ARTICLE

Diagnostic error, overconfidence and self-knowledge

Quassim Cassam[1]

ABSTRACT According to the overconfidence hypothesis (OH), physician overconfidence is a major factor contributing to diagnostic error in medicine. This article argues that OH can be read as offering a personal, a sub-personal or a systemic explanation of diagnostic error. It is argued that personal level overconfidence is an "epistemic vice". The hypothesis that diagnostic errors due to overconfidence can be remedied by increasing physician self-knowledge is shown to be questionable. Some epistemic vices or cognitive biases, including overconfidence, are "stealthy" in the sense that they obstruct their own detection. Even if the barriers to self-knowledge can be overcome, some problematic traits are so deeply entrenched that even well-informed and motivated individuals might be unable to correct them. One such trait is overconfidence. Alternative approaches to "debiasing" are considered and it is argued that overconfidence is blameworthy only if it is understood as a personal level epistemic vice rather than a sub-personal cognitive bias. This paper is published as part of a collection on self-knowledge in and outside of illness.

[1] University of Warwick, Department of Philosophy, Coventry, UK Correspondence: (e-mail: q.cassam@warwick.ac.uk)

Much has been written about the extent and causes of diagnostic error in medicine. Multiple studies suggest that levels of diagnostic error remain "disappointingly high" (Berner and Graber, 2008: S3) despite advances in medical technology. One analysis concludes that "while the exact prevalence of diagnostic error remains unknown, data from autopsy series spanning several decades conservatively and consistently reveal error rates of 10%–15%" (Schiff et al., 2009: 1881). Since autopsies are rare in many countries other methods of tracking diagnostic error rates have had to be employed but the overall picture is the same: diagnostic error is relatively common (Kuhn, 2002; Graber, 2013; Singh et al., 2013).

Among the multifarious causes of diagnostic error one that has attracted the attention of researchers is overconfidence. According to the *overconfidence hypothesis* (OH) physician overconfidence "is a major factor contributing to diagnostic error" (Berner and Graber, 2008: S6). A natural suggestion is that if (OH) is correct then one way to reduce levels of diagnostic error would be to make physicians aware of the full extent of their tendencies to err and the role of overconfidence in causing diagnostic error. In effect, the proposal is that overconfidence and diagnostic errors due to overconfidence can be addressed by increasing physician self-knowledge or self-awareness (Borrell-Carió and Epstein, 2004; Croskerry et al., 2013b).

In order to assess this proposal a better understanding is required of overconfidence, (OH) and self-knowledge. Overconfidence can be conceived of in different ways and (OH)'s explanation of diagnostic error can also be interpreted in different ways depending on one's view of overconfidence. In brief, overconfidence can be viewed as a "sub-personal" *cognitive bias* or as a personal or professional *epistemic vice*. Either way, it is questionable to what extent diagnostic errors due to overconfidence can be remedied by self-knowledge. The issue, to put it crudely, is whether the biases or vices that are responsible for diagnostic error are also obstacles to self-knowledge.

A proper understanding of overconfidence is also the key to another difficult issue: how blameworthy are diagnostic errors due to overconfidence? In his work on human error Reason distinguishes between the "person approach" and the "system approach" (Reason, 2000). The former focuses on the individual origins of error while the latter attributes errors to their system context. The person approach blames individuals or groups of individuals for medical errors, including diagnostic errors, and holds them personally responsible. It might seem that (OH) subscribes to the person approach but only if overconfidence is an epistemic vice rather than a sub-personal cognitive bias. On either reading physician overconfidence may have a systemic explanation that complicates the issue of blameworthiness.

The first section of what follows will discuss (OH) and the notion of overconfidence. The focus here will be on understanding this notion and identifying the type of explanation of diagnostic error that (OH) offers. The following section will discuss the suggestion that self-knowledge is a remedy for physician overconfidence and the diagnostic errors to which it gives rise. While this suggestion has obvious attractions it will be seen to underestimate the obstacles to self-knowledge while overestimating its benefits. The final section will reflect on the issue of blameworthiness for diagnostic error, on the assumption that (OH) is justified in attributing to such errors to physician overconfidence. Alternative explanations of diagnostic error will be considered along the way.

The overconfidence hypothesis (OH)
Overconfidence has been characterized in terms of calibration. Subjects answer multiple-choice questions and are asked to state their subjective probability that their answers are correct. Studies suggest that when subjects report 100% confidence the relative frequency of correct answers is 80%; when subjects express 90% confidence they are right only 70% of the time (Lichtenstein et al., 1982). Overconfidence can be defined as "the difference between mean confidence and overall accuracy" (Kahneman and Tversky, 1996: 587), such that the former exceeds the latter. *Diagnostic* calibration is "the relationship between diagnostic accuracy and confidence in that accuracy" (Meyer et al., 2013: 1952). Overconfidence occurs when "the relationship between accuracy and confidence is miscalibrated or misaligned such that confidence is higher than it should be" (Meyer et al., 2013: 1953). In what sense is confidence higher than it "should be"? An example makes the point: "A patient who is informed by his surgeon that she is 99% confident in his complete recovery may be justifiably upset to learn that when the surgeon expressed that level of confidence she is actually correct only 75% of the time" (Kahneman and Tversky, 1996: 588).

