in order to (i) make the choice of arguing in the proper strategic context, (ii) correctly assess how essential is argumentation to achieve the user's objective, and (iii) correctly assess opportunity costs suffered by the system, the user, or both.

(4) To ensure a convenient cost–benefit balance, the system needs accurate and updated information on the *costs suffered by arguing* (as well as some *predictive mechanism* capable of anticipating such costs before it is too late to avoid or reduce them), both in terms of lost opportunities (see above) and as resources consumed for performing the task over time.

(5) Since time is a precious commodity, for both the system and its users, some *temporal discounting mechanism* should be incorporated in the utility function, using either an hyperbolic function (if approximating human's preference dynamics is the aim) or an exponential one (if avoiding inconsistencies in decision-making is the paramount concern).

(6) If appropriate, the system may also be designed to *correct*, rather than replicate, known biases of human decision-making concerning costs and benefits: e.g. ignoring sunk costs

and/or looking for them in the user's decisions, to suggest how to avoid or neutralise them; discounting future utility exponentially and thus avoiding dynamic inconsistencies (see above); and so on.

(7) Finally, the system needs anticipatory information on the inherent *dangers* of arguing, either for the user or for the system itself, depending on the relevant context: e.g. a support system for CMC should pay great attention to risks of disagreement escalation, emotional disruption, and loss of reputation, whereas an expert system designed to assist in writing sound argumentative essays will put greater emphasis on dangers of backfiring arguments and thematic misdirection.

This list is not intended to be complete, and it is deliberately pitched at a high level of generality – anything more specific would in fact require discussing a specific argumentation technology as a case study, and this is not our purpose here. We rather prefer to put forward a more general message: abstracting from considerations of costs, benefits, and dangers in argumentation technologies leads not only to *lack of cognitive realism* (designing artificial agents that have little in common with human arguers), but also to *limitations in the computational efficacy* of the system – these agents miss valuable strategic resources that help human arguers to optimise their argumentative practice. In the absence of some cost–benefit compass, artificial arguers are condemned to charge blindly through the complex field of argumentation techniques. Our plea is to start building that compass, keeping an open mind on virtues and vices of arguing.

8. Conclusions and future work

The take-home message of this paper is that *arguing is not always the brightest solution* for the agent's predicament, be it a human or an artificial arguer. In everyday discourse, arguers are all keenly aware of this fact, and thus they treat the decision to argue as a *strategic problem*. We suggest more attention should be paid in argumentation theories to this strategic dimension of arguing and that lessons learned from studying the decision-making process of human arguers will provide useful guidance for the development of more effective argumentation technologies. In this paper, we moved some preliminary steps in that direction, by providing the bare outline of the strategic considerations that enter the decision to argue and tentatively discussing their implications for argument-based computational systems. The main points we considered include the need to *put every argumentative act in the broader context of the agent's practical reasoning* (typically, we do not argue just for the sake of it, but rather to foster further ends, and it is with respect to such extra-dialogical goals that the arguer's decision-making is to be understood), the possibility that *argumentation might produce disastrous results* (in the precise sense of worsening the arguers' initial predicament according to some standard of quality, e.g. credibility of the conclusion, level of agreement between the parties, etc.) and the threefold concerns that shape the arguer's decision – namely, expected *benefits*, likely *costs*, and foreseen *dangers*.

While the cost–benefit analysis of argumentation is without doubt still in its infancy, we hope to have shown that it has much potential to blossom into a rich, relevant, and highly interdisciplinary domain of inquiry. There is much virgin territory to be explored and charted, and we believe future studies should cover all levels of analysis. Taking the liberty of acting as pathfinders, here we suggest some promising 'tracks' that we see opening in front of argumentation scholars:

(1) *Theoretical track*: Efforts should be made to develop the somehow haphazardly observations presented in this paper and elsewhere in the literature (see Section 1) into a coherent and comprehensive *theory of argumentative decision-making*. Moreover, the

significance of argumentative decision-making should be discussed and clarified in the broader context of argumentation theories, that so far paid only lip service to this topic.

(2) *Empirical track*: The map of the factors affecting the arguer's decision is still too coarse-grained to provide effective guidance, and *empirical verification* is badly needed. We should tease out specific factors, to assess empirically (either by experiment or via ethnographic observation) what weight they have in the arguer's decision across different dialogical contexts and what kind of biases or distortions may also affect argumentative decision-making.[13]

(3) *Computational track*: We should design and test *argumentation technologies explicitly inspired by cost–benefit considerations*, possibly starting from well-defined and narrow domains where only few factors are likely to be relevant, and later on scaling up to open systems for complex, dynamic applications. The problems encountered and the results achieved will in turn provide indications on what is truly relevant for the decision-making of artificial arguers, as opposed to natural ones.

