Commentary on: Frank Zenker's "Know thy biases! Bringing argumentative virtues to the classroom"

STEVE OSWALD

Cognitive Science Centre
Dept of Language and Communication Sciences
University of Neuchâtel
Espace Louis-Agassiz 1, 2000 Neuchâtel
Switzerland
steve.oswald@unine.ch

1. INTRODUCTION

For quite some time now, cognitive and social psychology have been documenting the pervasiveness of cognitive biases in reasoning and decision making (among other dimensions of human cognition) but also how people struggle to get rid of them – achieving own bias awareness is a difficult task. Zenker's "Know thy biases!" describes a concrete teaching and learning activity (TLA) meant to achieve debiasing. Specifically, his contribution deals with the *polarization effect*, which causes people to overestimate the magnitude of opinion difference between them and their opponents. The TLA under discussion is grounded on the assumption that traditional classroom activities such as 'showing and telling' fail to enable students to spot their own biases, but that, according to empirical studies (e.g. Pronin, Puccio, & Ross, 2002), putting them in situations where they are asked to consider adversary positions seriously – and thus evaluate them – can yield better results.

While I have little doubt that the purpose of Zenker's paper is convincingly fulfilled, I believe that his description of the proposed TLA is ideally suited to kick-start a discussion about the cognitive underpinnings of reasoning and its relationship with biases and (corresponding) debiasing processes. The following remarks are thus meant as theoretical expansions of Zenker's point.

The reasons behind the success of the type of practical exercise envisaged to overcome the power of biases remain underexplored in Zenker's contribution. I will try to show here that the success of the type of practical proposal defended in this paper constitutes evidence of the social function of reasoning (see Mercier & Sperber, 2011), and that it is because the TLA prompts for reasoning to be triggered within its natural context of occurrence (i.e., an inherently argumentative social context) that it has prospects of being successful. To be more specific, I think that if Zenker showed *that* the TLA works and *how* it works, more can be said on the reasons *why* it works. To this end, I will draw on Mercier & Sperber's recent argumentative theory of reasoning (see Mercier, 2009, Mercier & Sperber, 2009, 2011).

Mohammed, D., & Lewiński, M. (Eds.). *Virtues of Argumentation. Proceedings of the 10th International Conference of the Ontario Society for the Study of Argumentation (OSSA), 22-26 May 2013.* Windsor, ON: OSSA, pp. 1-7.

2. "REASONING AS A SOCIAL COMPETENCE" (MERCIER & SPERBER, 2011)

Mercier and Sperber postulate that reasoning has evolved not for individual purposes but for social and communicative ones. The central function of reasoning would thus not be to enhance cognition (for instance by allowing individuals to improve their knowledge or reach better decisions) but to convince others. Reasoning so envisaged is therefore intimately linked to social needs, and

> contributes to the effectiveness and reliability of communication by enabling communicators to argue for their claim and by enabling addressees to assess these arguments. It thus increases both in quantity and in epistemic quality the information humans are able to share. (Mercier & Sperber, 2011, p. 71-72)

Within a massively modular cognitive framework, they postulate the existence of a domain-specific argumentative module that evolved to produce and evaluate arguments. The module is assumed to operate inferentially by taking as input a claim and relevant information and by yielding reasons to accept or reject that claim in its output; it is furthermore assumed to operate both intuitively and reflectively, the latter corresponding to what we call reasoning proper.

According to the theory, individuals engage in reasoning when they are motivated to do so: when someone explicitly or implicitly rejects their standpoint, when they anticipate disagreement, when they evaluate others' public arguments or when they want to communicate their decisions (see Mercier, 2009, p. 102 ff.). In all these scenarios reasoning is triggered by social motivations: individuals basically feel some kind of pressure to provide evidence in support of their position. This, following the theory, is the natural context of occurrence of reasoning, namely a context of "resolution of a disagreement through discussion" (Mercier & Sperber, 2011, p. 65). Inherently argumentative contexts are thus those where reasoning is at its best.

