Argumentation and Reasoned Action

Proceedings of the 1st European Conference on Argumentation, Lisbon 2015

Volume I

Edited by

Dima Mohammed

and

Marcin Lewiński

ISBN 978-1-84890-211-4

College Publications
Scientific Director: Dov Gabbay
Managing Director: Jane Spurr

http://www.collegepublications.co.uk

Original cover design by Orchid Creative www.orchidcreative.co.uk
Printed by Lightning Source, Milton Keynes, UK

35

Giving Reasons *Pro Et Contra* as a Debiasing Technique in Legal Decision Making

FRANK ZENKER
Department of Philosophy & Cognitive Science, Lund University, Sweden
frank.zenker@fil.lu.se

CHRISTIAN DAHLMAN
Law Faculty, Lund University, Sweden
christian.dahlman@jur.lu.se

RASMUS BÅÅTH
Department of Philosophy & Cognitive Science, Lund University, Sweden
rasmus.baath@lucs.lu.se

FARHAN SARWAR
Department of Psychology, Lund University, Sweden
farhan.sarwar@psy.lu.se

We report on the results of deploying the debiasing technique "giving reasons pro et contra" among professional judges at Swedish municipal courts (n=239). Experimental participants assessed the relevance of an eyewitness's previous conviction to his credibility in the present case. Results are compared to data from lay judges (n=372). The technique produced a small positive debiasing effect in the sample of Swedish judges, while the effect was negative among lay judges.

KEYWORDS: debiasing technique, heuristics and biases, legal decision-making, prior conviction, witness scenario

1. INTRODUCTION

How to improve decisions is a pertinent question whenever judgments are unavoidable. The decisions that judges and juries must reach

D. Mohammed & M. Lewiński (eds.) (2016). *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon, 2015. Vol. I, 807-820.* London: College Publications.

virtually every day provide a case in point, a fortiori when these bear strongly on the fates of individual and collective agents. Since biased reasoning and decision making is (rightly) thought to occur also in legal contexts (see, e.g., Langevoort, 1998 for a review; cf. Mitchell, 2002), no argument seems required that it ought to be reduced. Rather, empirical knowledge is wanted how reliable reductions may be achieved.

Professional judges tend to assume of themselves, firstly, that non-jurist decision makers regularly err in assessing the relevance of legal evidence; and, secondly, that judges reason in ways that reliably avoid such error. For some five decades, however, empirical research in the heuristics and biases tradition has supported the first assumption also for judges. Relevance-assessments may therefore be assumed to differ widely between intuitive and deliberative modes of reasoning and decision making, both between and within (groups of) agents.

Our research focuses on the second assumption, above. It addresses four related questions through controlled experimentation and interpretative analysis:

(1) What is the accuracy-difference between judges' and laypersons' assessments of the relevance of legal evidence (or: how much better are judges at activating 'system 2' in such assessments)?
(2) Do relevance-assessments improve across both groups subsequent to being instructed to deploy a debiasing technique?
(3) What is an optimal allocation between debiasing techniques and the bias(es) thus mitigated?
(4) How can debiasing techniques be improved?

This paper reports empirical results regarding the first two questions forthcoming from a pilot-study with a sample of professional Swedish judges and a sample of Swedish lay-judges (*nämdeman*). Experimental participants were asked to assess aspects of a mock legal-case that had been manipulated to contain bias-triggering information. In the experimental subgroup, the mock case was followed by explicit instructions "to give reasons pro/con"; in the control group it was not.

The purpose of this experimental set-up is to assess the positive, negative, or neutral effect(-size) of instructions to deploy a debiasing technique in a hypothetical legal decision making scenario vis-à-vis an established cognitive bias, on one hand, and a debiasing method, on the other—where the latter may count as far less established. The relevant bias is a "devil effect", insofar as an information-item about a person is of exaggerated importance in gauging her general credibility (see below). Such research contributes to assessing the average effectiveness of a debiasing technique, itself an instance of prescriptive ameliorative

intervention, if and insofar as a technique constitutes an efficient cause whose effect shows as an improved, or perfect, alignment between a normative standard and the decision outcome.

Following a brief introduction to biases and debiasing (Sect. 2), we present the method (3) and main results (4), offer a discussion (5), and close with brief conclusions (6).

2. BIASES AND DEBIASING

Biases are generally considered *latent*, that is, subjects tend to be unaware of them. By definition, a technique does debias when its deployment brings forth a decision that (i) differs markedly from one brought forth by deploying a heuristics, and (ii) also complies with a normative standard, e.g., as set forth by the law.

Broadly speaking, what authors such as Kahneman & Tversky (1982; 1996), or Kahneman (2011) call *biases*, philosophers and scholars of law associate with the *fallacies*. The latter fields share a tradition in Aristotelian scholarship, specifically the critiques of the Sophistic mode of audience persuasion. The 16th century Francis Bacon's delivering his *idolatry* or the 17th century John Locke naming of a range of fallacies fronted by "ad" (e.g., *ad hominem*) have continued this tradition into the modern age. Since Hamblin (1970), fallacies are standard fare in speech communication, rhetoric, and argumentation studies, among others. Around that time, moreover, the interpretation of fallacies as *reasoning errors* became separated from viewing fallacies as *problematic arguments* (e.g., van Eemeren & Grootendorst, 1984). Most psychologist and cognitive scientists, however, continue to strictly endorse the first interpretation.