As mentioned, the road ahead is long, unexplored, and full of bifurcations. Nevertheless, the benefits of taking it far exceed its costs and dangers. Studying argumentation without considering the decision to argue is like reverse engineering a bicycle without understanding its purpose – a hopeless endeavour at its worst, a mere intellectual exercise at its best. Argumentation theories should have a higher aim, and they will need to keep the arguer's decisions in the loop to successfully hit the target.

Acknowledgements

Notes

1. This lack of interest may largely depend on the fact that 'mainstream' work on computational models of argument does not consider the dialogical or procedural aspects of argument much at all, but focuses its attention on the rather limited view of argumentation as a method for defeasible reasoning with inconsistent information. For instance, the otherwise excellent monograph by Besnard and Hunter (2008) provides a vivid illustration of this narrow research focus. Thus, strategic consideration of costs and benefits still remains ahead of much of the field, since the issue becomes relevant only after acknowledging argumentation as an essentially dialogical process, involving communication between multiple agents.

2. Notably, Amgoud and Maudet (2002) are fully aware of this limitation in their work, and they freely admit the importance of extending similar considerations to resource-bounded agents:

    Ideal agents compute (...) for free. But computation takes resources, and for instance spending too much time trying to determine the acceptability of an argument may be a poor strategy. What if the agent cannot conclude within the bound of the resources? (...) The role of strategy is even more crucial when taking into account the resource-bounded nature of agents' (p. 406).

3. A survey of the literature in argumentation theories, illustrating both how widespread argumentative optimism is in general and some notable exceptions to it, is to be found in Paglieri (2009).

4. This is not a necessary distinction, insofar as both costs and dangers refer to negative utility and, in some circumstances, may even be difficult to tell apart in practice. However, we will endorse the distinction in this paper, since it is both useful for explanatory purposes, and nicely reflects different concerns arguers may have, when pondering the likely outcome of argumentation.

5. Clearly, expectations on likelihood will play a role again here: the probability that a given side benefit or collateral damage will be incurred by arguing is not necessarily 1, and arguers may take into account also this probabilistic factor in their cost–benefit analysis.

6. By definition, opportunity costs are negligible or even null for agents that do not have much else to do, aside from arguing. This is rarely the case with humans, but it could easily happen with software agents solely designed to argue in certain domains for a given set of purposes. Even in this case, however, it remains true that (i) these single-minded argumentative agents suffer direct costs due to argumentation and (ii) these costs increase over time – the longer the argument, the higher the costs.

7. Another exception to the law of constant benefits, which is independent from the type of dialogue, is the case where arguing provides the subject with some kind of *desirable social exposure*, e.g. it increases his reputation within a relevant social group. In similar cases, the social benefits of arguing are likely to increase over time, albeit the overall dynamics will not be linear – every audience will sooner or later tire to listen to even the most brilliant orator, and good speakers should know how to quit when their social image is at peak.

8. Once more, let us reiterate that economic considerations of costs may well be subject to several biases, and that we are not championing here any ideal of the so-called 'perfect rationality' for arguers. For instance, quite often we find ourselves keeping on a heated discussion even when it is perfectly clear that (i) the argument has outlived its usefulness, (ii) a satisfactory conclusion will never be reached, and (iii) on the contrary things will systematically worsen through prolonged debate. None-theless, we keep arguing with perverse gusto in such quandary, and one likely reason for it (aside from plain stubbornness) is *sunk costs*: that is, our tendency to consider past resources already spent on a given task as relevant factors in our decision to either persevere or abandon such a task, when this decision should instead depend only on present and future costs and benefits (Arkes and Blumer 1985). Paraphrasing the well-known financial advice, one should not throw new words into a bad argument, and yet we often do.

9. Notice that here only what we stand to lose in case of failure is increasing, not the likelihood of failure itself – that, on the contrary, will probably be affected positively by argument's efforts, quality, and, to some extent, duration. This implies that the arguer has to consider a delicate trade-off (one among many) between the magnitude of his potential epistemic losses, and the probability these losses will in fact occur.