"Overconfidence" can be used to refer to positive illusions or to "excessive certainty". The former is the tendency to have positive illusions about our merits relative to others. The latter "describes the tendency we have to believe that our knowledge is more certain that it really is" (Galloway, 2015: 16). The overconfidence that is at issue here is primarily the latter type. Overconfidence in this sense is related to arrogance and complacency (Berner and Graber, 2008: S6). The dictionary definition of "arrogant" is "excessively assertive or presumptuous; overbearing". Arrogance in this sense is not a necessary consequence of overconfidence. Similarly, a person can be overconfident without being complacent, that is, without being smugly self-satisfied. Nevertheless, overconfidence, arrogance and complacency are closely related in practice. Overconfidence can cause arrogance, and the reverse may also be true. Overconfidence and arrogance are in a symbiotic relationship even if they are distinct mental properties.

Diagnostic error is "any mistake or failure in the diagnostic process leading to a misdiagnosis, a missed diagnosis, or a delayed diagnosis" (Schiff et al., 2009: 1882). (OH) posits a causal relationship between diagnostic error and overconfidence: according to (OH) the latter is a major cause of the former. What is the causal pathway from overconfidence to diagnostic error? One suggestion is that overconfidence is related to a decreased likelihood of requesting additional diagnostic tests (Meyer et al., 2013: 1956). One study showed that in difficult cases, where there is a large reduction in diagnostic accuracy, physicians' confidence is only slightly lower than in less difficult cases with a lower likelihood of misdiagnosis. As a result of overconfidence, "physicians might not request the additional resources to facilitate diagnosis when they most need it" (Meyer et al., 2013: 1957). Diagnostic error is, in these circumstances, more likely. Other effects of overconfidence that may lead to diagnostic and other errors include widespread non-compliance with clinical guidelines and "the general tendency on the part of physicians to disregard, or fail to use, decision-support resources" (Berner and Graber, 2008: S7).

One way to summarize these causal pathways from overconfidence to diagnostic error would be think in terms of *insouciance*: physician overconfidence can lead to an attitude of insouciance when faced with difficult cases. This attitude causes behaviours (for example, failing to request additional diagnostic tests) that can in turn lead to diagnostic error. It is an empirical question whether this is the correct explanation of actual cases of diagnostic error but (OH) is a non-starter unless something along these lines is hypothesized. Causal explanations require the identification of causal pathways, and this means that it is not enough simply to point to a correlation between overconfidence and diagnostic error, even if such a correlation can be established.

It also needs to be explained *how* physician overconfidence is supposed to cause diagnostic error.

As noted above, Reason distinguishes two approaches to the explanation of error, a system and a person approach. In many cases of error, medical or otherwise, a more complex scheme is required, and it is helpful to distinguish at least four types of explanation:

(a) *Personal* explanations attribute error to the personal qualities of individuals or groups of individuals. Carelessness, gullibility, closed-mindedness, dogmatism, and prejudice and wishful thinking are examples of such qualities. These qualities are *epistemic vices* and explanations of error by reference to epistemic vices are *vice explanations* (Cassam, 2016).

(b) *Sub-personal* explanations attribute error to the automatic, involuntary, and non-conscious operation of hard-wired cognitive mechanisms. The distinction between personal and sub-personal explanations is due to Dennett. The personal level is the "explanatory level of people and their sensations and activities". The contrast is with the "level of brains and events in the nervous system" (Dennett, 2010: 105). In sub-personal explanations "the person, qua person, does not figure" (Elton, 2000: 2). These explanations are mechanistic in a way that personal explanations are not, and the mechanisms they posit are universal rather than person-specific.

(c) *Situational* explanations attribute error to contingent situational factors such as time pressure, distraction, overwork or fatigue. The aim of such explanations is to shift the explanatory focus away from the individuals towards the complex and demanding situations in which individuals find themselves.

(d) *Systemic* explanations attribute error to organizational or systemic factors such as lack of resources, poor training, or professional culture.

In terms of this taxonomy, which type of explanation of diagnostic error does (OH) offer? This question is of practical as well as theoretical importance since different interpretations of (OH) are likely to point to different practical measures for tackling diagnostic errors due to physician overconfidence.

The suggestion that (OH) proposes a situational explanation of diagnostic error can be dismissed without further discussion since overconfidence is not a "situational" factor in the relevant sense. The live options are that overconfidence is personal, sub-personal or systemic. Of these, the interpretation of (OH) as offering an explanation of diagnostic error in personal terms is arguably the most intuitive since overconfidence, arrogance and complacency have a strong *prima facie* claim to be classified as epistemic vices. Much depends, however, on how the notion of an epistemic vice is understood. What is the conception of epistemic vices on which (OH) can be regarded as proposing a vice explanation of diagnostic error?