Despite a vast number of empirical studies confirming the assumed operation of such biases for various groups of subjects, few studies pertain to contexts of legal decision making. Exceptions are, among others, Guthrie et al.'s (2007) study of anchoring, hindsight bias and base rate neglect, and English et al.'s (2006) study of the anchoring effect. Both particularly support that biases also influence legal decision making (for further references see Zenker et al., 2015).

Extant research moreover strongly suggests that humans are especially challenged in the application of debiasing methods, and more so in self-application (Pronin & Kugler, 2007; Pronin, Lin & Ross, 2002; Willingham, 2007; Kahneman, 2011; Kenyon, 2014). Self-assessment for biased thinking generally counts as a difficult cognitive ability to master; the primary challenge is the suspension of latency. But extant research (e.g., Guthrie et al., 2007; Irwin et al., 2010) also identifies debiasing techniques for legal decision making contexts, including the following.

Some of their underlying principles are already incorporated into procedural and substantial law. Debiasing effects thus brought should hence produce decisions that fall within the law.

- *Accountability*: legal decisions are subject to review by higher courts (Arkes, 1991).
- *Devil's Advocate*: Reminding subjects of the hypothetical possibility of the opposite standpoint (Lord et al., 1984; Mussweiler et al., 2000).
- *Giving Reasons* (Larrick, 2004, p. 323; Hodgkinson, 1999; Mumma & Wilson, 1995; Koriat et al., 1980).
- *Censorship*: When evidence counts as inadmissible, this may avoid biases triggered by such evidence.
- *Reducing Discretion*: Formulating legal norms that leave less room for a judge's interpretation (e.g., explicit checklists, or a pre-set damage amount).

An overview of extant research on debiasing in legal contexts including key methodological issues and additional references is provided in Zenker, Dahlman, and Sarwar (2015). As is argued there, successful debiasing techniques must simultaneously address aspects of cognition, motivation, and technology. They need to raise the agent's awareness of the bias (cognition) in ways that sustain or increase her impetus to avoid biased reasoning (motivation), while providing information that agents can in fact deploy to correct extant reasoning (technology).

Empirically testing a debiasing technique vis-à-vis a bias-triggering mock case serves to (i) empirically assess the extent to which a hypothetical (yet realistic) legal decision can be subject to biases, if and insofar as judges' and laypersons' hypothetical decisions "in the lab" are representative of those "outside the lab." Research further serves to (ii) estimate the potential of such instructions at mitigating biases, if and insofar as mitigation in the lab indicates that the same succeeds outside the lab. Finally, research eventually yields (iii) information on the optimal point at, and the optimal manner in, which decision makers would reasonably want to deploy a debiasing technique.

3. METHOD

By regular mail, *all* 667 professional judges at municipal courts in Sweden were asked to answer a pen-and-paper questionnaire that sought to assess whether, and if so to what extent, a previous conviction affects a witness's credibility. By way of a court's chief judge, moreover, 738 lay judges were asked to assess what one may generally call the "prior conviction relevance" in the following mock case.

> Sebastian P is charged for assault. According to the prosecutor's charge, Sebastian P assaulted Victor A, on July 20, 2012 at 23:30 outside a cinema in central Malmö, by repeated blows to the head. Sebastian P testifies that he acted in self-defense and denies the charges. One of the witnesses in the trial is Tony T, who was at the site on that particular evening. During the examination of the witness Tony T, it emerges that he had recently served a two-year prison sentence for illegal possession of weapons and arms trafficking.
>
> *Which of the following best describes your assessment? (Tick one option only)*
>
> - Tony T's previous conviction for illegal possession of weapons and arms trafficking affects the assessment of his credibility as a witness in the current trial. When various factors are weighed, the fact that he had previously been convicted of illegal possession of weapons and arms trafficking is *strongly* to his disadvantage.
>
> - Tony T's previous conviction for illegal possession of weapons and arms trafficking affects the assessment of his credibility as a witness in the current trial. When various factors are weighed, the fact that he had previously been convicted of illegal possession of weapons and arms trafficking is *clearly* to his disadvantage.
>
> - Tony T's previous conviction for illegal possession of weapons and arms trafficking affects the assessment of his credibility as a witness in the current trial. When various factors are weighed, the fact that he had previously been convicted of illegal possession of weapons and arms trafficking is *somewhat* to his disadvantage.
>
> - Tony T's previous conviction for illegal possession of weapons and arms trafficking *does not affect* the assessment of his credibility as a witness in the current trial.

In the experimental groups of both samples (professional and lay judges)—after the scenario, but before the central question and the four alternative answers were presented—participants were asked to state reasons why Tony T's convictions would affect his credibility as a witness in the present trial *and* to state reasons why his convictions would not affect his credibility in the present trial. No such instructions were included in the questionnaire given to control group-participants.

Of the professional judges, 40% returned the questionnaire (*n*=239), where 143 participants, i.e., 59.8% of the sample, had *not* received instruction to deploy any debiasing technique before answering the case (control group), while 96 participants, i.e., 40.2% of the sample, were instructed to state reasons for their assessment (experimental group; later referred to as "debias group"). Among lay judges, 52% returned the questionnaire (*n*=372), of which 171, i.e., 45.9%, belonged to the experimental group and 201, i.e., 54,1%, to the control group. In both samples, the response rate is unbalanced since participants were at liberty to return the questionnaire; they did not receive financial or other compensation for participating in this study.

