Critical Thinking Education and Debiasing

TIM KENYON

University of Waterloo
Department of Philosophy
200 University Avenue West
Waterloo, ON
N2L 3G1
Canada
tkenyon@uwaterloo.ca

GUILLAUME BEAULAC

Yale University
Department of Philosophy
344 College St.
New Haven, CT. 06511-6629
USA
guillaume.beaulac@yale.edu

Abstract: There are empirical grounds to doubt the effectiveness of a common and intuitive approach to teaching debiasing strategies in critical thinking courses. We summarize some of the grounds before suggesting a broader taxonomy of debiasing strategies. This four-level taxonomy enables a useful diagnosis of biasing factors and situations, and illuminates more strategies for more effective bias mitigation located in the shaping of situational factors and reasoning infrastructure—sometimes called "nudges" in the literature. The question, we contend, then becomes how best to teach the construction and use of such infrastructures.

Résumé: Des données empiriques nous permettent de douter de l'efficacité d'une approche commune et intuitive pour enseigner des stratégies de correction de biais cognitifs dans les cours de pensée critique. Nous résumons certains de ces résultats empiriques avant de suggérer une taxonomie plus étendue de ces stratégies de correction de biais. Cette taxonomie à quatre niveaux permet un diagnostic utile de facteurs causant les biais et elle met en évidence davantage de stratégies permettant la correction plus efficace de biais, stratégies situées dans des mesures modifiant les infrastructures et les environnements cognitifs ("nudge" dans la littérature). Nous soutenons que la question porte dès lors sur les meilleures façons d'enseigner la construction et l'utilisation de ces infrastructures.

Keywords: Critical thinking, biases, debiasing, education, nudges

1. Introduction

Teaching critical thinking is an undertaking that permits emphasis on many different combinations of elements, the most traditional of which are formal logic, informal logic, argumentation, fallacy theory, and rhetoric. Increasingly, too, critical thinking courses and texts include an explicit emphasis on the psychology of cognitive and social biases (see, for example, Kenyon 2008; Ruggiero 2004; Groarke & Tindale 2004; and Gilovich 1991). While they vary greatly in the length and detail of their treatments, a common feature of these texts is that they present names, taxonomies and definitions of some key biases, perhaps with some examples or explanations of the underlying empirical work included. Given their role in critical thinking didactics, it is safe to assume that these treatments are intended to foster practical reasoning skills of mitigating or forestalling the effects of biases – to enable students to identify biases in reasoning, and to minimize biases in their own thinking.

The overall aim is consonant with the general rationale for teaching critical reasoning courses in the first place. Yet these texts also commonly lack empirically-informed material, *distinct* from that already mentioned, that aims to teach students the skills of minimizing bias in their thinking or their actions. In other words, the combination of what such treatments do and do not contain reflects the assumption that simply teaching students about biases is an effective way of enabling them to reduce the distortions of biases in their own thinking. We identify this assumption as the *intuitive approach to teaching debiasing*, or IA.

> (IA) Teaching facts about biases, including a taxonomy of biases and their various propensities to distort reasoning, is a reasonably effective means of providing students in critical reasoning courses with skills enabling the detection and mitigation of biases, including students' own biases.

Something along the lines of IA informs the treatments that biases receive in the critical thinking texts already noted. It is also central to some reviews on the topic (e.g., Larrick 2004), while its influence can be seen also in training contexts beyond that of a critical thinking course. An example of this latter type of context is Croskerry, Singhal, & Mamede's (2013a, 2013b)

approach to cognitive debiasing for clinicians' medical judgments: even though Croskerry et al. record a "general pessimism […] about the feasibility of cognitive debiasing" (Croskerry et al. 2013a, p. ii63), they adopt the recommendation that clinicians "must be informed and recognise the need for constant vigilance and surveillance of their [own] thinking to mitigate diagnostic and other clinical errors" (Croskerry et al. 2013b, p. 6).[1]

It is noteworthy that IA characterizes much of how critical thinking education treats debiasing, we contend, because when one considers the empirical evidence bearing on it, the most plausible simple view of IA is that it is—at least in most cases—false.[2] At least, the practice of simply teaching students facts about biases is not as effective as one might hope. The literature on the cognitive and social psychology of debiasing indicates, on balance, that teaching people about biases does not reliably debias them. Indeed, the literature suggests that (for at least a wide class of biases) practically *any* debiasing strategy intended to be learned and subsequently self-deployed by individuals, acting alone and at the point of making a judgment, is unlikely to succeed in significantly minimizing biases.

In the following remarks, we briefly outline why this is so before moving on to consider the ramifications for critical thinking education. Vast resources are currently devoted to teaching critical reasoning worldwide. Does the implausibility of IA mean that these resources are misused, to the extent that they are predicated on IA? Should philosophers, psychologists, and other critical reasoning educators just stop including a focus on biases in critical thinking education?

We do not think so. Rather, we take the lesson to be that whole societies and polities have a major interest in promoting efficacious debiasing education—extending to population-level demographic scales and intergenerational time scales. The difficulty of teaching debiasing skills that could be deployed in a strictly atomistic or individualistic way counts in favor of teaching and investing also in more collective debiasing strategies and infrastructure that would serve the latter sorts of interests. This approach will encompass teaching not just individual skills and knowledge, but skills that enable the

[1] In fact the distinctions between approaches to critical thinking that we propose in the following remarks should be helpful in characterizing the kinds of clinical training strategies described by Croskerry et al. (2013a, 2013b).

