Jennifer Nagel, University of Toronto
Forthcoming in *Proceedings of the Aristotelian Society, Supplementary Volume*
February 6, 2014

Intuition, reflection, and the command of knowledge

Creatures like us can have a hard time doing the right thing. Not only are we conscious of struggling with untoward desires and conflicts in our beliefs, but we also face the silent menace of factors lying below the threshold of consciousness. According to Tamar Gendler, there is a deep distinction here: in these two ways of being pulled off the right path, we are wrestling with different animals. Looking closely at the struggle against what lies below the threshold of consciousness, Gendler contends that we can extract an important lesson about the limitations of knowledge. This lesson is supposed to be that knowledge does not have the power claimed for it by Socrates in the *Protagoras*, a 'commanding' power to bar us from wrong action. The subterranean realm of associative processes diverts us from our reflective commitments in ways that are not directly visible to us, and the failure of visibility here is taken to limit the rule of knowledge over action.

I share with Gendler the conviction that there is something epistemologically interesting about the line between what is consciously accessible and inaccessible; however, I am not convinced that the power of knowledge is restricted to the sunny side of this line. Indeed, I think that there is a version of the Socratic view of the commanding power of knowledge that remains defensible, even with respect to what is not apparent to consciousness. In my view, the best way to make sense of the problematic cases that Gendler highlights is not to see them as situations where the knowledge we have is powerless, but as situations in which some relevant knowledge is missing. The first section

re-examines the two Platonic pictures that Gendler sketches, and questions her reasons for rejecting each of them. The second section presents my view of the significance of conscious thought, in the context of a broader account of the contrast between intuitive and reflective thinking. The last section addresses the relationship between knowledge and self-regulation.

Section 1: The Platonic pictures

Gendler launches her discussion by contrasting two lines of thought advanced by Plato: according to the first, knowledge is sufficiently powerful to bar any contrary action, and according to the second, the rational part of the soul must struggle against the spirited and appetitive parts, and faces the live possibility of losing the fight. Gendler finds the second picture more appealing than the first, but goes on to argue for a view in which the soul or mind is further subdivided, a view in which rational control of action involves wrestling not only with consciously felt spirit and appetite, but also with a further force, the 'third horse' of unconscious implicit association. She argues that this last struggle is different in kind from the others, and suggests that it poses a deeper challenge to the original Platonic picture of knowledge as the commanding element in the soul. Before examining the distinctive significance of the implicit, it will be helpful to take a second look at the initial motivating framework drawn from Plato.

When Plato depicts the rational part of the soul as struggling and possibly failing to rule, does it follow that he is rejecting his earlier characterization of knowledge as commanding? The two images have different focal points: the claim that knowledge

commands does not speak directly to the relation among parts of the soul, and the model of the divided soul doesn't directly offer a rival picture of knowledge. Even if Plato holds that the rational part of the soul is the natural home for knowledge, the claim that knowledge 'cannot be overcome' is silent on whether parts of the soul other than the rational part might sometimes rule: we can distinguish the part of the soul that houses knowledge from the knowledge it contains. According to the picture sketched in the *Protagoras,* when knowledge is present, it must rule; however, if the rational part of the soul is overpowered by emotion or appetite, this outcome need not indicate any weakness in knowledge as such, as long as the rational part of the soul could lose a battle simply by lacking some relevant knowledge. To retain the view that knowledge is commanding, I could for example hold that if my knowledge of the exact health costs of the chocolate cake were rich and detailed enough, and my knowledge of the value of good health were deep and clear enough, then it would indeed become impossible for me to act on a desire to eat the cake. Or if cake and health-related knowledge is not sufficient to keep me on the right path, perhaps further knowledge about other things, including full knowledge of my own psychological vulnerabilities, would be enough to do the trick. Perhaps the cake could still end up in my mouth as the result of physical compulsion or a seizure, but if we restrict our evaluations to anything that truly deserves the name of action, it would take some argument to show that full knowledge is less than commanding.