Typical responses in both samples included the following *pro*/*con* reasons:

Prior conviction is relevant (pro)

- Tony T. has no barrier to breaking the law
- Tony T. may have an interest (e.g., revenge)
- Tony T. has reduced "citizenship-capital"
- Tony T. has a pro-attitude to violence

Not relevant (con)

- Unrelated event/circumstances
- No evidence that prior conviction matters
- Prior conviction *should* be irrelevant
- Current testimony occurs under oath

Prior to deploying the questionnaire, we did not formulate a point-hypothesis to code a normatively correct response. Rather, we assumed that obtaining differences between the experimental and the control group suggests that "giving reasons *pro et contra*" has a debiasing effect provided that participants in this group do on average display a lower assessment of the prior conviction relevance.

4. RESULTS

The effect of deploying the debiasing technique "giving reasons *pro et contra*" was *prima facie* miniscule. First looking at professional judges, fewer participants in the debias group than in the control group took the witness's previous conviction to be *clearly* or *strongly* to his disadvantage in the present case. Expressed in numbers, these were six and respectively one *vs*. zero participants (4.2% and 0.7% of the sample). This can provide at best *some* reason to believe that the

debiasing technique had an ameliorating effect on judges. Moreover, 28 judges in the control group (19.6% of the judges in the control group) register as finding the witness's prior conviction to be *somewhat* negatively relevant. Finally, 20 judges in the experimental group (12.8% of the judges in the control group) so register *despite* a debiasing technique being deployed.

Turning now to lay judges, by contrast, hardly any noteworthy differences arose between the control and the experimental group: 7% and 8% of the total number of lay judges found the prior conviction to be *clearly* or, respectively, *strongly relevant*; 30% in each group found the conviction to be *somewhat relevant*; 61% and 63%, respectively, found the prior conviction to be *not relevant*. Table 1 and Fig. 1 give the full results of the questionnaire.

Responses were coded on a four point ordinal scale (as *not relevant*, *somewhat*, *clearly* and *strongly to the witness's disadvantage*; see Table 1). In order to investigate the difference between the four groups, that is, the control and experimental groups, each consisting of either professional or lay judges data was then subjected to an ordered probit analysis.[1]

		not relevant	**somewhat relevant**	**clearly relevant**	**strongly relevant**	***N***
Judges	Control	108 (76%)	28 (20%)	6 (4%)	1 (1%)	141
	Debias	76 (79%)	20 (21%)	0 (0%)	0 (0%)	96
	Total	*184 (77%)*	*48 (20%)*	*6 (3 %)*	*1 (0 %)*	*239*
Lay Judges	Control	126 (63%)	60 (30%)	12 (6%)	3 (1%)	201
	Debias	105 (61%)	52 (30%)	11 (6%)	2 (2%)	171
	Total	*231 (62%)*	*112 (30%)*	*23 (6%)*	*5 (2%)*	*372*

Table 1. Responses from Swedish judges and lay judges (*n*=number of subjects)

[1] See Daykin and Moffat (2002) for paradigmatic applications of ordered probit analysis and its advantages over far-better known, but also less well-suited, linear regression analyses. For instance, ordered probit analysis is not open to the objection that the distances between any two ordinal data points are implicitly treated as being equal. The probit analysis was done in the R statistical environment using the polr function in the MASS package (Venables & Ripley, 2002).

This analysis assumes that underlying the ordinal scale, on which participants' responses are measured, is a continuous random variable representing participants' assessment of prior-conviction relevance (PCR). The value of this latent variable has no direct interpretation but is a relative measure of PCR, where a higher value implies that a prior conviction is deemed more relevant. Crucial for the following statistical analysis, the expected value of the latent variable can be taken as a measure of the general sentiment of the group and can thus be used in comparing the groups.

The distributional parameters of the latent variable were gauged through maximum likelihood estimation, yielding the parameter estimates under which the ordered probit model is most likely to generate the observed data in Table 1.

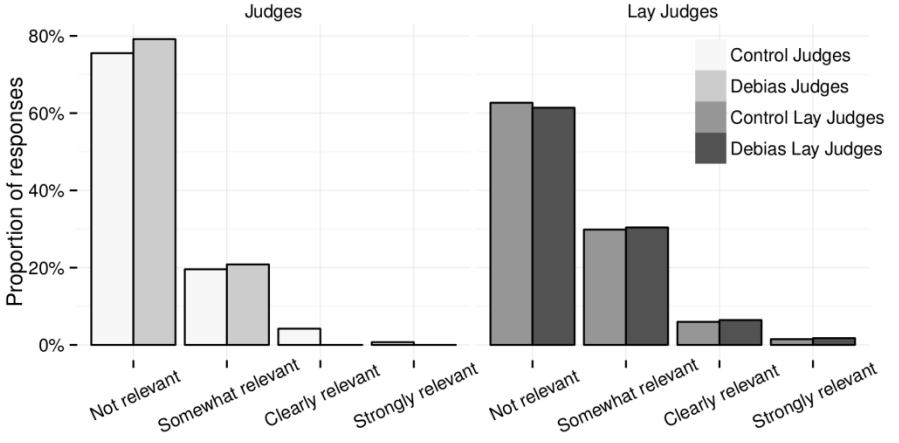

Fig. 1 Proportion of responses from Swedish judges and lay judges

In virtue of being maximally consistent with the original data, the hypothetical model may be interpreted as the *most probable continuous distribution of the latent PCR-variable* among respondents. In this sense, the hypothetical model can be viewed to have probably generated the original data. The shaded curves in Figure 2 show the maximum likelihood estimates of the latent PCR-variable among judges and lay judges in the control and the experimental group. The figure is divided into four regions corresponding to the four possible responses in the survey. The percentage of the area under the curve within each region corresponds to the model's estimate of the probability that a member of these groups produces the corresponding survey response. The dashed vertical lines mark the expected values of the latent PCR-variables, here taken as a measure of the general sentiment of groups.