Epistemic vices are personal qualities. These qualities constitute the building blocks of personal level explanations of a person's epistemic or other conduct. They include character traits, attitudes and ways of thinking. Epistemic vices are character traits, attitudes or ways of thinking that impede the acquisition, retention or transfer of knowledge. These are all ways in which epistemic vices "get in the way of knowledge" (Medina, 2013: 30). Closed-mindedness is an epistemically vicious character trait, prejudice an epistemically vicious attitude and wishful thinking an epistemically vicious way of thinking. Closed-minded is something that a person might be said to *be*, a prejudice is something that a person *has*, while wishful thinking is what a person *does*.

Epistemic vices can on occasion be conducive to knowledge, or at least to true belief. The proposal is that epistemic vices *normally* obstruct the acquisition, retention or transfer of knowledge, not that they invariably do so. Vices are harmful to those who have them, and perhaps to others, and it is in the nature of epistemic vices to be *epistemically* harmful. The *Concise Oxford Dictionary* defines "vice" as "evil or grossly immoral conduct" or, less dramatically, as a "defect or weakness of character or conduct". It is in the latter sense that epistemic vices are vices: they are defects of epistemic character or conduct. The description of closed-mindedness, wishful thinking and so on as epistemic *vices* implies that they are blameworthy, since "vice requires something for which we can be blamed" (Battaly, 2014: 62). Whatever the conditions for blameworthiness, epistemic vices must satisfy them.

The case for reading (OH) as proposing an epistemic vice explanation of diagnostic error is only as strong as the case for regarding overconfidence and the related qualities of arrogance and complacency as blameworthy personal qualities that are normally epistemically harmful. Overconfidence, arrogance and complacency are personal qualities in the sense that they are qualities of people rather than of their brains or nervous systems. It is *people* who are overconfident or arrogant or complacent, and explanations of their conduct by reference to these qualities are ones in which the people qua people figure. In addition, personal qualities are variable: they are not ones that every person has, or has to the same degree, and this is another respect in which overconfidence is personal. While some people are overconfident others are markedly underconfident. These qualities can either be viewed as character traits or as attitudes. Either way, they "express who we are as people" (Battaly, 2016: 100).

Is overconfidence epistemically harmful? Its role in causing diagnostic error is itself an illustration of the epistemic harm it does. The aim of a diagnosis is to determine the cause of a patient's symptoms. To determine the cause is to know the cause, and qualities that lead to diagnostic error are epistemically harmful to the extent that they obstruct knowledge of the cause. More generally, investigators who are excessively confident in their own judgments and abilities are less likely to seek additional evidence, even when such evidence is needed, or to enlist the help of others from whose expertise they might benefit. They are more likely to bring their inquiries to a premature conclusion and to take as settled questions that are not in fact settled. These are all ways in which overconfidence gets in the way of knowledge. If the claim that overconfidence is blameworthy can also be justified (see below) then the only reasonable conclusion is that (OH) is offering a personal level, vice explanation of diagnostic error.

This is not how all proponents of (OH) see things. Some argue that *bias* is "one of the principal factors underlying diagnostic error" (Croskerry *et al.*, 2013a: 1) and regard overconfidence as one such bias. Overconfidence bias has been defined as the "universal tendency to believe we know more than we do" (Croskerry, 2003a: 778). This proposal draws on dual-process theories of human psychology:

> These theories come in different forms, but all agree in positing two distinct processing mechanisms for a given task, which employ different procedures and may yield different, and sometimes conflicting, results. Typically, one of the processes is characterized as fast, effortless, automatic, nonconscious, inflexible, heavily contextualized, and undemanding of working memory, and the other as effortful, controlled, conscious, flexible, decontextualized, and demanding of working memory (Evans and Frankish, 2009: 1).

Some dual-process theories are also theories of mental architecture. They claim that "human central cognition is composed of two multi-purpose reasoning systems, usually called *System 1* and *System 2*, the operations of the former having

fast-process characteristics (….) and those of the latter slow-process ones" (Evans and Frankish, 2009: 1). Dual-process theories of this form are dual-system theories.

System 1 uses *heuristics*. A heuristic is a "simple procedure that helps find adequate, though often imperfect, answers to difficult questions" (Kahneman, 2011: 98). In the medical context heuristics have been described as "subconscious rules of thumb" that "clinicians use to solve diagnostic puzzles" (Berner and Graber, 2008: S8). Although heuristics are generally useful they "sometimes lead to severe and systematic errors" (Kahneman, 2011: 419). These errors are *biases*. In these terms, overconfidence can either be regarded as a bias in its own right or as caused by other more fundamental biases. Confirmation bias is the tendency to search selectively for evidence or information that confirms what one already believes. One proposal is that confirmation bias is the cause of overconfidence bias (Koriat *et al.*, 1980). Explanations of diagnostic error by reference to overconfidence bias are sub-personal to the extent that they attribute this bias to automatic, involuntary and unconscious System 1 processing. The explanation is mechanistic rather than personal and the posited explanatory bias is not a defect of epistemic character or conduct. It is a reflection of how the human cognitive apparatus works, and in all probability has its roots deep in our evolutionary history.