Gendler's effort to advance such an argument starts from the observation that we can imagine stronger and weaker versions of the thesis that knowledge is commanding. On the strongest version, knowledge of the right course of action suffices both to produce that action and to extinguish any contrary inclinations. Knowledge on this view commands not

only action but appearance. This view is not unanimously agreed to be implausible—one form of it is defended by John McDowell (1979), for example—but it is stronger than the position I will defend here. A more moderate version of the Socratic view allows that contrary inclinations could be felt, while insisting on the power of knowledge over action: knowledge will 'not allow a man, if he only knows the difference of good and evil, to *do* anything which is contrary to knowledge' (my emphasis). Gendler rejects this line with the comment: 'there are few things of which I am more certain than that realizing that some course of action is the one I *ought to* carry out is compatible with my actually *failing* to carry it out, even if I focus clearly and explicitly on the gap between what I ought to be doing, and what I am actually engaged in doing' (p.5). Failures matching this description are surely possible, but it is controversial whether this description actually guarantees a counterexample to the moderate Socratic view. There are various ways in which we could find ourselves failing to carry out a course of action recognized as what we ought to do, and not all of them would constitute a reason to reject the view of knowledge as commanding. To begin, there are some simple cases satisfying the description that Gendler herself would doubtless grant are unproblematic for the Socratic view. For example, if the failure to do what one knows one ought to do arises not from a poor judgment but from a break in the connection between judgment and action—I know I ought to speak up, but a sudden throat spasm makes it impossible to talk—then we do not have an instance of someone acting contrary to knowledge. We have a mere incapacity to act, or to act in the right way (I might still find myself doing something, opening my mouth in some futile way, but not the thing I know to be best). Equally, if my behaviour is produced by a blind compulsion—an outburst driven by Tourette's Syndrome or a movement governed by uncontrolled addiction—then

it is a questionable instance of action. To count as an uncontroversial instance of action, my behaviour must be the product of choice on my part, not simply something I find myself doing or failing to do for reasons utterly beyond my control as an agent.

What more is needed to count an action as a clear counterexample to the Socratic view? An important feature of the Socratic view is the very particular kind of knowledge involved: it is the agent who 'knows the difference of good and evil' who is unable to act contrary to knowledge, or as Socrates later elaborates the view, unable to do wrong knowingly. Here one might think that 'realizing that some course of action is the one I *ought to* carry out' will suffice to meet that requirement. But I think the most plausible versions of the Socratic view take 'knowing the difference of good and evil' to mean something quite strong: it is not enough to register that there is some difference between these terms in general, or that some perhaps vaguely articulated course of action is in some way good. 'Knowing the difference of good and evil' must apply to the particular options at hand, and under the relevant mode of presentation. If Oedipus knows that it is wrong to kill one's father, but lacks relevant knowledge about the stranger he meets on the road, then he would fail to know the difference of good and evil in this sense. Oedipus does not do wrong knowingly. For a clear counter-example to the Socratic view, I must know, of a specific course of action available to me, that it is good, and simultaneously make a contrary choice, contrary in the sense of being a choice to do what I know to be less good. There is something strange about choosing what one knows to be worse over what one knows to be better—indeed it is strange enough that one might argue that contrary choice would be a sign that the difference between good and evil was not properly known by the agent in this case. Perhaps the cases of weakness of the will that Gendler describes are better

understood as involving a conflict within the agent's evaluations of actions, and as the kind of conflict that would undercut the agent's claim to knowledge of the good in the case at hand (see Tenenbaum, Sergio 1999 for a guide to this strategy). Or perhaps the contrary choice is a sign that something is faulty in the agent's knowledge of the relationship between the general course of action that she knows to be good and the specific circumstances in which she finds herself. For example, if I deceive myself about my reasons for action ('it's just this once, and because I am feeling sad!') I could conceive of this one particular act of eating cake as compatible with my general recognition of the merits of a healthy diet. But if I am self-deceived about the character of the particular choice that now confronts me, it is not clear that I can count as retaining knowledge of the good in the moment of action. It remains possible that the akratic agent could *say* or *think*, of one course of action, that it was the one she ought to pursue, while doing otherwise;[1] later, she could come to realize or know that the road not taken was the right one, but it is not trivial to show that at the moment of contrary choice the agent really did have the appropriate knowledge of the good. Even if we allow that it is psychologically possible to have contradictory beliefs on a given point, it is hard to see how someone could count as *knowing* that *p* while making a contrary judgment under the same mode of presentation.

Not everyone will accept the claim that intentional action is invariably based on judgment. Gendler herself sees cases of weakness of the will as a reason to applaud (and then improve upon) Plato's model of the tripartite soul, where action is commanded by warring faculties of reason, spirit and appetite, and where in her reading the latter faculties are not seen as judgmental. In her view, spirit and appetite produce inclinations which can

[1] Plato himself seems to have been unwilling to make even this concession: see *Protagoras* 358b-d.

guide action despite their not being endorsed by the agent: the akratic agent is not making a problematic endorsement of a course of action suggested by appetite, but is acting in a way which is independent of what she has endorsed and of what she knows.[2] Appetite without judgment directly launches us on courses of action 'even in the face of knowledge of how the world actually is and which course of action is the correct one to pursue' (5).