[2] Below, we identify some methods that would fall under IA that we believe to be relatively promising.

construction of reasoning infrastructure, and effective participation in social and organizational reasoning processes and decision procedures.

What would these processes, strategies and infrastructure look like? A key first step here is to reflect on the breadth of what can count as debiasing from a critical thinking perspective. Our aim in this reflection is to help motivate and set the stage for creative and empirically guided work on how to teach debiasing in ways that might be efficacious, serving both private and public interests in minimizing distorted or unreliable reasoning. By focusing on choices, behavior, and agent-world interactions, we suggest a broader range of outcomes for critical thinking than that informing IA, and therefore a broader range of options for critical thinking education as well.

Those familiar with the critical thinking literature can skip section 2, in which we develop and justify our characterization of IA. In section 3, we present some reasons for pessimism towards IA. Finally, in sections 4 and 5, we introduce our positive proposal by first distinguishing different ways in which we can debias and then discussing how this impacts the way we conceive critical thinking education.

2. Characterizing the intuitive approach

First, an explanation and a caveat. By 'bias' we most generally mean something neutral with respect to both moral properties and questions of accuracy. A bias in this sense is simply a disposition, implicit or explicit, to reach a particular kind of conclusion or outcome, or to remain in one. This interpretation, common in the psychological literature, accommodates the idea that biases can skew a process in a way that makes its outcome inaccurate or otherwise wrong, but it also leaves open the prospect that biases play a role in truth-conducive reasoning processes and morally unproblematic judgments or attitudes. In common parlance, of course, one normally goes to the trouble of saying that some attitude, reasoning, or person is biased only if the operation of the bias is claimed to be problematic—a distortion, or a prejudice that amounts to a vice. Our focus on debiasing is one that presumes the former meaning: it is predicated on the thought that biases should be mitigated *when* they are problematic, and not because they are by definition problematic (e.g., Klein & Kahneman 2009 explore when and how heuristics and their associated biases can help us get things

right). Not everything that is a bias needs to be debiased; only biases manifesting in a problematic manner or degree.

The caveat is that the empirical literature on biases and on debiasing is massive and varied; even to summarize it comprehensively would be impossible for a single paper. The number of biases to consider is moreover increasing as psychologists discover or propose new ones.[3] We will use just a few results we believe to be representative to illustrate the grounds for thinking that teaching students about biases, and warning them to be on the lookout for biases, is unlikely to significantly reduce the generation of distorted judgments "in the wild," or to increase the likelihood that biased judgments will be recognized and remedied by the agent herself. Because we hope to spend some time on the *implications* of this fact for critical thinking education, we are compelled to move through the empirical evidence rather briskly. So our subsequent reflections will have to remain conditional, not just on the probity of the defeasible empirical literature, but on the accuracy of our depiction of that literature.[4]

The most general problem to emphasize about IA is illustrated in Baruch Fischhoff's (1982) influential work on mitigating the *hindsight bias*. Hindsight bias is the tendency to regard actual outcomes as inevitable outcomes, in retrospect, and to overestimate the extent to which one had antecedently expected the actual outcome. Fischhoff points out that a number of approaches to debiasing subjects for hindsight effects simply do not work very well under a wide range of experimental conditions (1982, pp. 427-431). These approaches, falling under IA, include:

- Explicitly explaining the bias to subjects, and asking them to avoid it in their own reasoning;
- Inducing subjects to value the accuracy of their performance;
- Encouraging subjects to think first in terms of diagnosing other subjects' biased reasoning, before turning to the question of their own reasoning.

These strategies (and others that Fischhoff describes) are motivated by some quite natural assumptions about the nature of learning, of cognition, and of error—the most basic one being the idea behind IA, that forewarned is forearmed. But the

[3] Stanovich reviews this expanding literature and offers a useful taxonomy with relatively few categories (2009, p. 182; 2011, p. 230-243).

[4] A somewhat more detailed examination of this evidence is provided in Kenyon (2014).

ineffectiveness of these strategies at mitigating hindsight bias, and many other biases, has been quite strongly confirmed by subsequent psychological work (Wilson et al. 2002).

3. Pessimism about teaching debiasing abilities

Fischhoff's studies and subsequent ones are, perforce, largely experimental designs that isolate specific instances of biased reasoning, rather than longitudinal analyses of learning outcomes in educational contexts. The latter type of study, though, tends to be hard to perform rigorously, and hard to interpret (Willingham 2007, p. 12). Experimental designs constitute the best evidence we now possess about the propensity for normally teachable information and skills to reduce biased reasoning in students, outside the classroom and in later life. How much confidence to place in the applicability of these results is a good question; but this is the evidence we have. On balance, it weighs against the thought that simply teaching and warning people about biases will successfully mitigate biased reasoning. IA is not well supported by evidence.

Perhaps the most significant factor explaining why teaching people about biases does not itself particularly reduce their biases is known in the literature as *bias blind spot* (Pronin and Kugler 2007; Pronin, Lin & Ross 2002).[5] Put simply: knowing that people in general are subject to a particular bias is consistent with one's believing that one is not subject to it. Indeed, more importantly, even knowing that one is *generally* susceptible to a bias is consistent with one's believing, on the specific occasion one considers the matter, that one is not displaying a bias. Bias blind spot thereby insulates one's judgments, in the event, from the application of whatever debiasing strategies might actually be effective.