Comparing panels A-B and C-D in Fig. 2, the displacement of the expected values of the PCR-variables indicates the impact of the debiasing intervention. While there is a visible difference in the general assessment of PCR between judges in the experimental and judges in the control group, there is hardly any difference between the lay judges in the experimental group and lay judges in the control group. But there was nevertheless a substantial overall difference between judges and lay judges: the former judged the prior conviction to be less relevant than the latter.

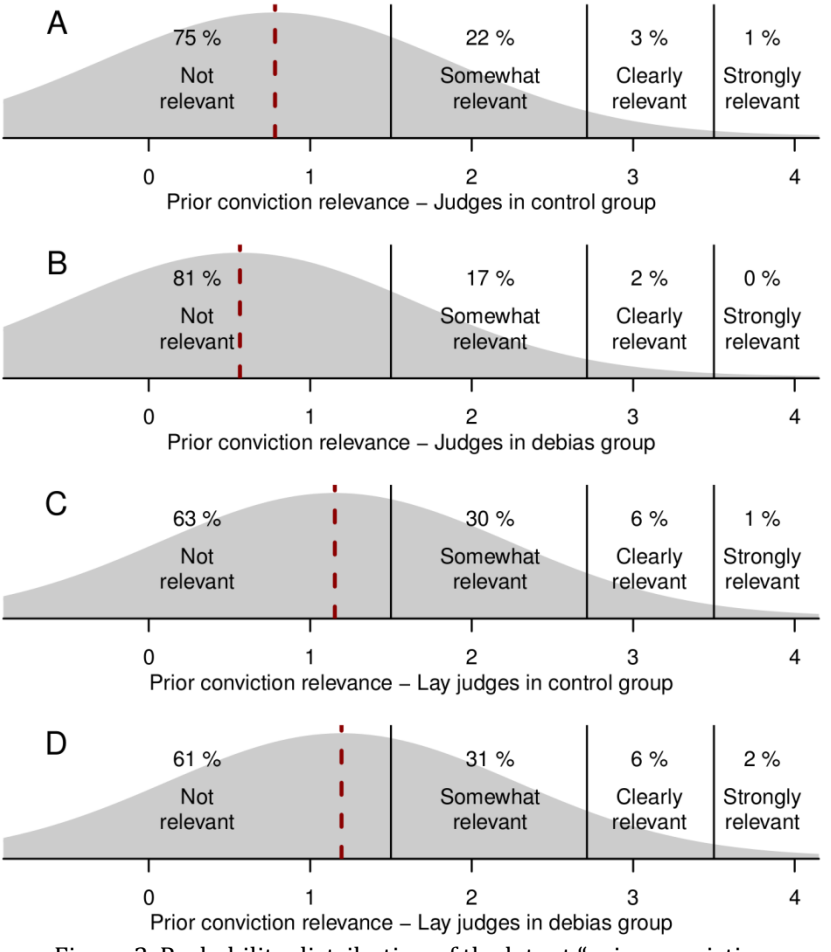

Figure 2. Probability distribution of the latent "prior conviction relevance"-variable for judges and lay judges in the debias and the control groups.

A Bayesian analysis was performed to gauge the uncertainty in the estimates from the ordered probit analysis, and to quantify whether the joint data from judges and lay judges in the experimental and the control group support, or undermine, the hypothesis that the debiasing technique "giving reasons pro/con" had an ameliorating effect.[2]