Gendler embraces the idea that we can act without endorsement, but there is something about this tripartite soul model that she still considers unsatisfactory: when we are driven by factors other than endorsement, she says, the model assumes that we will feel the force of those factors, and perhaps 'a distinctive phenomenology' of inner conflict as well—although this is characterized as something that is 'often associated' with conflict between knowledge and action, rather than being a criterion or invariable accompaniment of it (p.5). However, Gendler contends, 'there are also cases where such conflicts occur with no associated phenomenology' (p.7): these are cases in which we have encoded associations that 'end up guiding our attention and interpretation in ways we fail to even notice'. In cases where we would 'normatively disavow' these associations, they nonetheless 'redirect our evaluative and behavioural tendencies in ways that run contrary to our reasoned commitments'. This kind of conflict within the self is unlike the traditional conflicts between reason and spirit or appetite, Gendler maintains, because 'it pulls and points the chariot without the charioteer or the other horses even realizing there has been

[2] In characterizing Gendler as holding that appetite does not endorse, I am guided by her claim that familiar cases of succumbing to weakness of will are 'cases where the non-endorsed inclination guides actions' (p.5) These inclinations may however count as 'endorsed' in some weaker sense: on p.2 she describes weakness of will cases as involving 'a cognitive conflict between our reflective attitudes and our non-reflective endorsements'. It may be that she intends 'endorsement' in the fullest sense only to apply to reflective endorsement; I suggest some reasons to reject that view in Section 2 of this paper. Alternatively, if acting on appetite really does involve some endorsement, then it becomes even easier to argue that claims to reflective knowledge are undercut when we act on contrary appetite.

a process of redirection' (p.7). If rational deliberation ordinarily operates upon contents available to consciousness, then the influence of these implicit associations lies outside the reach of rational deliberation, threatening to hijack our behaviour away from our reflective commitments. Gendler concludes that associations demonstrate a limit to the powers of knowledge, and that they call for self-regulation of a kind that is different from the self-regulation we apply in cases of consciously felt desire.

The relationships between reason, consciousness, and knowledge are intriguing, but somewhat complex. Gendler evidently takes reason to be the faculty within us that manages conscious deliberation; this characterization departs from Plato's, but certainly corresponds to something psychologically real. If reason is the power to deliberate consciously, however, it should be noted that not everything produced by reason has the status of knowledge: sometimes our explicit reasoning starts from misleading evidence or ends up drifting into some fallacy. Alarm bells are not guaranteed to ring in these circumstances, but the possibility of erring without noticing is not a sign that some force other than reason—understood just as the power for conscious deliberation—is determining our judgment. It should therefore be possible to act on the basis of reason, and act badly, without thereby impugning the Socratic model. When we act badly on the basis of deliberation that has reached something less than knowledge—say, a false judgment produced as the result of our having reasoned fallaciously—this is not a sign that knowledge is unable to guide us to do the right thing. On the other side of the equation, if we allow that not every instance of knowledge is produced by conscious deliberation, then we cannot assume that circumventing or opposing conscious deliberation is always problematic. Exactly because conscious deliberation does not necessarily lead to

knowledge, there are circumstances in which failure to be guided by this type of deliberation can be a good thing. For example, consider the man who has homosexual desires but has convinced himself through bad reasoning that these desires are wrong and must be overcome through therapy. Such a person might do well to act on the basis of desire or intuition rather than 'reflective commitment', and it's not obvious that in doing so he'd be acting contrary to knowledge. The conflict between his reflective commitments and his desires might entail that even his unreflective acts could not count as performed on the basis of knowledge, but there is a distinction between acting in the absence of knowledge and acting contrary to it. Meanwhile, if well-founded intuitive judgments can constitute knowledge, one might act knowledgeably on the basis of intuition, either in acting in accordance with one's reflective position, or in the absence of a reflective commitment, for example on a question on which one had never engaged in conscious deliberation.