What is the relationship between sub-personal and personal explanations of diagnostic error? The recognition that there are two levels of explanation "gives birth to the burden of relating them" (Dennett, 2010: 107). Several views are possible. *Distinctivism* is the view that personal and sub-personal explanations are distinct and autonomous. Neither is reducible to the other and each is capable of explaining diagnostic error. This is a reflection of the fact that epistemic vices are distinct from sub-personal cognitive biases and not reducible to such biases. *Reductionism* is the view that personal explanations of diagnostic error are reducible to sub-personal explanations. On this view, the so-called "epistemic vices" that figure in personal explanations are *constituted* by sub-personal cognitive biases: to have a particular epistemic vice just is to have the corresponding sub-personal bias. A third approach is *eliminativism*. This holds that epistemic vices don't exist and talk of sub-personal biases should replace talk of epistemic vices.

There are difficulties with all three options. Distinctivism is problematic since it isn't credible that the personal and sub-personal levels are wholly autonomous. The sub-personal/personal distinction is closely related to, if not identical with, the System 1/ System 2 distinction, and System 2 is "heavily dependent on System 1" (Frankish, 2009: 97). Even if over-confidence is a personal-level epistemic vice it is highly likely to be influenced by sub-personal biases. Reductionism faces the difficulty that there are aspects of our epistemic conduct or character that cannot plausibly be explained in purely sub-personal terms, not least because these aspects are far from universal. They are not, in this sense, built into human cognitive apparatus. A case in point is arrogance, which is not a trait of all humans. There is no System 1 explanation of the fact that some people, including physicians, are overbearing, excessively assertive and presumptuous, while others are not. These are personal qualities that can neither be reduced to sub-personal biases nor eliminated if the aim is to explain the full variety of human epistemic conduct. Arrogance is not an illusion.

A way to reconcile the genuine insights of personal and sub-personal conceptions of overconfidence is to distinguish between overconfidence as a sub-personal cognitive bias and over-confidence as a personal epistemic vice. It is in the latter sense that overconfidence is closely related to arrogance and compla-cency since these are character traits rather than biases. This is not to say that biases or their underpinning heuristics can only be sub-personal. Overconfidence as an epistemic vice can be

regarded as a bias but overconfidence in this sense is a personal rather than a sub-personal bias. Biases and heuristics can be personal or sub-personal (Evans, 2009: 36; Kahneman, 2011: 120), and personal level overconfidence is the tendency to believe that one knows more than one does. This tendency is "prevalent but not universal" (Kahneman and Tversky, 1996: 587). It is caused in part by sub-personal biases but is neither identical with nor reducible to them. Personal and sub-personal overconfidence can both cause insouciance and thereby be responsible for diagnostic error.

On this pluralist account of overconfidence, (OH) can either be read as propounding a sub-personal or a personal explanation of diagnostic error. Both readings of (OH) are feasible and it needs to be made clear which reading is intended when diagnostic error is attributed to overconfidence in a given case. In either case, systemic factors may also come into play, but especially in relation to personal level overconfidence. There are several ways in which this might happen. The culture of medicine, which is a broadly systemic factor, might be thought to encourage over-confidence in physicians since "confidence is valued over uncertainty and there is a prevailing censure against disclosing uncertainty to patients" (Croskerry and Norman, 2008: S26). Particular specialisms might attract individuals who are highly confident or overconfident. It has been suggested, for example, that "a certain bravado goes with being a surgeon" and that "a surgeon has to have a high level of confidence to operate" (Groopman, 2008: 169). While patients "want their physicians to be highly competent and confident" (Kerr, 2007: 704), it is doubtful that they want them to be overconfident. Further research into what patients want is needed.

To the extent that overconfidence is an epistemic vice that is encouraged by the professional culture of medicine, it might be described as a "professional vice" (Greenhalgh, 2016). If there is evidence that overconfident individuals are more likely to be recruited into medical profession, and also more likely to be promoted, then diagnostic errors caused by overconfidence can be said to have a partly systemic explanation. However, it is possible to acknowledge the role of systemic factors in relation to diagnostic error while continuing to endorse (OH). The reason is not just that (OH) only identifies overconfidence as *a* major factor contributing to diagnostic error and so leaves room for other explanations. The reason is that physician overconfidence is itself, at least to some extent, a systemic factor.

In sum, personal, sub-personal and systemic factors all have a part to play in causing diagnostic error. These factors include overconfidence, which can be viewed as personal, as sub-personal or as systemic. The question whether, overconfidence is, as (OH) proposes, a *major* factor contributing to diagnostic error cannot be answered without further research and an examination of other causes of diagnostic error. Although such an examination cannot be undertaken here it should be remarked that among the other factors that may give rise to diagnostic and other medical errors is *underconfidence*, where mean confidence is lower than overall accuracy. Underconfidence has been described as a cognitive bias and can also be viewed as an epistemic vice. Whether underconfidence is more or less epistemically harmful than overconfidence cannot be known *a priori*. However, there is compelling evidence that underconfidence and overconfidence can co-exist. Reflection on this observation promises to cast further light on the nature of overconfidence and the appropriate measures for tackling it.