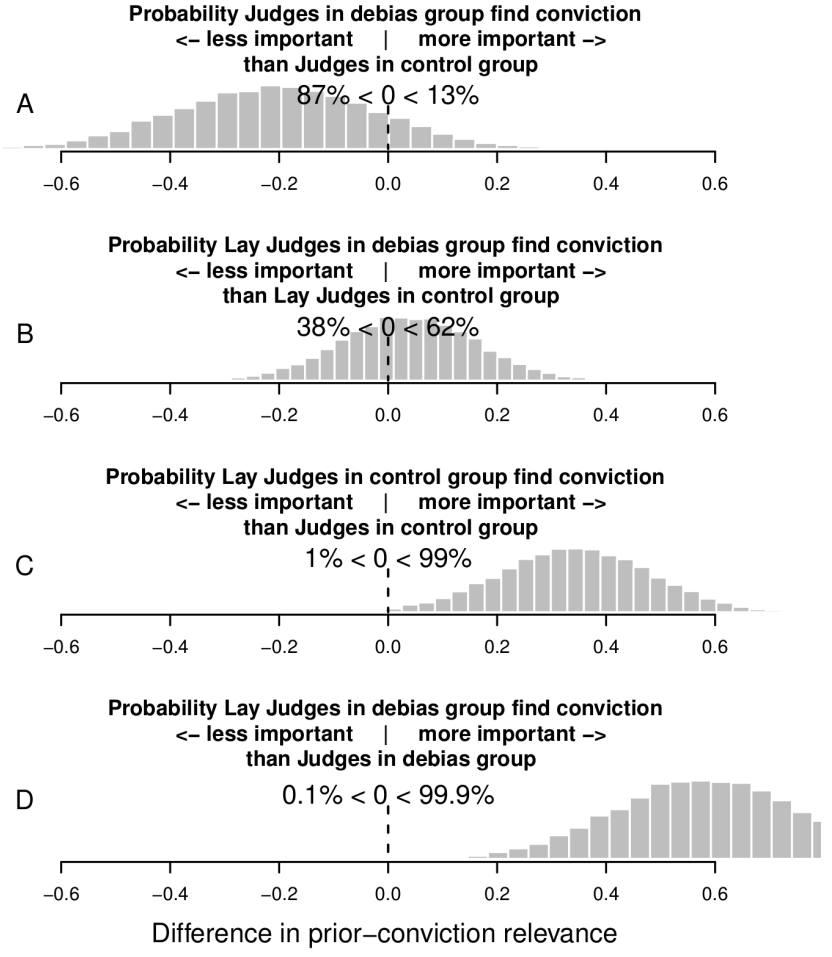

Fig. 3. Distribution of probabilities given model and evidence from professional and lay judges

[2] The analysis was performed in the R statistical environment using the MCMCoprobit function in the MCMCpack package (Andrew et al., 2011). The default priors of the MCMCoprobit function was used, which were non-informative uniform priors over all parameters.

Figure 3 shows the probable difference in the expected values of the PCR-variable (marked by a dashed line; Fig. 2) between all four groups. Given model and data, there is a 87% probability that judges in the experimental find the prior conviction less relevant (Fig. 3, panel A), compared to a 38% probability that lay judges in the experimental group find the prior conviction less relevant (Fig. 3, panel B).

This may be interpreted as *rather weak positive* evidence that deploying the relevant debiasing technique has a debiasing effect among judges, but not among lay judges. Moreover, comparing judges and lay judges in the control group (Fig. 3, panel C) and the debias group (Fig. 3, panel D) shows a probability larger than 99%—which may be interpreted as *very strong* evidence—that in both the control and in the experimental condition lay-judges assign a higher prior conviction relevance than judges, with evidence from the experimental condition registering slightly stronger yet. This, in fact, amounts to having observed an interaction of the professional status with the assessment of prior conviction relevance.

4. DISCUSSION

In the experimental data, strong evidence for a mitigating effect of the debiasing method "stating reasons pro/con" onto participants' responses has *not* been forthcoming. Rather, the study found an 87.1% probability for a mitigating effect. This can at best count as weak evidence. In the mock case, lay judges did overall assign a *greater* weight to the previous conviction of the witness than professional judges. Moreover—and perhaps disturbingly—compared to the relevant control group lay judges in the experimental group displayed an *increased* mean score.

Results are broadly negative in the sense that the "Tony T" mock case failed to trigger a *strong* bias among professional or lay judges. By and large, professional judges merely assigned *some* weight to the previous conviction, while lay judges assigned a greater weight. The debiasing technique "stating reasons *pro et contra*" in other words failed to meet with a strongly biased sample of judges and lay judges. The technique nevertheless appears to succeed in "taking the edge off," as it were. After all, compared to the relevant control group, the number of extreme judgements in the experimental group of professional judges is reduced. It stands to reason, of course, that "removing" but one extreme judgement through a debiasing intervention does already constitute an important and desirable outcome. This nonetheless remains a very small effect. And as the debiasing technique met with a comparatively more biased sample of lay judges, its deployment not only failed to

mitigate the bias; rather, it slightly worsened the judgement compared to the control group of lay judges. But also this result remains statistically insignificant, and so cannot easily be accounted for as an effect of deploying the technique.

To address the objection that additional data should have been collected in order to assess whether a statistically significant debiasing-effect would after all have been observed, consider that the sample of Swedish judges in the present study (n=239) represents no less than 40% of the relevant population(!). To increase this number would no doubt present greater practical difficulties. It remains correct, of course, that small experimental effects must always be confronted with large data-samples. But for the small effect here reported to *potentially* register as statistically significant does necessarily require a sample-size that exceeds the size of the relevant population! This fact hence entails that there might be biases whose presence, and debiasing techniques whose effect, can principally *not* be demonstrated by obtaining strong evidence for a difference between the control and the experimental group whenever the effect is too small to register as significant even against the size of the relevant population. For this reason, the "need more data"-objection is particularly weak in the present context.

Demonstrating the effectiveness of a debiasing technique at conventionally accepted levels of significance could instead be served by maximizing the difference between participants' ratings in the control and the experimental group. In the present study, as we saw, both groups displayed rather low degrees of biasedness. It therefore remains a challenge for future research to create experimental set-ups that induce stronger biases. In view of the idea that already a small ameliorating effect, if it is real, should be viewed as a desirable outcome of deploying a debiasing technique, we suggest that it can be reasonable to accept weaker forms of evidential support, rather than inferring that the debiasing technique was probably ineffective. Since this stance is unlikely to meet with wide acceptance, however, the key-task would remain to induce a stronger bias.

5. CONCLUSION

Among Swedish judges at municipal courts, the "Tony T" mock case failed to meet with "sufficiently biased" respondents, since few assigned a great(er) weight to the witness's prior conviction regarding his credibility in the present case. The debiasing technique "giving reasons *pro et contra*" could thus at best produce a small effect—too small to count as strong evidence relative to the sample or even the relevant population. Rather than inferring that the technique probably had no

effect, however, we submit these results as weak positive evidence in favor of the effectiveness of this debiasing technique.

