Cognitive debiasing 1: origins of bias and theory of debiasing

Pat Croskerry,1 Geeta Singhal,2 Sílvia Mamede3

ABSTRACT
Numerous studies have shown that diagnostic failure depends upon a variety of factors. Psychological factors are fundamental in influencing the cognitive performance of the decision maker. In this first of two papers, we discuss the basics of reasoning and the Dual Process Theory (DPT) of decision making. The general properties of the DPT model, as it applies to diagnostic reasoning, are reviewed. A variety of cognitive and affective biases are known to compromise the decision-making process. They mostly appear to originate in the fast intuitive processes of Type 1 that dominate (or drive) decision making. Type 1 processes work well most of the time but they may open the door for biases. Removing or at least mitigating these biases would appear to be an important goal. We will also review the origins of biases. The consensus is that there are two major sources: innate, hard-wired biases that developed in our evolutionary past, and acquired biases established in the course of development and within our working environments. Both are associated with abbreviated decision making in the form of heuristics. Other work suggests that ambient and contextual factors may create high risk situations that dispose decision makers to particular biases. Fatigue, sleep deprivation and cognitive overload appear to be important determinants. The theoretical basis of several approaches towards debiasing is then discussed. All share a common feature that involves a deliberate decoupling from Type 1 intuitive processing and moving to Type 2 analytical processing so that eventually unexamined intuitive judgments can be submitted to verification. This decoupling step appears to be the critical feature of cognitive and affective debiasing.

INTRODUCTION
Clinical decision making is a complex process. Clinical decisions about patient’s diagnoses are made in one of two modes: either intuitive or analytical, also referred to, respectively as Type 1 and Type 2 processes. The former are more commonly used. They are fast, usually effective, but also more likely to fail. As they are largely unconscious, mistakes—when they occur—are seldom corrected.1–3 In contrast, Type 2 processes are fairly reliable, safe and effective, but slow and resource intensive. The intuitive mode of decision making is characterised by heuristics—short-cuts, abbreviated ways of thinking, maxims, ‘seen this many times before’, ways of thinking. Heuristics represent an adaptive mechanism that saves us time and effort while making daily decisions. Indeed, it is a rule of thumb among cognitive psychologists that we spend about 95% of our time in the intuitive mode.4

We perform many of our daily activities through serial associations—one event automatically triggers the next with few events of deliberate, focused, analytical thinking. We have a prevailing disposition to use heuristics, and while they work well most of the time, they are vulnerable to error. Our systematic errors are termed biases,3 and there are many of them—over a hundred cognitive biases5 and approximately one dozen or so affective biases (ways in which our feelings influence our judgment).6 Bias is inherent in human judgment, and physicians are, of course, also subject to them.

Indeed, one of the principal factors underlying diagnostic error is bias.7–9 Post hoc analyses of diagnostic errors10 11 have in fact suggested that flaws in clinical reasoning rather than lack of knowledge underlie cognitive diagnostic errors, and there is some experimental evidence that, at least when problems are complex, errors were associated with intuitive judgments and could be repaired by analytical reasoning.12–14 Moreover, a few experimental studies have supported the claim that bias may misdirect diagnostic reasoning, thus leading to errors.15 16 While the evidence for the role of bias in medical diagnostic error is still scarce, research findings in other domains1 3 is sufficient
to justify concerns with the potential adverse influence of bias on diagnostic reasoning. Two clinical examples of biased decision making leading to diagnostic failure are given in Case 1 and Case 2, displayed, respectively, in boxes 1 and 2. From the case descriptions, the mode of decision making the physician was relying upon cannot be determined. This is indeed a limitation of studies on diagnostic reasoning, which only have indirect evidence or post hoc inference of reasoning processes, at least until other tools such as functional MRI (fMRI) can be experimentally employed.

In this first paper, we discuss how biases are generated, and situations that make physicians more vulnerable to bias. Building upon dual process theories (DPTs) of reasoning, we discuss the origins of

---

**Box 1 Example of biased decision making leading to diagnostic failure: case 1**

A 55-year-old man presents to a walk-in clinic towards the end of the evening. It has been a busy day for the clinic and they are about to close. His chief complaint is constipation. He has not had a bowel movement in 4 days, which is unusual for him. He complains of pain in his lower back and lower abdomen and also some tingling in his legs. He thinks that he will feel better with a laxative because what he has tried so far has not worked. He was briefly examined by the physician who did not find anything remarkable on his abdominal examination. There was some mild suprapubic tenderness, which was attributed to the patient’s need to urinate. Bowel sounds were good, his abdomen was soft and there were no masses or organomegaly. The physician prescribed a stronger laxative and advised the patient to contact his family doctor for further follow-up.