Self-knowledge
It is reasonable to assume that the role of overconfidence in causing diagnostic error is not widely recognized by physicians,

and that their knowledge of their own cognitive biases and epistemic vices is also limited. Physicians are, in this respect, no different from the wider population. It might be suggested that one measure to mitigate physician overconfidence and its adverse effects would be to fill in these knowledge gaps. Two types of knowledge are relevant. The first is working knowledge of the effect of cognitive biases and epistemic vices on clinical decision-making generally. The second is knowledge on the part of individual physicians of their own cognitive biases and epistemic vices. It isn't enough for them to know in the abstract that sub-personal biases and epistemic vices are causes of diagnostic error. They also need to accept that *they* have biases and vices that might be responsible for *their* mistakes. The first kind of knowledge will be referred to here as *working knowledge* and the second as *self-knowledge*.

Working knowledge and self-knowledge are only the start. A physician might have both types of knowledge but not be motivated to change their diagnostic practices or correct their own biases or vices. Self-knowledge doesn't automatically lead to self-improvement. A further relevant factor is the ability to change. The extent to which known unwanted biases can be avoided, controlled or corrected is unclear (Wilson and Brekke, 1994). In the case of entrenched biases a systemic response may be more appropriate, including what have been called "cognitive forcing strategies" (Croskerry, 2003b: 114). Properly understood, these strategies take bias-control out of the hands of individuals. Examples of such strategies will be given below, following consideration of the limitations of self-knowledge as a tool for mitigating physician overconfidence.

Some epistemic vices are "stealthier" than others: by their nature they evade detection by those who have them (Cassam, 2015). They are not undetectable but are harder to detect in oneself than other vices. The stealthiness of some epistemic vices can be accounted for as follows: detection of one's own epistemic vices requires a process of active critical reflection (Fricker, 2007: 97). Stealthy vices are ones that inhibit such reflection or otherwise reduce its effectiveness. All epistemic vices may be expected to some negative impact on a thinker's active critical reflection but some more than others. Effective critical reflection on one's own epistemic and other vices requires the possession and exercise of epistemic *virtues* such as open-mindedness, sensitivity to evidence, and the humility to acknowledge one's vices. One possibility is that the necessary virtues co-exist with the vices that critical reflection aims to uncover. For example, carelessness is an epistemic vice that can co-exist with epistemic virtues such as open-mindedness, sensitivity to evidence and humility. It is not in the nature of carelessness to evade detection by active critical reflection since being careless need not deprive one of the ability to detect one's carelessness. Other factors, such as a desire to think well of oneself, might get in the way of knowledge of one's own carelessness but this is not a case of an epistemic vice impeding its own detection.

In contrast, closed-mindedness is antithetical to the epistemic virtues required for the discovery of one's epistemic vices by reflection. If a person is closed-minded then their mind may well be closed to the possibility that they are closed-minded. This is not just a reflection of the general desire to think well of oneself but of the role of closed-mindedness in directly negating one of the key epistemic virtues required for active critical reflection, the virtue of open-mindedness. Complacency is another stealthy vice. Being complacent is hardly conducive to active critical reflection on one's complacency. This suggests a spectrum of epistemic vices, with less stealthy epistemic vices at one end and more stealthy vices at the other. In these terms, there is a case for locating overconfidence and complacency at the stealthier end of the spectrum, on the basis that they are closely associated with

lack of the epistemic humility that is required for active and honest critical reflection on one's epistemic defects.

Knowledge of one's sub-personal cognitive biases is, if anything, even harder to attain than knowledge of one's personal epistemic vices. Sub-personal biases are not introspectable and unlikely to be detected by unaided critical reflection, however earnest. Their nature, existence and influence has been brought to light by the work of cognitive psychologists, and a person with working knowledge of this research might infer that they are not immune to cognitive biases. Working knowledge of cognitive psychology cannot be assumed, however, and overconfidence might induce physicians who are aware of the potential impact of biases on clinical judgement to believe that they are not vulnerable to them (Croskerry *et al.*, 2013b: ii 66). This is a reflection of the fact that "the same kinds of biases that distort our thinking in general also distort our thinking about the biases themselves" (Horton, 2004).

Despite the obstacles to self-knowledge, knowledge of one's own cognitive biases and epistemic vices is not ruled out. Epistemic vices may be domain-specific and the domains in which a particular epistemic vice makes itself felt need not include active critical reflection on one's own vices and biases. In addition, active critical reflection on one's "cognitive sins" (Adams, 1985: 17) is not the only source of self-knowledge. Feedback is another potential source, although even this route to self-knowledge presupposes a willingness to listen that some may lack. An alternative route to self-knowledge is the occurrence of a single traumatic event, such as the unexpected death of a patient, that fundamentally changes a physician's self-conception and causes them to reflect on their epistemic vices and biases in ways that would not otherwise have been possible (Croskerry *et al.*, 2013b: ii 66). Such "transformative experiences" (Paul, 2014) may motivate change and self-improvement as well as lead to self-knowledge. It is when things go wrong that hitherto over-confident and complacent clinicians may be motivated to reflect on their defects and change their diagnostic practices. Flawed practices and procedures may need to be "unlearned" (Rushmer and Davies, 2004), and transformative experiences may bring about "transformative unlearning" (Macdonald, 2002).