As we also saw, results differed—yet in the normatively "wrong" direction—when the same technique was deployed vis-à-vis the same mock case among lay judges, who seem to have constituted a comparatively more biased sample than the professional judges. The debiasing technique had a weak *adverse* effect on lay judges; subsequent to deploying it, the latter assigned a slightly increased weight to the relevance of previous conviction. As we have stressed, however, this interpretation is subject to caveats as the effect remained too small.

Among all measures taken, we obtained very strong evidence merely for a relation between the profession and the level of biasedness, there being a probability greater than 99% that lay judges were more biased than professional judges. To test the effectiveness of debiasing methods against standard statistical assumptions, future studies seeking to produce strong(er) positive evidence are challenged to find ways of triggering strong(er) biases.

ACKNOWLEDGEMENTS: We thank audience members at the First European Conference on Argumentation, 9-12 June 2015, Lisbon, Portugal, for discussion and Fabrizio Macagno for his commentary. Research was funded by the Ragnar Söderberg Foundation. Rasmus Bååth acknowledges funding through Swedish Research Council grant number 349-2007-8695.

REFERENCES

Arkes, H.R. (1991). Costs and benefits of judgement errors: Implications for debiasing. *Psychological Bulletin*, *110*, 486–498.

Daykin, A.R., & Moffat, P.G. (2002). Analyzing ordered responses: a review of the ordered probit model. *Understanding Statistics*, *1*(3), 157–166.

Eemeren, F. H. van, & Grootendorst, R. (1984). *Speech acts in argumentative discussions: A theoretical model for the analysis of discussions directed towards solving conflicts of opinion*. Amsterdam: Walter de Gruyter.

English, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, *32*(2), 188–200.

Guthrie, C., Rachlinski, J. J., & Wistrich, A. J. (2007). Blinking on the bench: How judges decide cases. *Cornell Law Review*, *1*, 1–44.

Hamblin, C. (1970). *Fallacies*. London: Methuen.

Irwin, J., & Daniel, L.R. (2010). Unconscious influences on judicial decision-making. *McGeorge Law Review*, *43*, 1–20.

Kahneman, D., & Tversky, A. (1982). On the study of cognitive illusions. *Cognition*, *11*, 1123–1141.

Kahneman, D. & Tversky, A. (1996). On the reality of cognitive illusions: A reply to Gigerenzer's critique. *Psychological Review*, *103*, 582–591.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Strauss and Giroux.

Kenyon, T. (2014). False polarization: Debiasing as applied social epistemology. *Synthese*, *191*(11), 2529–2547.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory*, *6*(2), 107–118.

Langevoort, D. C. (1998). Behavioral theories of judgment and decision making in legal scholarship: A literature review. *Vanderbilt Law Review*, *51*, 1499–1540.

Lord, C.G., Lepper, M.R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Judgment*, *47*(6), 1231–1243.

Martin, Andrew D., Quinn, Kevin M., & Park, Jong Hee (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*, *42*(9), 1-21.

Mitchell, G. (2002). Why law and economics' perfect rationality should not be traded for behavioral law and economics' equal incompetence. *Georgetown Law Journal*, *91*, 67–167.

Mumma, G.H., & Wilson, S.B. (1995). Procedural debiasing of primacy/ anchoring effects in clinical-like judgments. *Journal of Clinical Psychology*, *51*(6), 841–853.

Mussweiler, T., Strack, F., & Pfeifer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin*, *26*(9), 1142–1150.

Pronin, E., Lin, D., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*, 369–381.

Pronin, E., & Kugler, M. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, *434*, 565–578.

Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition. Springer: New York.

Willingham, D. (2007). Critical thinking: Why is it so hard to teach? *American Educator*, *31*(2), 8–19. (reprinted as: Willingham, D.T. (2008). Critical thinking: Why is it so hard to teach? *Arts Education Policy Review*, *109*(4), 21–32.)

Zenker, F., Dahlman, C. (2015). Reliable debiasing techniques in legal contexts? Weak signals from a darker corner of the social science universe. In F. Paglieri (Ed.). *The psychology of argument: Cognitive approaches to argumentation and persuasion* (pp. xx-yy). London: College Publications (*forthcoming*).

Commentary on Zenker, Dahlman, Bååth and Sarwar's Giving Reasons *Pro Et Contra* as a Debiasing Technique in Legal Decision Making

FABRIZIO MACAGNO
ArgLab, Universidade Nova de Lisboa, Portugal
fabrizio.macagno@fcsh.unl.pt

1. INTRODUCTION

The paper presents an insightful and groundbreaking approach to legal reasoning and argumentation. The fundamental assumption of this work is that biases, or rather latent fallacious reasoning (Zenker, Dahlman, Bååth & Sarwar, 2016, p. 811-812) affect legal reasoning as well, and can result in unwarranted conclusions to be reached. Such biases can be un-triggered by specific techniques, and in this paper the authors assess the effect of a strategy of debiasing character attacks in witness testimony. To this purpose, the authors run a mock test in which in a hypothetical scenario in which a witness is shown to have been convicted for previous crimes. The decision-makers (judges and perspective jurors) are divided in two groups (the control and the experimental group), in which the experimental group is subjected to a debiasing technique (to give reasons for their decision). The authors use quantitative methods to assess the effectiveness of this strategy, but as they report, the results are statistically weak, even though relevant for the purpose of discussing the relationship between the technique and its effects. Despite the efforts of the authors to undermine the relevance of the paper for legal argumentation, this work sheds light on relevant theoretical and practical issues that need to be taken into account, and introduces an extremely interesting method of investigation.