10. Here we endorse a Reidian view on trust in social testimony (Reid 1970; see also Lewis 1969; Grice 1989; Goldman 1999): its applicability to argumentation is extensively discussed elsewhere (Paglieri 2007; Paglieri and Castelfranchi in press; Paglieri and Woods in press), so here it will be taken for granted. For a general outlook on social epistemology, see Govier (1997, 1998), Goldman (1999), and Origgi (2004).

11. The legitimacy of argument is also deeply affected by more specific socio-cultural constraints, as discussed by Paglieri (2009). However, the effects of social norms and cultural conventions on argumentation lack the level of generality that interests us here, so we will skip further reference to them.

12. It is worth noticing that many of the items on this list require some knowledge or information to be made available and modelled computationally in order to provide decision support. This may well conflict with the fact that, in expert systems, there is a 'knowledge-acquisition bottleneck' which can entail high costs. Insofar as considerations of costs and benefits should be global and not local, this raises the issue of what are the costs for those knowledge acquisition tasks needed to support the users in assessing the costs of arguing. Whenever knowledge acquisition costs are likely to be higher than argumentation costs, argument technology may actually benefit from ignoring the latter. Hence, our recommendations should not be taken as the Ten Commandments, and their feasibility (or lack thereof) should be assessed on a case-by-case basis.

13. Here future studies will profit from existing empirical research on argumentation strategies (for a review, see Hample 2005), in particular the work done by Hample and colleagues on *argument editing* (Hample and Dallinger 1990, 1992; Hample, Werber, and Young 2009), that is, the decision whether to use a specific argument that occurs to the arguer's mind, in light of its foreseen

consequences for the dialogue in progress. Hample's analysis concerns local decisions within an ongoing argumentation, whereas our focus here was on the global decision of either engaging in argument or avoiding it: nevertheless, there are important commonalities between these two lines of research, as discussed elsewhere (Paglieri 2009; Hample, Paglieri and Na in press).

References

Ainslie, G. (2001), *Breakdown of Will*, Cambridge: Cambridge University Press.

Amgoud, L., and Maudet, N. (2002), 'Strategical Considerations for Argumentative Agents', in *Proceedings of the 9th International Workshop on Non-Monotonic Reasoning (NMR)*, eds. S. Benferhat and E. Giunchiglia, Toulouse: IRIT, 399–407.

Andriessen, J., Baker, M., and Suthers, D. (eds.) (2003), *Arguing to Learn. Confronting Cognitions in Computer-supported Collaborative Learning Environments*, Dordrecht: Kluwer.

Arkes, H., and Blumer, C. (1985), 'The Psychology of Sunk Cost', *Organizational Behavior and Human Decision Process*, 35, 124–140.

Bench-Capon, T., and Dunne, P. (2007), 'Argumentation in Artificial Intelligence', *Artificial Intelligence*, 171, 619–641.

Berns, G., Laibson, D., and Loewenstein, G. (2007), 'Intertemporal Choice – Toward an Integrative Framework', *Trends in Cognitive Sciences*, 11, 482–488.

Besnard, P., and Hunter, A. (2008), *Elements of Argumentation*, Cambridge: MIT Press.

Castelfranchi, C. (1998), 'Modelling Social Action for AI Agents', *Artificial Intelligence*, 103, 157–182.

Castelfranchi, C., and Paglieri, F. (2007), 'The Role of Beliefs in Goal Dynamics: Prolegomena to a Constructive Theory of Intentions', *Synthese*, 155, 237–263.

Castelfranchi, C., and Falcone, R. (2010), *Trust Theory: A Socio-Cognitive and Computational Approach*, Chichester: John Wiley & Sons.

Cayrol, C., Doutre, S., and Mengin, J. (2003), 'On Decision Problems Related to the Preferred Semantics for Argumentation Frameworks', *Journal of Logic and Computation*, 13, 377–403.

Cohen, D. (2005), 'Arguments that Backfire', in *The Uses of Argument*, ed. D. Hitchcock, Hamilton: OSSA, pp. 58–65.

Conte, R., and Castelfranchi, C. (1995), *Cognitive Social Action*, London: UCL Press.

Conte, R., and Paolucci, M. (2002), *Reputation in Artificial Societies. Social Beliefs for Social Order*, Boston: Kluwer.

Dimopoulos, Y., Nebel, B., and Toni, F. (2002), 'On the Computational Complexity of Assumption-based Argumentation by Default Reasoning', *Artificial Intelligence*, 141, 57–78.