As noted above, being motivated to correct a bias one knows one has is no guarantee that one will be able to correct it. Some biases may be so deeply entrenched that even well-informed and motivated individuals might be unable to shake them off. Overconfidence is a case in point. This has been defined as the discrepancy between single-event confidences and the actual relative frequency of correct answers, such that the former exceeds the latter. For example, if a subject reports 100% confidence in a particular answer but the actual frequency of correct answers when they report this level of confidence is 75% then they are described as "overconfident". However, when the same subjects are asked several questions and invited to estimate the *frequency* of correct answers their estimated frequencies are "practically identical with actual frequencies, with even a small tendency towards underestimation" (Gigerenzer, 1991: 6). At this level there is no "overconfidence bias". Nevertheless, subjects can "maintain a high degree of confidence in the validity of specific answers even when they know that their overall hit rate is low" (Kahneman and Tversky, 1996: 588).

This phenomenon has been described as the "illusion of validity". In one well known example of this illusion, candidates for officer training in the Israeli army were observed and evaluated on the basis of performance in a group leadership exercise. Evaluators who made confident judgements regarding the suitability of individual candidates for officer training were told in later feedback sessions that their ability to predict

performance at officer training school was negligible. One of the evaluators commented:

> What happened was remarkable. The global evidence of our previous failure should have shaken our confidence in our judgments of the candidates, but it did not. It should also have caused us to moderate our predictions, but it did not. We knew as a general fact that our predictions were little better than random guesses, but we continued to feel and act as if each of our specific predictions was valid (Kahneman, 2011: 211).

Kahneman compares the illusion of validity with the Müller-Lyer illusion in which two lines that are known to be equal in length are seen as unequal. Neither knowing that the lines are equal nor being motivated to see them as equal is sufficient to bring it about that they are seen as equal. By the same token, in cases such as those described by Kahneman, neither knowing that one's track record is poor nor being motivated to do better may be sufficient to shake one's unwarranted confidence in one's judgement in an individual case. In such cases, other means of reigning in overconfidence will need to be found.

In his work on "cognitive debiasing", Croskerry proposes that clinicians can develop "cognitive forcing strategies" to "abort" the latent biases that cause diagnostic error (2003b: 110). A cognitive forcing strategy is "a deliberate, conscious selection of a particular strategy in a specific situation to optimize decisonmaking and avoid error" (Croskerry, 2003b: 115). Cognitive forcing is essentially a form of self-regulation: physicians are urged to self-monitor for bias and follow rules that are designed to counter cognitive pitfalls. Yet in the case of stealthy and entrenched biases this is unlikely to be sufficient since self-monitoring and self-regulation are not immune to the cognitive pitfalls and illusions they are designed to counter. There is, however, another way of understanding the notion of a cognitive forcing strategy on which this approach offers greater promise.

The alternative reading of cognitive forcing draws on the notion of a forcing function. Forcing functions are "built into system design" with a view to minimizing error (Croskerry, 2003b: 115). An example is the design of vehicle locking systems to prevent car drivers from locking themselves out of their cars: it is not possible to lock the car door as long as the key has not been removed from the ignition. Similarly:

> A person might engineer her "external" epistemic environment in other ways to ensure that her intentions and values are more fluidly expressed in her actions and judgements, and not distorted by the operation of implicit biases. For instance, if one is (justifiably) worried about implicit biases corrupting the assessment of candidates in a job search, one can take measures to remove information from application dossiers that may trigger those implicit biases in the first place (Holroyd and Kelly, 2016: 121–2).

What is described here is a programme of bias-control rather than debiasing. The primary aim is not to correct implicit biases but to mitigate their consequences by measures designed to prevent them from taking effect. The removal of certain types of identifying information from application dossiers is a forcing function, albeit one engineered by the subject herself. In this scenario, knowledge of the harmful effects of cognitive biases and a favourable disposition towards bias control are both presupposed. However, forcing functions need not be self-imposed and can be built into system design by external agencies.

The main attraction of forcing functions is that they take bias-control out of the hands of bias-prone individuals who may not see the need for special measures to tackle their biases or who may lack the motivation to self-monitor and follow rules designed to avoid cognitive pitfalls. Forcing functions will be especially appropriate in relation to biases that even well-informed and epistemically conscientious individuals find it difficult to shake off. The precise design of such forcing functions in clinical contexts cannot be discussed here but the proposal can be succinctly stated: to the extent that overconfidence is a major cause of diagnostic error steps need to be taken either to reduce physician overconfidence or lessen its harmful effects. While working knowledge and self-knowledge might prove helpful in both regards, there are good reasons to be sceptical about their efficacy in bringing about debiasing or epistemic vice reduction. That being the case, forcing functions represent a promising alternative.