During the night, the patient was unable to urinate and went to the emergency department. On examination, his lower abdomen appeared distended and he was found to have a residual volume of 1200cc on catheterisation. His rectum was markedly distended with soft stool that required disimpaction. He recalled straining his lower back lifting about 4 days earlier. The emergency physician suspected cauda equina syndrome and this was confirmed on MRI. He was taken immediately to the operating room for surgical decompression. He did well postoperatively and regained full bladder control.

**Comment:** The patient was initially misdiagnosed and might have suffered permanent loss of bladder function requiring lifelong catheterisation. The principle biases for the physician who saw him in the clinic were framing, search satisficing and premature diagnostic closure. These may have led to the incomplete history-taking, an incomplete physical exam, failure to consider symptoms that appeared discordant with constipation (back pain, leg paresthesias) and to consider other diagnoses. Fatigue may have been a contributing factor; it is known to increase the likelihood of defaulting to System 1 and vulnerability to bias (for a description of particular cognitive biases, see ref. 7 17 18).

---

**Box 2 Example of biased decision making leading to diagnostic failure: case 2**

A mildly obese, 19-year-old woman is admitted to a psychiatric hospital for stabilisation and investigation. She has suffered depressive symptoms accompanied by marked anxiety. Over the last week she has had bouts of rapid breathing which have been attributed to her anxiety. However, she has also exhibited mild symptoms of a respiratory infection and the psychiatry resident transfers her to a tertiary care department (ED) of a tertiary care hospital to ‘rule out pneumonia’. She is on no medications other than birth control pills. At triage, she is noted to have an elevated heart rate and respiratory rate. She is uncomfortable, anxious and impatient, and does not want to be in the ED. She is noted to be ‘difficult’ with the nurses.

After several hours she is seen by an emergency medicine resident who finds her very irritable but notes nothing remarkable on her chest or cardiac examination. However, to ensure that pneumonia is ruled out he orders a chest X-ray. He reviews the patient with his attending noting that the patient is anxious to return to the psychiatric facility and is only at the ED ‘because she was told to come’. He expresses his view that her symptoms are attributable to her anxiety and that she does not have pneumonia. Nevertheless, he asks the attending to review the chest X-ray to ensure he has not missing something. The attending confirmed that there was no evidence of pneumonia and agreed with the resident that the patient could be returned to the psychiatric hospital.

While awaiting transfer back to the psychiatric hospital the patient requests permission on several occasions to go outside for a cigarette and is allowed to do so. Later, on the sidewalk outside the ED, the patient has a cardiac arrest. She is immediately brought back into the ED but could not be resuscitated. At autopsy, massive pulmonary saddle emboli are found as well as multiple small emboli scattered throughout both lungs.

**Comment:** The patient died following a diagnostic failure. Various cognitive and affective biases are evident. The principle ones are framing, diagnostic momentum, premature diagnostic closure and psych-out error (see ref. 7 for a description of the biases). The patient’s demeanour towards staff and the resident may have engendered some negative antipathy which may have further compromised decision making.
Biases, and the theoretical basis of how cognitive debiasing actually works. In the second paper, we summarise the known strategies for cognitive debiasing that have been proposed to counteract different types of biases.

Biases are ‘predictable deviations from rationality’. Many biases that diagnosticians have can possibly be recognised and corrected. Essentially, this is the process that underlies learning and refining of clinical behaviour. We may have acquired an inappropriate response to a particular situation that, in turn, leads to a maladaptive habit. Through feedback, however, or other processes, some insight or revelation occurs and we are able to change our thinking to achieve a more successful outcome. The basic premise is that if we can effectively debias our thinking, from innate and learned biases, we will be better thinkers and more accurate diagnosticians.