A related explanation for the relative ineffectiveness of teaching information about biases is that we can easily think that we have debiased when we have not. A theoretical knowledge of the need to adjust for bias does not reduce this problem, since the problem is precisely that one falsely believes oneself to have addressed that need. Indeed, merely thinking about debiasing can enable the problem! By thinking over the details of the case at hand, and considering the prospect of being biased, one may

[5] We would encourage the use of another expression that could communicate this idea just as well, without using the ableist expression *blindness* to denote a type of ignorance (cf. Schorr 1999).

simply give one's biases more raw material to operate on (Thompson 1995). Here, then, another appealing thought about clear thinking meets unwelcome data: the idea that one can debias by firmly *thinking it over*, that debiasing can be a matter of having a stern word with oneself about not being biased, is mistaken (Frantz & Janoff-Bulman 2000). In fact, attempting to self-debias in this way can even make one's biases worse (Hirt and Markman 1995; Sanna, Stocker & Schwartz 2002). Telling ourselves that we have debiased, we can come to hold our attitudes and views more strongly—convinced that they have been vetted for distortion. As Frantz (2006, p. 165) observed, merely to ask a question like "Am I being fair?" is to provide an additional opportunity for a bias to operate, accompanied by a greater conviction that one's judgment is unbiased.

This conveys a sense of the kinds of evidence speaking against the idea that we can teach people to be significantly less biased reasoners simply by teaching and warning them about biases. But this is not to say that *no* debiasing strategies have been shown to work in this literature. A range of strategies work to varying degrees, depending on the bias, with the single most effective (and most *generally* effective) strategy being for the subject to explicitly consider and entertain a range of alternative perspectives or counterfactual outcomes, and what would have had to happen in order for those outcomes to occur (Pronin, Puccio and Ross 2002; Wilson et al 2002; Anderson & Sechler 1986; Fischhoff 1982). So we do have at least one mitigation strategy with a significant prospect of success*, taken as an experimental treatment*.

The problem is that the strategy is extremely difficult to implement as a self-deployed skill. Existing biases and attentional limits can easily make themselves felt as an unwillingness or inability to generate plausible alternative scenarios (O'Brien 2009, pp. 329-330); and even a willingness to do so is no guarantee that the generation and consideration of alternatives will be sufficiently disciplined or constrained to actually lead to a less distorted judgment (Tetlock 2005, p. 199). Absent the sort of facilitation or guidance by assistants that tends to characterize the experimental contexts in which "consider the opposite" is an effective strategy, there is little reason to expect it to be employed with regularity by individual agents in normal contexts, nor to work well when it is employed.

Roughly and readily, then, there is a seeming dilemma for those who wish to teach debiasing as part of critical thinking. The things that are most easily teachable and open to long-term

retention by learners—what biases are and how they work; and that their distortive influences are to be avoided—are not in themselves very effective at debiasing people's judgments; while the things that are rather effective at debiasing judgments—counterfactual or opposite-scenario consideration—are not very teachable as individual skills to be recalled and applied when needed, nor to be implemented easily even when attempted. In either case, IA does not deliver.

Again, we do not take this as grounds to doubt that there is still a rationale for focusing on biases in critical thinking education. Rather, we believe that the difficulty of teaching effective debiasing strategies under the assumption of IA is really an invitation to a broader and more fine-grained taxonomy of debiasing outcomes than is presupposed in IA. This larger terrain of debiasing outcomes should in turn create space for additional strategies to mitigate biases in outcomes so construed. If teaching debiasing looks too hard in light of the data just described, it is because IA focuses on doing it *at the least plausible levels*: by giving students propositional knowledge that will enable them to debias, or to debias their own thinking at the point of bias manifestation. Put differently, IA requires students to debias in the most cognitively demanding way.

We propose to alleviate this problem by distinguishing further levels or domains of debiasing. These levels, we believe, are partly anticipated in the extant literature; various authors allude to some of the strategies we will explore below. But there exists at the moment no taxonomy of the kind we outline here, illuminating a wider range of overlapping skills and habits that can more plausibly be taught and implemented, with the aim of addressing biases at those different levels. We see this as supplementing existing strategies and, hopefully, as offering new ways of thinking about the challenges we outlined in this section.

4. The scope of debiasing

While knowing about a bias is no prophylactic in itself, it may serve as one of many steps along a path to debiasing (Stanovich & West 2008; Wilson & Brekke 1994). For example, Wilson & Brekke's model lists the awareness of an unwanted process as the first step in debiasing. Of course, one must also be motivated to correct the bias, know the *direction* and the *magnitude* of this bias, and be sufficiently in control, with sufficient mental

resources, to be able to adjust the response (Wilson & Brekke 1994, p. 119). Now, we have briefly reviewed grounds to believe that no individually portable suite of skills seems very apt to put these cognitive and affective resources to work at the right times. But what if we did not limit ourselves to the goal of preventing biased judgments, nor even that of unskewing judgments after they are made?

We take the problem thus far to be an artefact of the level at which we have been considering both biases and debiasing strategies. We submit that the core issues of interest from a critical thinking perspective are broader—including not simply what one thinks, but how one acts.[6] This opens up the scope of what will count as a debiasing strategy in the relevant sense. It holds out the promise that a more variegated conception of bias-reduction will offer a range of strategies that limit bias at different levels and in different ways.

In effect, we propose swapping a teaching approach that is simple in presentation but has little hope of success for an approach that will certainly be more complex in presentation, but has a greater chance of bearing fruit. The implausible approach is IA: teaching information about biases in such a way that learners will somehow subsequently recall that information, recognize its situational relevance, and act on it appropriately in bias-fraught contexts of thought and action. We advocate not only teaching information about biases, but also teaching and ingraining the habits, skills and dispositions that facilitate adopting general reasoning and decision-making principles, which nudge agents away from biased reasoning and filter its effects out of their actions.