Dung, P. (1995), 'On the Acceptability of Arguments and Its Fundamental Role in Non-monotonic Reasoning, Logic Programming, and N-Person Games', *Artificial Intelligence*, 77, 321–357.

Dung, P., Mancarella, P., and Toni, F. (2007), 'Computing Ideal Sceptical Argumentation', *Artificial Intelligence*, 171, 642–674.

Dunne, P. (2007), 'Computational Properties of Argument Systems Satisfying Graph-theoretic Constraints', *Artificial Intelligence*, 171, 701–729.

Dunne, P., and Bench-Capon, T. (2002), 'Coherence in Finite Argument Systems', *Artificial Intelligence*, 141, 187–203.

Gabbay, D., and Woods, J. (2003), *Agenda Relevance: A Study in Formal Pragmatics*, Amsterdam: North-Holland.

Gigerenzer, G., and Selten, R. (eds.) (2001), *Bounded Rationality: The Adaptive Toolbox*, Cambridge: MIT Press.

Gilbert, M. (1997), *Coalescent Argumentation*, Mahwah: LEA.

Gilbert, M. (1999). 'Agreement/Disagreement', in *Argumentation at the Century's Turn: Proceedings of OSSA 1999* eds. H. Hansen, C. Tindale, and E. Sveda, Windsor: OSSA, CD-ROM, pp. 1–9.

Goldman, A. (1999), *Knowledge in a Social World*, Oxford: Oxford University Press.

Govier, T. (1997), *Social Trust and Human Communities*, Montreal: McGill-Queen's University Press.

Govier, T. (1998), *Dilemmas of Trust*, Montreal: McGill-Queen's University Press.

Grice, P. (1989), *Studies in the Way of Words*, Harvard: Harvard University Press.

Guerini, M., Stock, O., and Zancanaro, M. (2007), 'A Taxonomy of Strategies for Multimodal Persuasive Message Generation', *Applied Artificial Intelligence Journal*, 21(2), 99–136.

Hample, D. (2005), *Arguing: Exchanging Reasons Face to Face*, Mahwah: Lawrence Erlbaum Associates.

Hample, D., and Dallinger, J. (1990), 'Arguers as Editors', *Argumentation*, 4, 153–169.

Hample, D., and Dallinger, J. (1992), 'The Use of Multiple Goals in Cognitive Editing of Arguments', *Argumentation and Advocacy*, 28, 109–122.

Hample, D., Paglieri, F., and Na, L. (in press), 'The Costs and Benefits of Arguing: Predicting the Decision Whether to Engage or Not', in *Proceedings of ISSA 2010*, eds. F. van Eemeren, B. Garssen, A. Blair, and G. Mitchell, Amsterdam: Sic Sat.

Hample, D., Werber, B., and Young, D. (2009), 'Framing and Editing Interpersonal Arguments', *Argumentation*, 23, 21–37.

Johnson, R. (2000), *Manifest Rationality: A Pragmatic Theory of Argument*, Mahwah: Lawrence Erlbaum Associates.

Joiner, J., Jones, S., and Doherty, J. (2008), 'Two Studies Examining Argumentation in Asynchronous Computer Mediated Communication', *International Journal of Research & Method in Education*, 31, 243–255.

Kahneman, D., Slovic, P., and Tversky, A. (1982), *Judgment Under Uncertainty: Heuristics and Biases*, New York: Cambridge University Press.

Karunatillake, N. (2006), 'Argumentation-Based Negotiation in a Social Context', unpublished Ph.D. thesis in Computer Science', University of Southampton, School of Electronics and Computer Science.

Karunatillake, N., and Jennings, N. (2004). 'Is it Worth Arguing?' in *Argumentation in Multi-Agent Systems (Proc. of ArgMAS'04)*, eds. I. Rahwan, P. Moratïs, and C. Reed, Berlin: Springer-Verlag, pp. 234–250.

Karunatillake, N., Jennings, N., Rahwan, I., and McBurney, P. (2009), 'Dialogue Games that Agents Play within a Society', *Artificial Intelligence*, 173, 935–981.

Kirby, K., and Herrnstein, R. (1995), 'Preference Reversal Due to Myopic Discounting', *Psychological Science*, 6, 83–89.

Laibson, D. (1997), 'Golden Eggs and Hyperbolic Discounting', *Quarterly Journal of Economics*, 112(2), 443–477.

Lewis, D. (1969), *Convention: A Philosophical Study*, Cambridge: Harvard University Press.