Clinical decision making is a complex process. Besides the overall vulnerability of the human mind towards biases in decision making, it is generally appreciated that the quality of decision making is also influenced by ambient conditions: prevailing conditions in the immediate environment—context, team factors, patient factors, resource limitations, physical plant design and ergonomic factors. Individual factors such as affective state, general fatigue, cognitive load, decision fatigue, interruptions and distractions, sleep deprivation and sleep-debt, are influential too. Other individual factors such as personality, intelligence, rationality, gender and other variables also impact decision making. Since our judgments are so vulnerable to biases and so many different factors affect decision making, the challenge of developing clinical decision makers who consistently make optimal and reliable decisions appears daunting. Two questions need to be answered: can we improve our performance by using cognitive debiasing to repair incorrect judgments made under the influence of bias? This means appropriately alerting the analytical mode to situations in which a bias might arise so that it can be detected and a debiasing intervention applied. Second, can we mitigate the impact of adverse ambient conditions, either by improving conditions in the decision making environment, or by changing the threshold for detection of bias and initiating debiasing strategies?

THE ORIGINS OF COGNITIVE BIASES

There appear to be a prevailing assumption that biases are all created equal, that all are difficult to overcome and that some common debiasing strategy might work. However, as Larrick points out, many biases have multiple determinants, and it is unlikely that there is a ‘one-to-one mapping of causes to bias or of bias to cure’; neither is it likely that one-shot debiasing interventions will be effective. From DPT and other work of cognitive psychologists we know that, although biases can occur in both types of processes, most biases are associated with heuristics and typically are Type 1 (intuitive) processes. Other theories of reasoning exist, but DPT has become prevalent, gaining increasing support including from functional MRI studies. DPT has been used as a template for the diagnostic process (figure 1).

In the figure, the intuitive system is schematised as Type 1 processes and the analytic system by Type 2 processes. There are eight major features of the model:

1. Type 1 processing is fast, autonomous, and where we spend most of our time. It usually works well, but as it occurs largely unconsciously and uses heuristics heavily, unexamined decision making in the intuitive mode is more prone to biases.
2. Type 2 processing is slower, deliberate, rule-based and takes places under conscious control, which may prevent mistakes.
3. The predictable deviations from rationality that eventually lead to errors tend to occur more frequently in the Type 1 processes, in line with findings of dual-process researchers in other domains.
4. Repetitive processing using Type 2 processes may allow processing in Type 1. This is the basis of skill acquisition.
5. Biases that negatively affect judgments, often unconsciously, can be overridden by an explicit effort at

![Figure 1](image1.png)

**Figure 1** Dual process model for decision making. From: Croskerry (T is the toggle function, which means that the decision maker is able to move forth and back between Type 1 and Type 2 processes).
reasoning. Type 2 processes can perform an executive override function—which is key to debiasing.

6. Excessive reliance on Type 1 processes can override Type 2, preventing reflection and leading to unexamined decisions—this works against debiasing.

7. The decision maker can toggle (T) back and forth between the two systems—shown as broken line in figure 1.

8. The brain generally tries to default to Type 1 processing whenever possible

These operating characteristics have been described in more detail elsewhere. The model does not imply that one single reasoning mode accounts for a diagnostic decision or that a particular mode is always preferable over the other one. Current thinking is that making diagnoses usually involves some interactive combination of intuitive and analytical processing in different degrees. And whereas in some circumstances a high degree of System 1 processing may work well or be even lifesaving, such as in imminent life-threatening conditions, in others a high degree of reflection (System 2) may be required. Optimal diagnostic reasoning would appear to be a blend of the two reasoning modes in appropriate doses.

Further, not all biases originate in Type 1 processing, but when a bias does occur it can only be dealt with by activating Type 2 processing. Thus, a good balance of Type 1 and Type 2 processes is required for a well-calibrated performance.

Importantly, the intuitive processes are multiple and varied. Stanovich has recently categorised these ‘autonomous’ Type 1 processes according to their origins, and describes four main groups (figure 2).

1. Processes that are hard-wired. These were naturally selected (in the Darwinian sense) in our evolutionary past for their adaptation value. Examples of such ‘innate’ heuristics that may induce biases are: the metaheuristics (anchoring and adjustment, representativeness, availability), search satisficing, overconfidence and others.

2. Processes that are regulated by our emotions. These too may be evolved adaptations (hard-wired) and are grouped into six major categories: happiness, sadness, fear, surprise, anger and disgust. Fear of snakes, for example, is universally present in all cultures. Or they may be socially constructed (acquired, learned), or combinations of the two—hard-wired modified by learning for example, visceral reactions against particular types of patients.