When we talk of a *nudge*, we mean the term in the sense advanced by Thaler & Sunstein (2009). A nudge is a strategy or an infrastructure put in place in order to minimize or to eliminate a set of cognitive biases by using aspects of the environment. Changing the way information is presented to participants or changing what the default option is are common examples of nudges. A striking example from Thaler & Sunstein's discussion is the way food is displayed in a buffet: depending on where certain food items are placed, their "popularity" as a choice can increase by 50%. The idea here is to pre-emptively construct situations in order to minimize biases. From an individual agent's perspective, this presents two

[6] Cf. Beaulac & Robert (2011) on critical thinking *attitudes*. One strategy they deem particularly promising is that of *epistemic caution* (prudence épistémique).

dimensions of action to mitigate bias: exploiting existing nudges in the environment, and constructing nudges of one's own—either individually or collaboratively. Both the ability and the need to do these things generally are potential learning outcomes for a critical thinking course, outcomes that are insufficiently explored in the critical thinking literature at the moment. We believe our outlook gives better tools to integrate these ideas in the curriculum.

While a more fine-grained analysis is surely possible, we will for the sake of brevity limit the current discussion to four broad levels at which debiasing can be implemented, once it is taken to span the distinction between thought and action. We provide both a general description and an example for each level. It is worthy being clear, however, that this is meant as a broad characterization of these levels. We recognize that the divisions between the levels are not razor-sharp; there may be borderline cases, and complex examples might bridge across levels.

> Level 1 debiasing: Owing to general education, environment (family, neighbourhood, education, etc.), habits, critical thinking education and training over a long period, an agent has no disposition to produce a particular sort of biased judgment; that is, the bias does not arise. This sort of debiasing process is implemented during education (mostly on a very long period) and applies to individual agents' judgments.
>
> E.g., A hiring committee member does not notice or attend to racial differences, and shows no bias in reasoning about the quality of candidates from visible minority groups in hiring contexts because she grew up and still lives in a multicultural neighborhood and has not been markedly influenced by the media characterization of some groups.

> Level 2 debiasing: A biased judgment occurs or is incipient, but critical thinking education and training facilitate the agent's deployment of cognitive or behavioral strategies that lead to a revision of the judgment in context. Debiasing of this kind is implemented within the context of judgment-fixation, is initiated and mediated by agents' psychological processes, and applies to individual agents' judgments. (E.g., models by Stanovich & West 2008, Wilson & Brekke 1994)

E.g., A hiring committee member's first reaction is to assign an unwarrantedly low rating to a dossier from a candidate with a name connoting ethnic minority status. On second thought, though, she wonders whether she is being biased by the character of the name, and reflects on the positive features of the file. Eventually she comes to think of the candidate in more accurate terms.

Level 3 debiasing: A biased judgment occurs or is incipient, but critical thinking education and training (individual or collective) leads (or has led) to the creation of situational "nudges" that debias the agent's judgment in context. This sort of debiasing process is implemented within the context of judgment-fixation, is initiated or mediated by environmental cues or infrastructure, and applies to individual agents' judgments.

E.g., A hiring committee is given a preliminary presentation about the prospects for biased reasoning in hiring contexts. Notes and other guidelines from this presentation are kept in the meeting room, in a red folder on the table around which committee members sit. Later, a hiring committee member encounters a dossier from a candidate with a name connoting ethnic minority status. The visual salience of the red folder reminds her to attend to the significance of the candidate's name. She would otherwise have assigned an unwarrantedly low rating to the file, but owing to the earlier presentation she makes a point of reflecting on the candidate's positive features, considers how those features would appear if part of a privileged candidate's application, and ranks the file more accurately.

Level 4 debiasing: A biased judgment occurs, and is not significantly remedied, but situational constraints nevertheless debias the action or outcome. This type of debiasing process is implemented over time, both in advance of and during the context of judgment-fixation. It is initiated or mediated by environmental cues or infrastructure, and applies to group judgments, or to actions and outcomes.

E.g., A hiring committee member has an uncorrected bias of judgment against women in the profession; but

> anonymized applications hide candidates' gender information, and the committee member ultimately (unknowingly) votes to hire a superior woman candidate.
>
> E.g., A hiring committee member displays uncorrected biased reasoning in judging that a superior candidate should not be hired because of her sexual orientation; but declines to voice this view in light of the negative responses it would draw from colleagues, and ultimately votes in favor of the candidate.
>
> E.g., A hiring committee member displays uncorrected biased reasoning in judging that a superior candidate has an inferior track record, but the majority vote of the hiring committee favors the candidate, and she is offered the job anyhow.

The levels represent a way of carving up the gradient from the most individualist and internalist character-driven approaches, to the most outcomes-oriented and externally-mediated approaches. We can characterize Levels 1 and 2 as the more individualistic levels; they essentially treat the particular agent as both the source and the focus of debiasing outcomes. Levels 3 and 4 appeal to external, situational factors to a greater extent.

Level 3 debiasing retains a crucial individualistic component, since the "nudges" or external aids to reasoning that it postulates are devoted to mitigating biases in the individual agent. The humble notion of a *reminder* generalizes this approach to contexts far beyond that of debiasing, whether in form of a string tied around one's finger, or in the government-mandated installation of seatbelt reminder lights and noises in motor vehicles. Just like in those more general cases, the main advantages of a Level 3 approach to debiasing have to do with reducing the cognitive load on the individual agent faced with a bias-detection problem. As Stanovich & West (2008) and Wilson & Brekke (1994) aptly observe, factors bearing on the detection of the bias are some of the most important reasons why agents end up following their biased judgments. Level 3 debiasing strategies place this crucial *detection* stage outside the agent's mind, making this strategy cognitively easier than Level 2 strategies are. This makes the detection of the bias more likely.