Mercier & Sperber's extensive article (2011) convincingly and systematically shows that the wealth of empirical research in the psychology of reasoning can be accommodated within this theory. To give but two examples, puzzling results of the Wason Selection Task, where groups are shown to outperform individuals, can now be explained in terms of the concrete argumentative nature of the selection task in that particular version: subjects presenting their solution are bound to face criticism (unless everyone agrees with them) and hence need to come up with arguments to support it, which constitutes reasoning incentive. A second example of the theory's explanatory power is its novel construal of the confirmation bias, usually deemed to be one of the most important flaws in reasoning. This bias now becomes a *feature* of argument production: the theory predicts – and existing literature confirms it – that in contexts devoid of argumentative stakes, the bias will lead to poor outcomes (to the extent that the individual's arguments are neither critically evaluated nor challenged), but that in social contexts of problem resolution, it will help individuals to find (good) arguments to support their claims (Mercier & Sperber, 2011, p. 63-66).

3. REASONING AND DEBIASING: COUNTERING THE POLARIZATION EFFECT

One prediction of the argumentative theory is particularly relevant to assess Zenker's proposal, namely the prediction that individuals are better at reasoning (in terms of its outcome) when they are encouraged through interaction either to defend their position or to attack their opponent's (to potentially convince the latter of their own position). In short, the idea is that group discussions in which individuals find it relevant to produce and evaluate arguments provide appropriate conditions for effective reasoning to be deployed. But this is not yet equivalent to assuming that such scenarios are ideal for debiasing purposes: we first need to say more about the relationship between reasoning and debiasing.

One way of going about people's natural propensity for bias is to consider that they are "nearly incorrigible cognitive optimists", because "they take for granted that their spontaneous cognitive processes are highly reliable, and that the output of these processes does not need re-checking" (Sperber *et al*., 1995, p. 90). In the absence of critical challenge, people will expectedly stick to their mental states, even if these are normatively disadvantageous in some respect. The first step in making them aware of this disadvantage would thus be to challenge them to reflect on the reasons they have for holding these mental states: if they are unable to come up with solid justifications or if the justifications they produce are shown to be flawed, chances are that they will recognise the weakness of their position – provided they are not hopelessly dogmatic and that they allow for the possibility of reasonable criticism. Gaining awareness that the reasons for which they held their initial position are weak might in turn allow them to explain their 'error' as the result of bias. In other words, encouraging people to engage in reasoning may bring about a personal experience of being wrong, which is an indication that reasoning may be used for debiasing purposes. Additionally, and everything else being equal, representing and making available multiple reasons supporting a standpoint makes it *ipso facto* more vulnerable to criticism, since necessarily more information is available and exposed to challenge.

Zenker's TLA is designed to prevent the polarization effect by alerting participants onto their own biases. For this participants are instructed to first state their opinion on a debatable issue and to "call out reasons pro/con one or the other position" (p. 6), then to discuss, with a similarly opinioned group of students, the arguments supporting a different position in order to estimate their strengths and weaknesses and finally to publicly report "the order and structure over pro/con reasons" (p. 6). The task features three constraints of capital importance to trigger reasoning:

- The topic discussed in the TLA needs to be one students polarise over (step 2 of the sequential description, p. 6). Difference of opinion is a necessary condition for argumentative exchanges to take place, so in this configuration students will have motives to reason.
- The alternative position under discussion is said to entail that the students' own position is false (step 5). This parameter is of crucial

importance because it provides students a motivation to argumentatively engage with said position and to resort to their reasoning abilities. Put more simply, the TLA puts students in a situation with stakes: they are requested to take seriously a position challenging theirs as being false, and this will arguably motivate them to treat counterarguments critically.