Blameworthiness

To what extent are diagnostic errors caused by physician overconfidence or other epistemic vices or biases blameworthy? The issues here are too complex to be discussed in depth but it would nevertheless be appropriate to note the following: on a "voluntarist" conception of moral responsibility and blame-worthiness we are blameworthy only for what is within our voluntary control. It might seem, therefore, that cognitive biases and epistemic vices cannot be deemed blameworthy since they are not within our voluntary control. Where an individual physician's overconfidence leads to diagnostic errors with serious consequences the physician is causally responsible for those errors and their consequences but not blameworthy in a deeper sense unless their overconfidence is within their voluntary control.

One response to this line of argument is to suggest that at least some of a person's cognitive biases and epistemic vices are, in the relevant sense, within their voluntary control. The control might be indirect rather than direct but sufficient for blameworthiness. Another view is that voluntary control, direct or indirect, is not necessary for blameworthiness. This is the view to be defended here. It has been noted that there many blameworthy states of mind that are involuntary. Such objectionable states of mind include hatred, jealousy and contempt for other people (Adams, 1985: 4). Other "involuntary sins" are *cognitive* failures:

> Some examples are: believing that certain people do not have rights that they do in fact have; perceiving other members of some social group as less capable than they actually are; failing to notice indications of other people's feelings; and holding too high an opinion of one's own attainments. These failures are not in general voluntary (Adams, 1985: 18).

In that case, what makes such cognitive failures blameworthy? According to the "rational relations view" (Smith, 2008: 382), only features of a person that reflect her rational activity, that is, her judgments or evaluative assessments are genuinely blameworthy. A state of mind is blameworthy, on this account, only if its subject can be asked to justify it and to reassess it if an adequate justification cannot be provided. This explains, for example, why selfishness is blameworthy whereas stupidity is not.

This account implies that sub-personal cognitive biases may not be blameworthy in so far as they do not reflect a person's judgements or evaluative assessments. Personal level epistemic vices, in contrast, can be regarded as genuinely blameworthy

regardless of whether they are within our voluntary control. For example, the arrogant person "has a high opinion of himself" and treats others with disdain because he believes "he is a better person according to the general standards governing what counts as a successful human specimen" (Tiberius and Walker 1998: 382). On this view, arrogance plainly reflects a person's evaluative assessment of himself; it *is* an evaluative assessment of himself and others. As such, a person can be asked to justify his self-assessment and his disdain. Even if he has talents that few others have this does not make him a better person or justify his low opinion of those who lack his talents.

As noted above, overconfidence can either be regarded as a sub-personal cognitive bias or as a personal level epistemic vice. Overconfidence in the first sense may not be blameworthy but overconfidence in the second sense is a different matter. It is reflective of a person's evaluative assessment of how much they know and, in this sense, reflective of their rational activity. They can be challenged to justify their evaluation and "to acknowledge fault if an adequate defense cannot be provided" (Smith, 2008: 370). Regardless of whether overconfidence is a professional epistemic vice, it is no less blameworthy than the arrogance to which it is closely related. Both arrogance and overconfidence might be viewed as objectionable in themselves or an objectionable in virtue of their harmful effects. In neither case does the fact —if it is a fact—that these states of mind are involuntary imply that they are not blameworthy.

From the fact that a person is blame*worthy* it does not follow that it is appropriate to blame them. It has been argued that naming, blaming and shaming healthcare professionals for diagnostic and other medical errors is largely ineffective and counterproductive if the priority is to improve patient safety (Reason, 2013: 101). From this perspective the causes of medical error are largely systemic rather than personal and the appropriate response to such errors is not to blame individuals but to construct adequate system defences. However, physician overconfidence is an example of a cause of error that is both systemic *and* personal. To the extent that physician over-confidence is a fact of medical life a systems response that incorporates forcing functions might indeed be the most appropriate response. Nevertheless, regardless of whether blaming improves patient safety, it needs to be recognized that overconfidence and arrogance are objectionable traits in a physician, and that it is not unreasonable to regard them as, at least to some extent, blameworthy.

References

Adams R (1985) Involuntary sins. *The Philosophical Review*; **94** (1): 3–31.

Battaly H (2014) Varieties of epistemic vice. In: Matheson J and Vitz R (eds). *The Ethics of Belief: Individual and Social*. Oxford University Press: Oxford, pp 51–76.

Battaly H (2016) Epistemic virtue and vice: Reliabilism, responsibilism, and personalism. In: Mi C, Slote M and Sosa E (eds). *Moral and Intellectual Virtues in Western and Chinese Philosophy*. Routledge: New York, London, pp 99–120.

Berner E and Graber M (2008) Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*; **121** (Suppl. 5): S2–S23.

Borrell-Carió F and Epstein R (2004) Preventing errors in clinical practice: A call for self-awareness. *Annals of Family Medicine*; **2** (2): 310–316.

Cassam Q (2015) Stealthy vices. *Social Epistemology Review and Reply Collective*; **4** (10): 19–25, [Open access].

Cassam Q (2016) Vice epistemology. *The Monist*; **99** (2): 159–180, [Open access].

Croskerry P (2003a) The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine*; **78** (8): 775–780.