In Level 4 debiasing, this individual aspect is minimized, in some cases to the point of being eliminated altogether. This breadth of degree makes Level 4 a relatively broad and complex

class of debiasing approaches, ranging from those that forestall or minimize individual biases to those that tolerate the occurrence of individually manifest biased judgments, but minimize their significance in determining actions or outcomes. What distinguishes even those at the former sort end of this spectrum from the debiasing strategies of Level 3 is that they do not, at the point of decision-making, operate by debiasing the judgments of individual agents through the cognitive operations of those agents.

In the Level 3 example provided, an object in the environment leads to less distorted reasoning by being noticed, and by focusing the agent's attention in a potentially corrective way. While the first two Level 4 examples are also somewhat focused on individual agents, they do not involve the use of facilitated individual cognition and perception to promote debiased individual judgment. The efficacy of anonymizing applications in hiring, as in the first Level 4 example, is manifest in individual judgments, but does not require the agent to reflect on or notice the anonymization. Conversely, in the second Level 4 example, the agent who notices and reflects on the social costs of displaying prejudice during group decision-making, and elects not to do so, need not be debiased in judgment in order for the relevant outcome to be debiased. Thus we propose a gradient of Level 4 strategies, some of them resembling Level 3 strategies in their scope and orientation, but with an internal unity and distinctiveness all the same. The external factors invoked in Level 4 strategies are essentially oriented towards debiasing decisions, actions, and outcomes—including group outcomes—without specific reference to the dispositional properties of any particular agent.[7] Level 4 debiasing will of course still have individualistic overtones dynamically, since an agent may learn what and how better to think about an issue by seeing a debiased outcome and process. Indeed, this may well be a valued feature of such debiasing efforts and infrastructure over the longer term. But it is not a defining feature of Level 4 debiasing success.

It is worth noting that some approaches to debiasing already exist in the psychological literature, having some overlap with elements of our taxonomy. The particular granularity of our formulations strike us as more felicitous, however, and the employment of *levels* is a key refinement. For example, Croskerry et al. (2013b) place the strategy of "training

[7] Bishop & Trout (2005) discuss the wider ramifications of such strategies for epistemology.

on theories of reasoning and medical decision making" on a par with creating "supportive environments" for sound reasoning (pp. ii67-ii68). This tends to obscure not only the significantly distinct causal domains associated with these strategies—the reasoner's psychology versus the reasoner's environment—but also the correspondingly different material, structural and educational preconditions they require. The latter kinds of difference are especially critical if one's aim in both cases is to impart knowledge and training that will enable the relevant strategies.[8] Teaching theories of reasoning is plausibly a very different undertaking than teaching the skills of creating and employing a supportive reasoning environment. Similarly, even though Larrick (2004) and Soll et al. (forthcoming) distinguish between modifying *the person* and modifying *the environment*, we suggest here a more fine-grained analysis that can reveal more alternative strategies for mitigating distorted outcomes.

Our conjecture is that, when it comes to biases, an approach animated by IA treats Level 1 outcomes as the ideal (the bias should not come up at all), and strives at least to bring about Level 2 outcomes (if a bias comes about, the agent can correct it). We think this is practically impossible; if such education is ever effective, it is more likely because elements of the education itself are acting as persistent nudges to create occasional Level 3 outcomes, while the value of Level 4 outcomes is learned by trial and error, if at all, and is implemented relatively haphazardly. The impetus to treat Levels 1 and 2 as the real aim of critical thinking education depends, we think, not on evidence that this is a practical possibility, but substantially on a deep-seated intuition that critical thinking is properly implemented only in the minds and choices of specific agents.

The three distinct examples for Level 4 debiasing reflect both the flexibility of the individuation of actions, and the range of points at which debiasing action can take place. The first example proposes an intervention affecting the agent's judgment of candidates; with the second, the intervention debiases the agent's act of voting; while the third describes a mitigation of nothing more specific than the committee's collective hiring actions. The anonymized hiring protocol, the perception of social disapproval of prejudice, and the committee voting

[8] Croskerry et al. (2013a, 2013b) may well be contemplating a clinical administration that *teaches* theories of reasoning while itself directly *implementing* supportive infrastructure such as decision check-lists; we are contemplating how to bring about both kinds of outcome through education.

structure each count as an element of contextual engineering that effectively debiases the Level 4 scenario, even though all Level 4 cases *by definition* count as failures of debiasing by the purely individualistic cognitive standards we originally considered.

Clearly, then, this more fine-grained analysis reveals more opportunities to debias by clarifying the number of stages open to intervention in thinking, preparing, deciding and acting. Variations on the theme are not hard to find, moreover, including some that span the levels we have sketched. For example, Uhlmann and Cohen (2005) found that, if the notion of merit were left undefined for a hiring process, it would tend to become the vehicle of gender-biased decision making. That is, merit would be operationalized distinctly from case to case, with the overall effect of promoting hiring along gender lines—and particularly the hiring of men over women.[9] But eliciting a commitment to some hallmarks of merit from the evaluators *prior* to revealing information about the people being evaluated reduced this biased "moving goalposts" approach in their judgments (2005, p. 478). The example provides further empirical support for the idea that education about the advance construction and acceptance of such policies and organizational structures should fall within the core mandate of education for reaching more appropriately reflective and reliable outcomes in reasoning.