What special significance, if any, attaches to reflective commitments? Consciousness is widely agreed to have a key role in demarcating reflection from intuition, but there is controversy over exactly what that role might be. Gendler emphasizes the extent to which implicitly encoded associations may be hidden from the agent, but it is not generally true that intuitive cognition has hidden products. I think the decisive contrast relevant to consciousness has to do with the processing rather than the products; the next section sets out my view of this contrast.

Section 2: Intuition and reflection

In presenting scientific accounts of intuitive judgment, psychologists are not introducing a neologism: what they aim to explain is the psychological reality underlying a category recognized for a long time, if only roughly, in ordinary language. Our pre-scientific grasp of this category is enabled by the role played by consciousness in its demarcation. The Oxford English Dictionary marks the relevant sense of the word 'intuitive' as dating back to 1645: 'Of knowledge or mental perception: That consists in immediate apprehension, without the intervention of any reasoning process.' John Locke (1689) picks up roughly this sense in contrasting intuitive with demonstrative knowledge, where demonstrative knowledge is produced through a series of consciously accessible stages. We judge at once, in an 'irresistible' manner, that a triangle is not a circle, but we need to go through a series of stages (each individually conscious and intuitive) in order to reach the judgment that the internal angles of a triangle sum to two right angles (Locke, John 1689, IV.ii).

By the last decades of the twentieth century, a number of originally independent research programs in social and cognitive psychology had converged in the broad outlines of their approach to the distinction between intuitive and reflective thought. Psychologists from these various programs used a variety of labels and diagnostic criteria for the contrasting sides, however: Jonathan Evans's 2008 survey of the field lists 14 sets of labels, ranging from Reber's 'implicit/tacit' vs. 'explicit' to Schneider's 'automatic' vs. 'controlled', to Sloman's 'associative' vs. 'rule-based' (Evans, Jonathan 2008, Table 1). To talk about the possibility that these various programs were driving at the same thing, some neutral terminology was needed, and in this spirit Keith Stanovich introduced the generic labels

'System 1' and 'System 2' (Stanovich, Keith 1999). Stanovich worried from the start that the label 'System 1' might suggest a commitment to a single architectural structure underpinning the full range of intuitive judgments, where most theorists, including Stanovich himself, hold that these judgments are produced by a variety of distinct modules and systems. The real contrast here concerns the kind of processing involved, and not the system(s) doing the processing. In recent years both Stanovich and Evans have stopping using the labels 'System 1' and 'System 2'; they now officially categorize their research programs as Dual Process Theory (DPT) rather than Dual Systems Theory (Evans, Jonathan 2010; Stanovich, Keith 2011). The Evans and Stanovich programs converged in 2013 with their first co-authored article, a defence of DPT that uses the labels 'Type 1 (intuitive)' and 'Type 2 (reflective)' to describe the two different kinds of processing (Evans, Jonathan and Stanovich, Keith 2013). In what follows, I use their labels 'intuitive' and 'reflective'.

Over the years, various clusters of features have been contrasted as researchers have attempted to characterize the split between intuitive and reflective judgment. To give a partial list, intuitive processing has been described as fast, heuristic, associative, low-effort, automatic, biased, evolutionarily old, shared with animals, and unconscious, where reflective has been considered slow, analytic, rule-based, high-effort, controlled, normatively correct, evolutionary new, uniquely human, and conscious (from Evans, Jonathan 2008, Table 2). It's clear now that there is no single kind of processing that has all and only the characteristics that various researchers over the years have pinned on either the intuitive or the reflective side of the line (Evans, Jonathan and Stanovich, Keith 2013). But it would be premature to take this result as a defeat for DPT. First, some efforts at explaining what is distinctive about intuitive judgment have been better than others, and

our picture has become more accurate as new empirical results have emerged. It is clear that some of the features associated with each type of processing are just frequent correlates of that type, or rough diagnostics rather than definitive features. There are various reasons why a feature may be correlated with one or the other type of processing. In some cases, the common association between a characteristic and a mode of processing is at least partly an artefact of experimental design. For example, many of the early investigations of the contrast between intuitive and reflective judgment worked, for reasons of experimental convenience, with problems for which the two types of processing give different results, and more particularly, problems for which only the reflective result was correct: Kahneman and Tversky's influential work on the Linda problem, for example, falls in this category (Tversky, Amos 1973). Participants who solved the problem intuitively were guided by the match between the description of Linda and the stereotype of a feminist, and gave an answer that could not be correct; participants who solved the problem reflectively attended to the logical structure of the case and invoked the conjunction rule of probabilistic reasoning. Reasoning the second way gets the right answer in this case, but takes longer and is less likely to be executed when working memory is burdened (De Neys, Wim 2006). But it would be a mistake to conclude that intuitive processing generally delivers a biased or incorrect answer, or that reflective processing is always superior: for many problems the two types of processing deliver the same answer, although in different ways. There are problems, for example certain non-causal base rate problems, for which the intuitive answer is more likely to be correct than the reflective one (Stanovich, Keith and West, Richard 1998). The impression that intuitive thinking is bad and reflective thinking is good is according to Evans and Stanovich 'perhaps