Mazzotta, I., de Rosis, F., and Carofiglio, V. (2007), 'Portia: A User-adapted Persuasion System in the Healthy Eating Domain', *IEEE Intelligent Systems*, 22(6), 42–51.

Origgi, G. (2004), 'Is Trust an Epistemological Notion?' *Episteme*, 1, 61–72.

Paglieri, F. (2007), 'No More Charity, Please! Enthymematic Parsimony and the Pitfall of Benevolence', in *Dissensus and the Search for Common Ground*, eds. H. Hansen, C. Tindale, R. Johnson, and A. Blair Winder: OSSA, CD-ROM, pp. 1–27.

Paglieri, F. (2009), 'Ruinous Arguments: Escalation of Disagreement and the Dangers of Arguing', in *Argument Cultures: Proceedings of OSSA 2009*, eds. H. Hansen, C. Tindale, R. Johnson, and A. Blair. Windsor: OSSA, CD-ROM.

Paglieri, F., and Castelfranchi, C. (2006), 'The Toulmin Test: Framing Argumentation within Belief Revision Theories', in *Arguing on the Toulmin Model*, eds. D. Hitchcock and B. Verheij, Berlin: Springer, pp. 359–377.

Paglieri, F., and Castelfranchi, C. (in press), 'In Parsimony We Trust: Non-cooperative Roots of Linguistic Cooperation', in *Perspectives on Language Use and Pragmatics,* ed. A. Capone, München: Lincom Europa.

Paglieri, F., and Woods, J. (in press), 'Enthymematic Parsimony', *Synthese*.

Parisi, D., and Castelfranchi, C. (1981), 'A Goal Analysis of Some Pragmatic Aspects of Language', in *Possibilities and Limitations of Pragmatics,* eds. H. Parret, M. Sbisà, and J. Verschueren, Amsterdam: John Benjamins, pp. 551–567.

Rahwan, I., and Simari, G. (eds.) (2009), *Argumentation in Artificial Intelligence*, Berlin: Springer.

Reed, C., and Norman, T. (2004), *Argumentation Machines: New Frontiers in Argument and Computation*, Berlin: Springer.

Reid, T. (1970), *An Inquiry into the Human Mind,* ed. by T. Duggan, Chicago: University of Chicago Press.

Rubinstein, A. (1998), *Modelling Bounded Rationality*, Cambridge: MIT Press.

Sabater, J., and Sierra, C. (2005), 'Review on Computational Trust and Reputation Models', *Artificial Intelligence Review*, 24, 33–60.

Sillince, J., and Saeedi, M. (1999), 'Computer-mediated Communication: Problems and Potentials of Argumentation Support Systems', *Decision Support Systems*, 26, 287–306.

Simon, H. (1955), 'A Behavioral Model of Rational Choice', *Quarterly Journal of Economics*, 69, 99–118.

Simon, H. (1956), 'Rational Choice and the Structure of the Environment', *Psychological Review*, 63, 129–138.

Strotz, R. (1956), 'Myopia and Inconsistency in Dynamic Utility Maximization', *Review of Economic Studies*, 23(3), 165–180.

van Eemeren, F. (ed.) (2009), *Examining Argumentation in Context: Fifteen Studies on Strategic Maneuvering*, Amsterdam: John Benjamins.

van Eemeren, F., and Grootendorst, R. (2004), *A Systematic Theory of Argumentation: The Pragma-dialectical Approach*, Cambridge: Cambridge University Press.

van Eemeren, F., and Houtlosser, P. (2002), 'Strategic Maneuvering in Argumentative Discourse: Maintaining a Delicate Balance', in *Dialectic and Rhetoric: The Warp and Woof of Argumentation Analysis,* eds. F. van Eemeren and P. Houtlosser, Dordrecht: Kluwer Academic, pp. 131–159.

Walton, D. (1998), *The New Dialectic: Conversational Contexts of Argument*, Toronto: University of Toronto Press.

Walton, D. (2005), *Argumentation Methods for Artificial Intelligence in Law*, Berlin: Springer.

Walton, D., and Krabbe, E. (1995), *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*, Albany: SUNY Press.

Walton, D., Reed, C., and Macagno, F. (2008), *Argumentation Schemes*, Cambridge: Cambridge University Press.

Weinberger, A., and Fischer, F. (2006), 'A Framework to Analyze Argumentative Knowledge Construction in Computer-supported Collaborative Learning', *Computers and Education*, 46, 71–95.