Arguably this counts as a remedy that straddles the border between Levels 3 and 4, since the incipient bias is corrected in judgment, not merely in action or outcome; yet in practice the mechanisms achieving this outcome will be thoroughly environmental and causally remote. That is, somebody has to decide (presumably well in advance, in the case of policy-making) to set out clear rubrics for merit, and to ensure a hiring process structured so that evaluators review the hallmarks of merit before they review the details of applicants. So not every case of debiasing falls entirely within one such level; but we do think that this particular way of carving up of levels helps illuminate relevant features of even those cases spanning levels.

[9] There was also weak evidence that female evaluators would similarly construct merit in a gender-biased way to devalue male applicants, if the job were sufficiently stereotypically associated with women's gender roles—e.g., that of a Women's Studies professor (2005, p. 478).

5. Teaching debiasing as teaching acceptance of influences on cognition and constraints on action

There is, then, a very broad recipe for achieving better odds of teaching successful debiasing strategies: first broaden the conception of what counts as debiasing, and then be open to exploiting the full spectrum of opportunities to mitigate bias, from antecedent reasoning dispositions to the broadest conception of an action in context. We close with some schematic remarks about putting debiasing, so construed, into a typical critical thinking curriculum.

The approach we suggest becomes more plausibly effective than the debiasing strategies described earlier when it motivates us to subject ourselves to nudges, infrastructure and institutions in advance of the circumstances of bias that will make those things effective debiasing aids. That is, we hold that knowledge of biases has the best chance of effectiveness when it leads one generally to accept and construct nudges or contextual engineering of one's own. In that case it supports the adoption of general debiasing strategies that might simply be encoded in the lived environment, rather than holding out the hope that one can learn to debias in a series of contextual one-offs, as the need arises.

Of course, it is also important to note that, at this time, there are few direct empirical grounds for confidence that teaching skills and attitudes specifically to promote Level 3 and Level 4 debiasing in critical thinking courses will be easy or highly effective. We do not claim to show this; only that it is worth trying. Perhaps the most certain line of reasoning at our disposal is probabilistic. Unless the probability of success under our broader construal of debiasing outcomes is literally zero, the addition of this slate of options can only improve the chances of overall success in teaching debiasing skills. How close to zero that probability could be while yet justifying the effort of the attempt is a good question that we will not attempt to answer beyond offering three observations: first, that testing such payoffs is what pilot projects and exploratory studies are for; second, that critical thinking education incorporating IA already consumes many resources when its low chances of debiasing success *are* known; and third, that the chances of success, on our account, are unlikely to be that low. After all, the creation of and deference to bias-reducing infrastructure is palpably something that can spread through professional mentorship and collegial training. For example, instructors demonstrably can acquire from their peers various debiasing practices such as

anonymizing student work. When these outcomes are the explicit objects of training, they seem teachable and learnable; we think this shows that one would need a special reason for thinking that they are not largely or significantly teachable in courses that explicitly aim to teach them, rather than a special reason to think that they are. So we do not claim that our approach is sure bet to succeed, but we do claim that it is a reasonable bet, and in any case a better bet than what IA represents.

On our view, an education in debiasing includes an education in how to administer decision-making contexts and actions in a manner consonant with Level 3 and Level 4 debiasing. Analogies and antecedents might be found with a range of cases of training in controlling one's environment and actions, rather than merely one's internal states. For example, clinical psychologists and psychotherapists refer to the strategy of controlling one's environment in order to regulate thoughts and behaviour as *stimulus control*. When manifest as a kind of self-regulation it is a familiar and central element of many forms of (teachable, learnable) therapies, including Cognitive-Behavioural Therapy (Karoly 2012, p. 201). Roughly speaking, rather than trying merely to teach patients suffering from alcoholism or gambling addictions how to avoid drinking *while at a bar*, or how to avoid gambling *while at the casino*, mitigation strategies include also teaching the ability to avoid the bar and the casino in the first place. Choices that determine one's environmental stimuli have profound influences on the sort of thoughts and actions that follow.

Similarly, choices ranging from how to form committees, how to solicit information, which buttons to push on the television remote control, and whether to ask about someone's personal details during a job interview can all powerfully influence the opportunities for biases to be reflected in our actions. It follows that the knowledge (both knowledge-that and knowledge-how) associated with those activities are reasonable components of an education in critical thinking. This knowledge will include skills of creating and maintaining physical, institutional and social infrastructure that facilitates more truth-conducive reasoning. But often this infrastructure already exists when students and former students encounter contexts of judgment and action; in those cases, the relevant skill will be that of deferring to such truth-conducive mechanisms. How to teach this knowledge and these action principles is a good question. Its feasibility, though, seems far more promising than the mere hope of IA, that some

combination of knowledge of biases and mental continence will be both effective and learnable.

It is worth considering a potential objection to the Level 4 style of debiasing education proposed here, proceeding from an amalgam of epistemological and pedagogical scruples. The worry is this: whatever the didactic barriers to focusing on Levels 1, 2 and 3 as debiasing strategies, addressing one's teaching to these levels at least promotes the right connection between methods and outcomes. By teaching students to recognize bias-inducing situations and to mobilize appropriate debiasing strategies in context as individuals, one would be teaching students to make cogent *inferences* regarding the need for unbiased or less-biased reasoning. Level 4's blunt focus on debiased *outcomes* does not require anyone in the context to appreciate the problem, nor why it is a problem, nor how the debiasing mechanisms will address the problem. For all that a Level 4 approach tells us, successful debiasing processes can be entirely arational from the perspective of the agents in the situation.