2. THE BIASED REASONING: CHARACTER ASSASSINATION

The authors took into account a specific type of biased reasoning, the commonly called "character assassination" in law that can be "devastatingly effective" (Cantrell, 2003, p. 534; Solomon, 2003, pp. 7–8). By showing that a witness (or a defendant) committed a previous crime, the decision-maker (the judge or the jury) is led to conclude that

D. Mohammed & M. Lewiński (eds.) (2016). *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon, 2015. Vol. I, 821-827.* London: College Publications.

his testimony is less reliable (or that he committed also the crime he is accused of). This type of attack is commonly analyzed in argumentation theory as the *ad hominem* argument (Walton, 1998, pp. 198–199, p. 217; 2002, p. 51). *Ad hominem* arguments consist in showing that the interlocutor's argument should not be accepted based on a negative judgment on different aspects of his or her character, such as logical reasoning, perception, veracity, or cognitive skills (Macagno, 2013). Clearly, the reasonableness of type of argument depends on the type of argument it is aimed at undermining. *Ad hominem* attacks are often reasonable when they undermine arguments based on the expertise or the position to know of a source, namely authoritative arguments. In these cases, if the truth, or rather acceptability, of the testimony depends on some of the character features attacked, the argument can be reasonable. Otherwise, would be simply irrelevant to the conclusion.

Despite their unreasonableness, often irrelevant *ad hominem* arguments have great impact on the evaluator of an argument, and more specifically in law on the judge or especially the jury. Attacking character does not simply amount to showing a flaw in an argument. It means showing that a person's character is somehow negative, which can lead to negative emotions. Emotions such as indignation, fear, contempt, or hate divert the interlocutor's attention from the rational and systematic assessment of the attack, leading to a conclusion based on fast associations between the emotion and a possible immediate reaction (Blanchette & Richards, 2004; Blanchette, 2006; Macagno, 2014).

Attacking a witness's character is allowed by the rules of evidence at common law. According to rule 609 of the Federal Rules of Evidence, it is possible to introduce evidence of the witness's past convictions in order to impeach his character for truthfulness: "One way of discrediting the witness is to introduce evidence of a prior criminal conviction of the witness, which affords the jury a basis to infer that the witness's character is such that he would be less likely than the average trustworthy citizen to be truthful in his testimony" (*State v. Nash*, 475 So. 2d 752, at 754, 1985). In this sense, evidence of prior misconduct can be a rational ground for assessing the trustworthiness of a witness, one of the possible dimensions that need to be taken into account when judging his testimony. However, this evidence often risks becoming a trigger of a fallacious conclusion reached by means of "fast" reasoning (Kahneman, Slovic, & Tversky, 1982; Tversky & Kahneman, 1974). As pointed out by McCormick (McCormick, 1972, p. 104) "a slashing cross-examination may carry strong accusations of misconduct and bad character, which the witness's denial will not remove from the jury's mind." One of the clearest and most famous examples of the force of this

type of character attack is *People v. Simpson* (No. BA 097211, 1995), in which the previous misconduct of the witnessing detective (Fuhrman), combined with proof of racial behavior, led the jury to believe that he was not reliable. This character assassination led to the acquittal of O.J. Simpson.

3. DEBIASING CHARACTER ASSASSINATION

Biased reasoning in law can occur both when the judgment is made by a professional judge, and when it is rendered by a jury of laypeople. In this latter case, in particular, the possibility of the jurors being unaware of the possible fallacious or weak reasoning becomes even higher, as they are not trained to make legal decisions and evaluate objectively the various factors of the case. To this purpose, the authors have first analyzed the so-called debiasing techniques, namely strategy used to un-trigger the latent mechanisms leading to a fallacious conclusion. Such strategies have different foci, depending on the dimension of the automatic reasoning that they intend to address. They can be aimed at turning a latent (implicit) mechanism into an explicit one, or leading the decision-maker to a careful assessment of the force of his conclusion, or simply preventing the decision-making from making some inferences. We can classify the techniques in Table 1:

Making the reasoning explicit	**Assessing the conclusion carefully**	**Preventing inferences**
Giving reasons	Accountability	Devil's advocate
	Censorship	Reducing discretion

Table 1: Debiasing techniques

These techniques clearly are some of the possible ones that can be used to reduce the possibility that the decision-maker comes to a conclusion based on problematic implicit arguments. Other possible techniques used in law are confronting the interlocutor with a biased but contrary conclusion or argument, so that the reasoning process becomes subject to careful assessment (Macagno & Walton, 2012).

The authors chose to test the debiasing technique of giving reasons, and they proved that its effects, even though not statistically significant, however indicate that there is a high probability that the difference between the control group and the one subjected to the

debiasing intervention is due to the intervention itself (5.6 times more probable than not). However, as the authors point out, this outcome does not meet the statistical requirements for significance. Moreover, when they took into account only the lay judges, they noticed that the debiasing technique worsened the judgment compared to the control group.

4. POSSIBLE PROBLEMS

One of the possible criticisms on the experiment can be addressed to the very mock case that the judges had to assess. Perhaps one of the causes of a lower variability is due to external factors and variables that the authors could not control, due to the vagueness of the case. The case reads as follows:

> Sebastian P is charged for assault. According to the prosecutor's charge, Sebastian P assaulted Victor A, on July 20, 2012 at 23:30 outside a cinema in central Malmö, by repeated blows to the head. Sebastian P testifies that he acted in self-defense and denies the charges. **One of the witnesses in the trial** is Tony T, who was **at the site** on that particular evening. During the examination of the witness Tony T, it emerges that he had recently served **a two-year prison sentence for illegal possession of weapons and arms trafficking**.