- In step 6, the TLA requires students to dialectically relate arguments supporting their own position (pro-reasons) to corresponding counterarguments (con-reasons): this might lead students to relativize the weight of their own arguments by realising that the same aspect of the issue can be equally justified or attacked. Furthermore, the dialectical matrix of arguments they are instructed to elaborate will allow them not only to gauge arguments and counterarguments against each other, but also, more simply, to quantitatively compare the number of pro-reasons and con-reasons.

The three constraints identified above make sense within the argumentative theory of reasoning because they provide ideal conditions for reasoning to take place: (i) the TLA makes sure that a debatable issue is on the table, thereby allowing for the *possibility of argumentative exchanges*; (ii) by explicitly calling their views into question, it prompts participants to *look for arguments* in favour of their position but also against alternative positions; (iii) it requires participants to explicitly relate pro and con reasons, thereby allowing for the *possibility of relativizing opposing positions*. The TLA therefore has good prospects of being successful because it meets the conditions for engaging the argumentative module. Let us also add that if at the end of the TLA debiasing has occurred and has successfully prevented the polarization effect, this is not a direct consequence of reasoning. Rather, reasoning has provided the means to lay all relevant information on the table for debiasing to occur. Arguably, students will be able to get rid of their bias only after they have been made aware, through direct perception of a difference between initial and final assessment of their position, that they were biased in the first place.

It is interesting to note that both the research referred to by Zenker and his own take on the issue implicitly converge towards the predictions of the argumentative theory. Pronin and colleagues specifically point to the ineffectiveness of group discussions which are devoid of dialectical argumentative engagement when they suggest that

> [a]sking opposing partisans to sit down together, and inviting them to share their views and the reasons they hold them, might actually prove counterproductive, because such exchanges are apt to reinforce rather than weaken presumptions of extremity and intractability. (Pronin *et al*., 2002, p. 653)

If you do not instruct participants to seriously engage their opponents and simply ask them to share (i.e., to tell) their views, their reasoning abilities may lack incentive. When they conclude that "[i]t is unsurprising, perhaps, that participants

saw peers and adversaries as less extreme after those individuals had articulated and even acknowledged the other side's arguments" (ibid.), Pronin *et al*. are unwittingly echoing the idea that reasoning underpins the ability to articulate and acknowledge counterarguments. Similarly, when Zenker asserts that the TLA allows students to benefit from "a model for de-biasing that implicates charitable engagement with interlocutors' views, rather than a mere self-check of one's presumably good intentions" (p. 8), he is indirectly pointing to the usefulness of dialectifying one's position instead of confining students into introspection, as predicted by the theory. In arguing that "[t]hrough such engaging, one's own biases may be better discerned than without such engaging" (p. 8), Zenker also expresses the main tenet of the argumentative theory of reasoning, namely that argumentative contexts (denoted by "such engaging") trigger reasoning, which, as we have seen, can contribute to debiasing.

4. FURTHER REMARKS AND CONSTRAINTS ON THE TLA

A few remaining questions and remarks can be added, mainly regarding the material used in the task and the nature of the interaction (collaborative/competitive) between participants of the TLA.

One may wonder to what extent competitive attitudes might perturb the TLA. That is, would the TLA still be successful if students expected (and ended up adopting) competitive rather than cooperative attitudes? Zenker reports the findings of Kennedy & Pronin (2008) that "bias ascription correlates with opponent's mutual expectations that the disagreement will escalate" (p. 4), thereby allowing us to draw the conclusion that under these conditions the TLA might fail. Mercier (2009, pp. 136-137) interestingly treats such cases as consequences of the confirmation bias: subjects who, for whatever reason, do not find it in them to be willing to critically evaluate their own claim might limit themselves to looking for evidence that confirms it. This takes us to a second question.