Thinking back to our examples of Level 4 debiasing, therefore, one might ask: How can these be critical thinking strategies, strictly speaking, when they encompass solutions that do not involve thinking about the problem at all? On this objection, such an approach to debiasing in critical thinking education misses something valuable about students' understanding of the rational connections between reasons and outcomes – something that students in a critical reasoning course should be taught to entertain, not to elide.

The worry is based on an overly narrow conception both of the scope of the problem and of the problem-solving context. Here it may be useful to return to the analogy with addiction patients who avoid pathological activities by avoiding situations that lead to those activities. The gambling addict need not avoid gambling situations solely by reflecting on the evils of gambling, nor need she choose to do some other activity on the basis of such reflections *at the time of the engaging in the alternative activity*. She might engage in a non-gambling activity out of sheer habit; but if she originally cultivated that habit as a means of avoiding gambling, then any particular case of avoidance by way of that activity reflects her considered judgment and her autonomy.

The proposal at hand puts forward a similar (minimally sufficient) connection between education about biases and students' subsequent participation, possibly just from habit, in cognitive and social routines and practices that promote reliable

reasoning. A rational and appropriately agent-endorsed connection between outcomes and methods is established when students are educated about the need to form such habits, or to defer to truth-conducive judgment and action mechanisms. Subsequently acting on those habits need not itself be an exercise in reasoning or inference at the point of action in order to be an exercise of the agent's commitment to critically informed reasoning. Indeed, consciously reviewing one's reasons at the point of decision-making might even disrupt the debiasing process. To deny that this represents the exercise of critical reasoning is to deny, by parity of reasoning, that the lifelong alcoholic who cultivates a preference for badminton as an alternative to hanging out in a pub is not demonstrating a willful continence regarding alcohol when she remains sober for years on end by spending time at the gym.

Examples of a similar shape are already implemented in some institutions, with the case of anonymized musical auditions being particularly telling. Women have long been underrepresented in orchestras around the world, comprising fewer than 10% of musicians in major American orchestras prior to the 1970s and little more than 20% in the 1980s, a much lower proportion than their availability in the hiring "pipeline" (Goldin & Rouse 2000). Of the various practices introduced by orchestras to reduce biases that might account for this imbalance, the most common means is to ensure the anonymity of candidates during auditions, by placing the musician behind a screen where he or she plays for 5 to 10 minutes (Goldin & Rouse 2000, p. 722). The screen is not used uniformly across orchestras; only three of the 11 orchestras discussed in by Goldin and Rouse use it all the way through the process (2000, p. 723). The effects of the screen, however, are remarkable: when the screen was used throughout the process, the probability that a woman would be offered the job was 60% higher than without it.

The use of the screen is clearly a Level 4 debiasing strategy in our taxonomy: its success in debiasing the outcome of the decision process does not require reduction of the dispositional or occurrent biases of the individual deciders at the point of evaluating candidates. Yet the prior decision to implement a general policy of anonymized auditioning is plausibly driven by just the sort of empirical details about biases, and commitments to erring on the side of caution, that a sound critical thinking education may inculcate. *This* sort of decision, made well in advance on the basis of general principles, is not hostage to the need for agents to recognize in

the context of judgment that they are biased. Nevertheless, it is a touchstone case of a critical thinking strategy that depends crucially on agents' thinking about the problem. It just enables them to think at arm's length from the situations in which the bias itself will disrupt their capacities to mitigate it. Teaching within critical thinking courses how effective such approaches are will hopefully increase the proportion of students trying to solve problems by implementing such Level 3 and 4 strategies within their own (current or future) workplace.

6. Conclusion

There is a familiar learning model associated with propositional knowledge of the "Paris is the capital city of France" sort. There is also a familiar set of habits of learning and application that enables students (and former students) to apply that knowledge over the longer term of their lives. The problem with IA is that this sort of knowledge—that biases operate in particular ways, that they occur in situations like the one at hand, and that one is susceptible to them in contexts like this—does not reliably issue in debiasing behaviours at the point of decision or judgment. A wider view of what counts as successful debiasing indicates a richer class of ways to apply teachable knowledge to the project of debiasing.

Our view, then, is that critical thinking education should include extensive practical guidance on how to structure and engage with one's environment to promote good reasoning. This will include teaching how and why to adopt decision-making policies and evidence-gathering practices that do not require the virtuoso ability to rise above invisible and subtle biases. And it will offer learners the opportunity to practice and experiment with infrastructure creation and reasonable epistemic deference. The intended learning outcomes on our model do include individual learners' coming explicitly to reason more truth-conducively in specific cases. But they also include, and place great emphasis upon, outcomes that are implicit and habitual from the individual's perspective, and which have their main intended effect over the long term and at group levels.

What kind of information, advice, guidance and practice will critical thinking courses of this sort offer? How are these things best taught? These, we think, are among the next big questions in critical thinking education.

Acknowledgements: We would like to acknowledge the helpful input of Veromi Arsiradam, Frédéric-I. Banville, Gillian Barker, Frédéric Bouchard, Samantha Brennan, David DeVidi, Carla Fehr, Christine Logel, Carolyn McLeod, Chris Viger, Audrey Yap, Frank Zenker, and two anonymous referees for this journal. Thanks also to Catherine Hundleby and the students in the 2014 "Fallacies and Bias" seminar at the University of Windsor. This work was supported in part by the Faculty of Arts, University of Waterloo, and by Social Sciences and Research Council of Canada Grant 410-2011-1737 and Postdoctoral Fellowship 756-2014-0319.