3. Processes that become firmly embedded in our cognitive and behavioural repertoires through overlearning. These might include explicit cultural and social habits, but also those associated with specific knowledge domains. An example of a bias acquired through repetitive exposures might be a ‘frequent flyer’ in a family doctor’s office or in the emergency department where the bias may be the expectation that no significant new diagnosis will be found.

4. Processes that have developed through implicit learning. It is well recognised that we learn in two fundamental ways. First, through deliberate explicit learning such as occurs in school and in formal training, and second, through implicit learning which is without intent or conscious awareness. Such learning plays an important role in our skills, perceptions, attitudes and overall behaviour. Implicit learning allows us to detect and appreciate incidental covariance and complex relationships between things in the environment without necessarily being able to articulate that understanding. Thus, some biases may be acquired unconsciously. Medical students and residents might subtly acquire particular biases by simply spending time in environments where others have these biases, even though the bias is never deliberately articulated or overtly expressed to them that is, the hidden curriculum. Examples might be the acquisition of biases towards age, socioeconomic status, gender, race, patients with psychiatric comorbidity, obesity and others.

Although Type 1 processes appear the most vulnerable to bias and suboptimal decision making, they are not the only source of impaired judgment. Cognitive error may also arise through biases that have become established through inferior strategies or imperfect decision rules. Arkes points out that error due to biases also occurs with Type 2 processes, that is, even though the decision maker may be deliberately

---

**Figure 2** Origins of biases in Type 1 processes. This is a modified section of the dual process model of diagnosis expanding upon the origins of Type 1 processes (based on Stanovich).
and analytically applying accepted strategies or rules, they may be flawed. Thus, there may have been a problem in the initial selection of a strategy which may then underestimate or overestimate a diagnosis. Of the two, it would seem preferable to always overestimate (as in the forcing strategy ‘rule out worst case scenario’) so that important diagnoses do not get missed; however this can sometimes be wasteful of resources. Generally, suboptimal strategies get selected when the stakes are not high.

**Situation-dependent biases:** An important question is: are there situations in which biases are more likely? Evidence suggests that certain conditions such as fatigue, sleep deprivation and cognitive overload, predispose decision makers to using Type 1 processes.\(^1\) In addition, specific clinical situations might increase vulnerability to specific biases. Some will set the physician up for exposure to particular biases whereas others will produce exposure to a wide range of biases. Some common situations are described in table 1.

**How does debiasing work?** While debiasing is an integral part of everyday living, some will do better than others. Those who are successful learn the consequences of their actions and take steps to avoid falling into the same thinking traps. Often this can be done using forcing strategies or deliberately suppressing impulsivity in certain situations. We can’t find our car keys at a time when we are in a hurry, so many of us learn the forcing strategy of always putting them in a specific place as soon as we arrive home. In some situations, we can adopt simple, protective forcing rules whenever we are going to do something irreversible for example, by following the maxim ‘measure twice, cut once’. In other domains, we have come to know that it is a good idea to suppress belief and be sceptical when we are offered deals that are too good to be true such as the email notifying us we have just won a large sum of money. Interestingly, increased intelligence does not protect against such follies.\(^{29}\)

Wilson and Brekke,\(^{35}\) in their extensive review, refer to cognitive bias as ‘mental contamination’ and debiasing as ‘mental correction’. They suggest an algorithmic approach, delineating a series of steps to avoid bias (figure 3). Bazerman sees the key to debiasing is first that some disequilibrium of the decision maker needs to occur so that the individual wants to move from a previously established response and change.\(^{36}\) This could come about by the individual simply being informed of a potential bias, or that their past judgment has raised the possibility they might be biased, or by developing insight into the adverse consequences of bias. This critical step may be more than simply becoming aware of the existence of biases and their causes; sometimes a vivid, perhaps emotion-laden experience needs to occur to precipitate cognitive change. The next step involves learning how the change will occur and what alternate strategies need to be learned. Finally, the last step occurs when the new approach is incorporated into the cognitive make-up of the decision maker and (with maintenance) becomes part of their regular thinking behaviour. An algorithmic approach has also been proposed by Stanovich and West,\(^{37}\) in which they further delineate characteristics of the decision maker needed to inhibit bias. Importantly, the decision maker must (1) be aware of the rules, procedures and strategies (mindware)\(^{38}\) needed to overcome the bias, (2) have the ability to detect the need for bias override, and (3) be cognitively capable of decoupling from the bias. Stanovich has examined the theoretical basis of