In those cases, the TLA's failure could follow from the participants' general failure to reason properly. This is why the topic must be carefully selected: on the one hand it should be an issue students polarize over, but on the other it should not prevent them from "taking the other side seriously". To take a (perhaps too) extreme example, a topic such as alcohol consumption and driving might trigger highly emotional and one-sided responses in someone who has lost a relative in a drinking-related car accident; in turn, this could be too strong to allow her/him to even consider the possibility of entertaining, for the purpose of the exercise or simply for the sake of it, an alternative position. This would then deny the possibility of taking that extra step required to provisionally give some merit to an opponent's alternative position. In those cases, furthermore, we should expect the confirmation bias to be deployed in full effect: the participant would presumably not have it in her/him to seriously engage with alternative positions. This is thus something for the instructor to take into account in the selection of the topic.

A second line of inquiry concerns other potential debiasing techniques. While empirical evidence suggests that 'show and tell' methods are unsuccessful, one may wonder if something closer to argumentative self-control strategies, within some

sort of "ethics of argumentation" (see Correia, 2012, section 4), may be useful. I am not aware of any empirical studies on these issues, but training people to develop specific skills such as argumentative reconstruction, abstract thinking (e.g., logic and statistics), playing the devil's advocate, conscious elaboration of heuristics meant to impose limits onto what the subject allows herself to do argumentatively-wise or even the ability to be charitable, could be experimentally tested to assess their usefulness. This could represent additional directions of research for future studies of debiasing techniques.

Finally, while the success of Zenker's TLA seems to be interpretable in terms of its ability to trigger reasoning, the question of the generalizability of this explanation remains: would it be possible to design additional TLAs which do not depend on the virtues of reasoning? The polarization effect typically belongs to a class of biases arising within an interactive scenario where opposing views are discussed. But what about more 'solitary' biases? Are they all prone to debiasing when envisaged in a critical discussion? I am under the impression that the answer is positive, to the extent that the act of calling into question and that of justifying one's position are somehow independent from the contents under consideration (in principle any piece of information can be critically called into question). So if debiasing is systematically a result of someone making their reasons for believing, acting, etc. manifest, chances are that reasoning will always be a privileged path to debiasing.

5. CONCLUSION

I have tried to show that the main reason behind the success of Zenker's TLA is to be found in its argumentative nature: within such context, reasoning is bound to take place and debiasing likely to occur. The cognitive complexity of the task should also be noted: in the TLA, subjects need at the same time to be able to produce and evaluate arguments after having been motivated to do so and it could be argued that this TLA provides ideal conditions to engage the argumentative module in its most natural domain of action.

Beyond theoretical considerations, I also believe that the kind of research described here is a clear illustration of the merits of multidisciplinary convergence, as I have tried to show that social and cognitive psychology have much to share with argumentation theory the moment we address argumentative reality in the way Zenker does.

REFERENCES

Correia, V. (2012). The ethics of argumentation. *Informal Logic, 32*(2), 222-241.

Kennedy, K. A., & Pronin, E. (2008). When disagreement gets ugly: Perceptions of bias and the escalation of conflict. *Personality and Social Psychology Bulletin, 34*, 833–848.

Mercier, H. (2009). *La théorie argumentative du raisonnement*. PhD thesis, EHESS, Paris, 376p, ms.

Mercier, H., & Sperber, D. (2009). Intuitive and reflective inferences. In J. St. B. T. Evans and K. Frankish (Eds.), *In Two Minds: Dual Processes and Beyond* (pp. 149-170). Oxford: Oxford University Press.

Mercier, H. & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioural and Brain Sciences, 34*, 57–111.

Pronin, E., Puccio, C., & Ross, L. (2002). Understanding misunderstanding: Social psychological perspectives. In: T. Gilovich, D. Griffin, and D. Kahneman (eds). *Heuristic and Biases: The Psychology of Intuitive Judgement* (pp. 636–665). Cambridge: Cambridge University Press.

Sperber, D., Cara, F. & Girotto, V. (1995). Relevance Theory explains the Selection Task. *Cognition, 57*, 31-95.