References

Anderson, C. & E. Sechler. 1986. Effects of explanation and counterexplanation on the development and use of social theories. *Journal of Personality and Social Psychology* 50: 24-34.

Beaulac, G., & S. Robert. 2011. Théories à processus duaux et théories de l'éducation : le cas de l'enseignement de la pensée critique et de la logique. *Les ateliers de l'éthique* 6.1: 63–77.

Bishop, Michael A, & J. D Trout. 2005. *Epistemology and the Psychology of Human Judgment*. New York: Oxford University Press.

Croskerry, P., G. Singhal, & S. Mamede. 2013a. Cognitive Debiasing 1: Origins of Bias and Theory of Debiasing. *BMJ Quality & Safety* 22 (Suppl 2): ii58–ii64. doi:10.1136/bmjqs-2012-001712.

Croskerry, P., G. Singhal, & S. Mamede. 2013b. Cognitive Debiasing 2: Impediments to and Strategies for Change. *BMJ Quality & Safety* 22 (Suppl 2): ii65–ii72. doi:10.1136/bmjqs-2012-001713.

Fischhoff, B. 1982. Debiasing. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press: 422-444.

Frantz, C. 2006. I AM being fair: The bias blind spot as a stumbling block to seeing both sides. *Basic and Applied Social Psychology* 28.2: 157–167.

Frantz, C. & R. Janoff-Bulman. 2000. Considering both sides: The limits of perspective-taking. *Basic and Applied Social Psychology* 22: 31–42.

Gilovich, T. 1991. *How We Know What Isn't So*. New York: The Free Press.

Goldin, C., & C. Rouse. 2000. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *The American Economic Review* 90.4: 715–741.

Groarke, L. & C. Tindale. 2004. *Good Reasoning Matters*. Toronto: Oxford University Press.

Hirt, E. R., & K.D. Markman. 1995. Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology* 69: 1069–1086.

Kahneman, D., & G. Klein. 2009. Conditions for Intuitive Expertise: A Failure to Disagree. *American Psychologist 64* (6): 515–26. doi:10.1037/a0016755.

Karoly, P. 2012. Self-regulation. In *Cognitive Behavior Therapy: Core Principles for Practice*. W. O'Donohue & J. Fisher (eds). Hoboken, NJ: Wiley. 183-123.

Kenyon, T. 2008. *Clear Thinking in a Blurry World*. Toronto: Nelson Academic

Kenyon, T. 2014. False polarization: Debiasing as applied social epistemology. In *Synthese* 191.11: 2529-2547.

Larrick R. 2004. Debiasing. In *The Blackwell Handbook of Judgment and Decision Making*. D. Koehler & N. Harvey (eds). Oxford: Blackwell Publishing, 2004:316–37.

Lilienfeld, S., R. Ammirati & K. Landfield. 2009. Giving debiasing away. *Perspectives on Psychological Science* 4.4: 390-8.

O'Brien, B. 2009. Prime suspect: An examination of factors that aggravate and counteract confirmation bias in criminal investigations. *Psychology, Public Policy, and Law* 15.4: 315-334.

Pettigrew, T. F. 1998. Intergroup Contact Theory. *Annual Review of Psychology* 49.1: 65–85.

Pronin, E. & M. Kugler. 2007. Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology* 43. 4: 565–578.

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

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

Reed, R., & A. Sharp (eds.) 1992. *Studies in Philosophy for Children*: *Harry Stottlemeier's Discovery*. Philadelphia: Temple University Press.

Ruggiero, V.R. 2004. *The Art of Thinking*. New York: Pearson Longman.

Sanna, L., S. Stocker & N. Schwarz. 2002. When debiasing backfires: Accessible content and accessibility experiences in debiasing hindsight. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28.3: 497-502.

Schorr, N. 1999. Blindness as Metaphor. *Differences: A Journal of Feminist Cultural Studies, 11*(2): 76-105.

Soll J. B., Milkman K. L., Payne J. W. (in press). A user's guide to debiasing. In Wu G., Keren G. (Eds.), Blackwell Handbook of Judgment and Decision Making (2nd edition). New York, NY: Wiley. (Full ToC http://faculty.chicagobooth.edu/george.wu/teaching/public/38913/handbook2014.htm)

Stanovich, K. E. 2009. *What Intelligence Tests Miss : the Psychology of Rational Thought*. New Haven: Yale University Press.

Stanovich, K. E. 2011. *Rationality and the Reflective Mind*. New York: Oxford University Press.

Stanovich, K. E., & West, R. F. 2008. On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology* 94.4: 672–695.

Thaler, R. H., & Sunstein, C. R. 2009. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven: Yale University Press.

Tetlock, P. 2005. *Expert Political Judgment*. Princeton, NJ: Princeton University Press.

Thompson, L. 1995. "They saw a negotiation": Partisanship and involvement. *Journal of Personality and Social Psychology* 68: 839–853.

Uhlmann, E. & G. Cohen. 2005. Redefining merit to justify discrimination. *Psychological Science* 16.6: 474-480.

Willingham, D. 2007. Critical thinking: Why is it so hard to teach? *American Educator* 31.2: 8-19.

Wilson, T. D., & Brekke, N. 1994. Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin* 116.1: 117–142.

Wilson, T. D., Centerbar, D. B., & Brekke, N. 2002. Mental Contamination and the Debiasing Problem. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.) *Heuristic and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press. 185-200.