Croskerry P (2003b) Cognitive forcing strategies in clinical decisionmaking. *Annals of Emergency Medicine*; **41** (1): 110–120.

Croskerry P and Norman G (2008) Overconfidence in clinical decision making. *The American Journal of Medicine*; **121** (Suppl. 5): S24–S29.

Croskerry P, Singhal G and Mamede S (2013a) Cognitive Debiasing 1: Origins of bias and theory of debiasing. *BMJ Quality & Safety*; **22** (Suppl. 2): ii58–ii64.

Croskerry P, Singhal G and Mamede S (2013b) Cognitive debiasing 2: Impediments to and strategies for change. *BMJ Quality & Safety*; **22** (Suppl. 2): ii65–ii72.

Dennett D (2010) *Content and Consciousness*. Routledge Classics: Abingdon, UK.

Elton M (2000) The personal/ sub-personal distinction: An introduction. *Philosophical Explorations*; **3** (1): 2–5.

Evans J (2009) How many dual process theories do we need? One, two, or many?. In: Evans J and Frankish K (eds). *In Two Minds: Dual Processes and Beyond*. Oxford University Press: Oxford, pp 33–54.

Evans J and Frankish K (2009) *In Two Minds: Dual Processes and Beyond*. Oxford University Press: Oxford.

Frankish K (2009) Systems and levels: Dual system theories and the personal-subpersonal distinction. In: Evans J and Frankish K (eds). *In Two Minds: Dual Processes and Beyond*. Oxford University Press: Oxford, pp 89–107.

Fricker M (2007) *Epistemic Injustice*. Oxford University Press: Oxford.

Galloway M (2015) Challenging diagnostic overconfidence. *Excellence in Medical Education*; **6**, 16–19.

Gigerenzer G (1991) How to make cognitive illusions disappear: Beyond "heuristics and biases". In: Stroebe W and Hewstone M (eds). *European Review of Social Psychology*. John Wiley & Sons: Chichester, UK, pp 83–115.

Graber M (2013) The incidence of diagnostic error in medicine. *BMJ Quality & Safety*; **22** (Suppl. 2): ii21–ii27.

Greenhalgh T (2016) 'Virtues and Vices in Evidence Based Clinical Practice', http://www.cebm.net/5395-2/#.Vq6l1WsGI_s.twitter.

Groopman J (2008) *How Doctors Think*. Mariner Books: Boston, MA.

Holroyd J, Kelly D (2016) Implicit bias, character, and control. In: Masala A and Webber J (eds). *From Personality to Virtue*. Oxford University Press: Oxford, pp 106–133.

Horton K (2004) Aid and bias. *Inquiry*; **47** (6): 545–561.

Kahneman D (2011) *Thinking, Fast and Slow*. Allen Lane: London.

Kahneman D and Tversky A (1996) On the reality of cognitive illusions. *Psychological Review*; **103** (3): 582–591.

Kerr J (2007) Confidence and humility: Our challenge to develop both during residency. *Canadian Family Physician*; **53** (4): 704–705.

Koriat A, Lichtenstein S and Fischoff B (1980) Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*; **6** (2): 107–118.

Kuhn G (2002) Diagnostic errors. *Academic Emergency Medicine*; **9** (7): 740–750.

Lichtenstein S, Fischoff B and Phillips L (1982) Calibration of probabilities: The state of the art to 1980. In: Kahneman D, Slovic P, Tversky A (eds). *Judgment Under Uncertainty: Heuristics & Biases*. Cambridge University Press: Cambridge, UK, pp 306–334.

Macdonald G (2002) Transformative unlearning: Safety, discernment and communities of learning. *Nursing Inquiry*; **9** (3): 170–178.

Medina J (2013) *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations*. Oxford University Press: Oxford.

Meyer A, Payne V, Meeks D, Rao R and Singh H (2013) Physicians' diagnostic accuracy, confidence, and resource requests: A vignette study. *JAMA Internal Medicine*; **173** (21): 1952–1958.

Paul LA (2014) *Transformative Experience*. Oxford University Press: Oxford.

Reason J (2000) Human error: Models and management. *BMJ*; **320** (7237): 768–770.

Reason J (2013) *A Life in Error: From Little Slips to Big Disasters*. Ashgate: Farnham.

Rushmer R and Davies H (2004) Unlearning in healthcare. *Quality and Safety in Healthcare*; **13** (Suppl II): ii10–ii15.

Schiff G et al (2009) Diagnostic error in medicine. *Archives of Internal Medicine*; **169** (20): 1881–1887.

Singh H, Giardina T, Meyer A, Forjuoh S, reis M and Thomas E (2013) Types and origins of diagnostic errors in primary care settings. *JAMA Internal Medicine*; **173** (6): 418–425.

Smith A (2008) Control, responsibility, and moral assessment. *Philosophical Studies*; **138** (3): 367–392.

Tiberius V and Walker J Arrogance (1998) *American Philosophical Quarterly*; **35** (4): 379–390.

Wilson T and Brekke N (1994) Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*; **116** (1): 117–142.

Data availability

Data sharing is not applicable to this article as no datasets were generated or analysed.

Additional information