The judges had to assess whether the previous conviction of Tony T affects the credibility of his testimony (in a strong, clear, some, or no way). However, the case is too generic and leaves too much room to narratives and possible reconstructions of background information. Considering that a very low percentage of judges indicated a strong or clear effect on the credibility of the witness, a low one cannot be excluded if the judge is allowed to reconstruct information that the case does not specify. Did the witness know the defendant? Was the witness involved in other criminal activities after his conviction? Was the witness somehow related to with the defendant or the victim? All such factors cannot be excluded, and are likely to be reconstructed or taken into account as possible reasons of a biased testimony. At common law, a real case usually involves a cross-examination of the key witnesses, especially when their credibility can be undermined by character issues. While professional judges are trained to evaluate the various factors of the case without making additional hypotheses, this could be not the case for laypeople who simply relate the story with the most accessible narratives (the witness may know the defendant, since they are

allegedly both violent, and the witness may want to cover for his friend). If the debiasing technique consists in giving reason, some factors that can be used in such reasons need to be taken into account and controlled.

A second possible problem is the language of the variables. The authors indicate a four-point scale, but they fail to define clearly what "somehow" means when referred to "affecting the witness's credibility." As pointed out by common law cases, evidence of prior convictions is allowed because it is an element that the jury may want to take into account when assessing a witness's trustworthiness compared to an average citizen. Clearly, when no other elements are present, this piece of evidence can be irrelevant for evaluating character. However, when combined with an extremely succinct narration of the circumstances, it is not unreasonable to think that an untrained judge may find possible additional reasons making the testimony somehow less reliable. Perhaps the authors could refine their tests introducing more variability. They could formulate clearer hypotheses, less subject to personal interpretations, including or excluding some circumstantial factor, and then use a Likert scale with more or more definite levels (strongly disagree… strongly agree) to assess them. For example, the test could read as follows:

> •"Tony T's previous conviction for illegal possession of weapons and arms trafficking strongly affects the assessment of his credibility as a witness in the current trial." (strongly agree … strongly disagree).
> •…
> •"Considering that Tony T had never met the defendant or the victim before, Tony T's previous conviction for illegal possession of weapons and arms trafficking strongly affects the assessment of his credibility as a witness in the current trial." (strongly agree … strongly disagree).

In this fashion, they could measure how much the interpretation of the event can affect judgment and more importantly the reasons underlying it.

5. CONCLUSION

The authors have focused their self-criticisms on the scarce significance of the overall results. However, this study shows clearly how lay judges and professional ones differ concerning the assessment of a case. This difference becomes even more relevant when we consider the fact that

the debiasing technique has opposite effects on laypeople, who provided even worse results when they had to give reasons. This effect should be analyzed in depth, and related to the problem of prejudices and background knowledge. How do prejudices affect the reconstruction of a state of affairs? Perhaps it would be interesting to investigate how a layperson can reconstruct the narrative underlying the whole case, including the relationship between the witness and the defendant, and compare them with the reconstruction of professional judges.

To conclude, the authors perhaps failed to make a statistical point, but opened a very broad range of fundamental questions that should be addressed with the method that they used and that is revolutionary in the field of legal argumentation.

ACKNOWLEDGEMENTS: I would like to thank the Fundação para a Ciência e a Tecnologia for the research grant no. IF/00945/2013.

REFERENCES

Blanchette, I. (2006). The effect of emotion on interpretation of and logic in a conditional reasoning task. *Memory and Cognition*, *34*, 1112–1125.

Blanchette, I., & Richards, A. (2004). Reasoning about emotional and neutral materials - Is logic affected by emotion? *Psychological Science*, *15*(11), 745–752. http://doi.org/10.1111/j.0956-7976.2004.00751.x

Cantrell, C. (2003). Prosecutorial Misconduct: Recognizing Errors In Closing Argument. *American Journal of Trial Advocacy*, *26*, 535–562.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under Uncertainty: Heuristics & Biases*. Cambridge: Cambridge University Press.

Macagno, F. (2013). Strategies of Character Attack. *Argumentation*, *27*(4), 369–401. http://doi.org/10.1007/s10503-013-9291-1

Macagno, F. (2014). Manipulating Emotions. Value-Based Reasoning And Emotive Language. *Argumentation & Advocacy*, *51*(2), 103–122.

Macagno, F., & Walton, D. (2012). Character Attacks as Complex Strategies of Legal Argumentation. *International Journal of Law, Language & Discourse*, *2*(3), 1–58.

McCormick, C. (1972). *McCormick's handbook of the law of evidence*. St. Paul, Minnesota: West Publishing Company.

Solomon, R. (2003). *Not Passion's Slave*. New York: Oxford University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science (New York, N.Y.)*, *185*(4157), 1124–1131. http://doi.org/10.1126/science.185.4157.1124

Walton, D. (1998). *Ad Hominem Arguments*. Tuscaloosa: University of Alabama Press.

Walton, D. (2002). *Legal argumentation and Evidence*. University Park: The Pennsylvania State University Press.

Zenker, F. Dahlman, C. Bååth, R. & Sarwar, F. 2016. Giving *Reasons Pro et Contra* as a Debiasing Technique in Legal Decision Making. In D. Mohammed & M. Lewiński (eds.), *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon, 2015.* Vol. I, 809-822. London: College Publications.