---

**Table 1** High-risk situations for biased reasoning

<table>
<thead>
<tr>
<th>High-risk situation</th>
<th>Potential biases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Was this patient handed off to me from a previous shift?</td>
<td>Diagnosis momentum, framing</td>
</tr>
<tr>
<td>2. Was the diagnosis suggested to me by the patient, nurse or another physician?</td>
<td>Premature closure, framing bias</td>
</tr>
<tr>
<td>3. Did I just accept the first diagnosis that came to mind?</td>
<td>Anchoring, availability, search satisficing, premature closure</td>
</tr>
<tr>
<td>4. Did I consider other organ systems besides the obvious one?</td>
<td>Anchoring, search satisficing, premature closure</td>
</tr>
<tr>
<td>5. Is this a patient I don’t like, or like too much, for some reason?</td>
<td>Affective bias</td>
</tr>
<tr>
<td>6. Have I been interrupted or distracted while evaluating this patient?</td>
<td>All biases</td>
</tr>
<tr>
<td>7. Am I feeling fatigued right now?</td>
<td>All biases</td>
</tr>
<tr>
<td>8. Did I sleep poorly last night?</td>
<td>All biases</td>
</tr>
<tr>
<td>9. Am I cognitively overloaded or overextended right now?</td>
<td>All biases</td>
</tr>
<tr>
<td>10. Am I stereotyping this patient?</td>
<td>Representational bias, affective bias, anchoring, fundamental attribution error, psych out error</td>
</tr>
<tr>
<td>11. Have I effectively ruled out must-not-miss diagnoses?</td>
<td>Overconfidence, anchoring, confirmation bias</td>
</tr>
</tbody>
</table>

Adapted from Graber\(^{34}\) General checklist for AHRQ project. A description of specific biases can be found in Croskerry.\(^7\)

---

**Figure 3** Successive steps in cognitive debiasing (adapted from Wilson and Brekke).\(^{35}\) Green arrows=yes; Red arrows=no
Debiasing in considerable depth. He proposes that a critical feature of debiasing is the ability to suppress automatic responses in the intuitive mode by decoupling from it. This is depicted in figure 1 as the executive override function. The decision maker must be able to use situational cues to detect the need to override the heuristic response and sustain the inhibition of the heuristic response while analysing alternative solutions, and must have knowledge of these alternative solutions. These solutions must of course have been learned and previously stored in memory as mindware. Debiasing involves having the appropriate knowledge of solutions and strategic rules to substitute for a heuristic response as well as the thinking dispositions that are able to trigger overrides of Type 1 (heuristic) processing.

Caveats are abundant in medical training and at its completion we are probably at our most cautious due to lack of experience and high levels of uncertainty. Experience subsequently accumulates but does not guarantee expertise. However, many clinicians will develop their own debiasing strategies to avoid the predictable pitfalls that they have experienced, or have learned through the experience of others. Morbidity and mortality rounds may be a good opportunity for such vicarious learning, provided that they are carefully and thoughtfully moderated. These rounds tend to inevitably remove the presented case from its context and to make it unduly salient in attendees’ minds, which may hinder rather than improve future judgment.

Although a general pessimism appears to prevail about the feasibility of cognitive debiasing, clearly people can change their minds and behaviours for the better. While evidence of debiasing in medicine is lacking, shaping and otherwise modifying our behaviours, extinguishing old habits, and developing new strategies and approaches are features of everyday life. Overall, we are faced with the continual challenge of debiasing our judgments throughout our careers. In our second paper, we review a number of general and specific strategies that have been grouped under the rubric of cognitive debiasing.

Contributors The article is based on a workshop conducted by the authors during the 2011 Diagnostic Error in Medicine Annual Conference in Chicago. PC conceived the work with substantial contributions from GS and SM. PC wrote the draft of the manuscript, and GS and SM revised it critically for important intellectual content. All authors approved the final version of the manuscript.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Open Access This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/3.0/
27 Croskerry P. Context is everything or How could I have been that stupid? *Healthcare Quarterly* 2009;12:167–73.