the most persistent fallacy in the perception of dual-process theories' (2013, 229). For anyone non-sceptical enough to grant that intuitive processes such as face recognition can yield knowledge, the line between knowledge and ignorance is orthogonal to the line between the intuitive and the reflective.

The fundamental basis of the contrast between intuitive and reflective thinking does not concern knowledge, speed, accuracy, emotional affect, biases or heuristics: the contrast is at the level of processing itself, and concerns the manner in which each interacts with working memory. Working memory is a single limited-capacity resource whose contents are posted to consciousness and therefore available as input to a variety of mental modules, and typically available for explicit report (Baddeley, Alan 2007). A classic example from Steve Sloman (1996) illustrates the way in which intuitive and reflective thinking engage working memory: Sloman invites us to consider the contrast between the task of solving an easy anagram in which just a few letters of a long word are transposed ('involnutray') with the task of solving a much harder anagram ('uersoippv'). For the first task, it is enough to train one's conscious attention on the jumbled letters to have the answer ('involuntary') spring to mind through some autonomous process whose workings are closed to us. For the second, we go through partially introspectable cycles of mental activity, experimenting with different consciously available contents involving various transpositions of letters until we find the solution. The second type of thinking depends on working memory not just to present the original problem, but also to take us via a series of stages towards a solution. Each of these individual stages is itself intuitive, just as Locke observed.

Reflective processing calls on working memory, but it is not entirely executed within working memory: the intermediate stages of the anagram solution are consciously available to modules whose inner workings are closed to us, whose processing is still needed to advance forward. For example, once we consciously entertain an arrangement like 'pupersoiv', we intuitively judge that this is still not a word, through processing that is not itself consciously accessible. Other things being equal, purely intuitive processing is faster: we can solve even the easy anagram by deliberate step-by-step reflective transposition of its letters, but it is faster to let our intuitive functions deliver the answer. The fact that Sloman's example happens to contrast an easy problem that can be solved swiftly with a harder problem that takes time should not however make us think either that intuitive cognition is generally restricted to solving easy problems, or that it always delivers its answers quickly. Because of its broader capacity, intuitive cognition may be particularly well-suited to answering difficult questions that require the integration of a great deal of subtle information; such conditions often enable it to outperform reflective cognition in accuracy, but if the tasks are hard enough, intuitive processing can be quite time-consuming (Reber, Arthur S 1996). Finally, heuristics can apply at either level: we can have a consciously accessible heuristic that we use in discursive reasoning—an explicit rule of thumb, for example—or a heuristic that applies entirely to the selection of responses outside of working memory (Shah, Anuj and Oppenheimer, Daniel 2008).

Various aspects of dual process theory are controversial, but the core idea behind the approach mapped out by Evans and Stanovich is widely shared. For example, Peter Carruthers, who as recently as 2009 was defending a theory he characterized as a version of DPT, now argues that dual process theory is mistaken about the intuitive/reflective split

(2009; 2013). However, his reasons for rejecting DPT are reasons for rejecting older versions of the theory, versions that mark the contrast along lines now classified as non-essential. So, in rejecting what he calls 'dual systems theory' while continuing to accept the reality of a distinction between intuition and reflection, Carruthers argues that intuitive reasoning can be slow, taking as an example an intuitive process that compiles an extensive range of information; he notes that heuristics can be conscious as well as unconscious; he notes that intuitive reasoning can outperform reflective reasoning on some tasks, and indeed can measure up to high normative standards. All of these moves are compatible with the leading current versions of DPT, which locate the decisive difference in the involvement of working memory in processing. Carruthers further emphasizes that on his view there is an important commonality in the kind of cognition applied on both sides of the intuitive/reflective contrast: he rejects views according to which intuitive thinking must be associationist, for example, pointing out that even conditioned behaviour in animals is better understood as generated by rule-governed computation (citing e.g. Gallistel, Charles R and King, Adam Philip 2009).[3] In his view, as in Evans and Stanovich's, what distinguishes reflective from intuitive thinking is the involvement of working memory. Carruthers presents a more specific characterization of its role:

> In outline, the proposal is that reflection operates like this: action-schemata are selected and activated, and are mentally rehearsed (with overt action suppressed); this gives rise to conscious images which are globally broadcast (in the manner of Baars, 1988) and thus made available as input to the full suite of intuitive systems… . (…) The result is sequences of motor images, visual images, or auditory images

[3] In engaging with Gendler's position I pick up her use of 'association' to describe the formation of intuitive impressions, but it is not clear to me whether she is using that term in the strong sense, to rule out computational accounts such as Gallistel's, or just in some weaker sense to describe learning from repeated exposure to stimuli, without conscious deliberation, where this learning could still have a complex information-processing structure. I find computational accounts plausible, but the question of whether they are correct is beyond the scope of this paper.

> (often in the form of so-called 'inner speech', when the actions rehearsed are speech actions), which serve as the conscious components of reflective thought. (…) [W]ith each iteration of mentally rehearsed action the various System 1 systems that 'consume' the globally broadcast images become active, sometimes producing an output that contains or contributes towards the solution. (Carruthers 2013)

On this view, reflective thinking is not something that happens in a separate system independent of the operation of intuition; reflection is realized in successive cycles of intuitive thought whose outputs are posted to consciousness. This characterization also answers a natural question about the distinction between the two modes of thought: if each step of reflective cognition is itself intuitive, the distinction between intuition and reflection looks like a difference in quantity rather than quality. However, not just any string of intuitions posting to consciousness will count as reflective thought. Reflection requires the goal-directed production of a series of mental images or inner speech: it is enabled by a question-directed state of mind (see Friedman, Jane 2013 for an account of such states). A mere succession of conscious mental images—as in daydreaming—may involve the successive operation of individually intuitive processes posting to consciousness, but does not constitute reflective thought.[4] In a reflective process like geometric proof or complex anagram solution, on the other hand, the sequence of individually intuitive steps is controlled to solve a problem, with the output of some stages serving as input to others. This sequential structure is important, and goes beyond the broader phenomenon of maintaining control over attention in the service of answering a question (cf. Koralus, Philipp 2014). In the course of explaining why a sense of effort is insufficient to demarcate reflective thought, Hugo Mercier and Dan Sperber (2009) give an example which illustrates this last point. If you stand in the train station searching for a

[4] I am grateful to Zachary Irving for discussion on this point.

friend among the arriving passengers, it takes effort to sustain attention on scanning all the faces in the crowd. Your interrogative frame of mind—'where is my friend?'—governs your attention to the long string of intuitive impressions formed by face recognition; still, this is not reflective cognition. Reflective cognition requires the sequential use of a progression of conscious contents to generate an attitude, as in deliberation.

Intuitive cognition can also generate an attitude on the basis of a series of conscious contents, where these are not generated in cycles of the agent's own mental activity. To count as intuitive, an attitude does not need to have been formed on the basis of subliminal impressions. In fact, it is disputed whether we *ever* learn on the basis of the subliminal (e.g. in Newell, Ben R and Shanks, David R 2014). In the classic study of implicit learning—Arthur Reber's experiments on artificial language learning—participants consciously attended to strings of letters governed by a mathematically complex 'grammar', and then asked to sort newly presented strings according to whether they fit the earlier grammar. Participants were able to sort the new strings at a rate well above chance despite a near-complete inability to articulate the rules they were following in doing so (Reber, Arthur S 1967). The stimuli were consciously presented, and the intuitive attitude that a new string fit the grammar was consciously available; participants did not however have conscious access to the relationships they were computing in the task. Indeed, participants instructed to reason explicitly about the task underperformed participants given neutral instructions to simply attend to the strings, apparently because the limited capacity of working memory was insufficient to master the mathematical complexity of the grammar (Reber, Arthur S 1989).

Implicit learning can be a powerful source of knowledge; however, in climates where our conscious experiences are misleading, as Gendler observes, implicit learning can leave us with intuitive attitudes that fall short of knowledge. The intuitive judgment that airline travel is riskier than car travel, for example, is generated by a misleading pattern of conscious experiences involving many sensational media reports of plane crashes with car crashes largely passing unreported; similar distortions contribute to the intuitive attitudes characteristic of racism and sexism. Generally good intuitive processes can yield poor results when they start from misleading materials, just as generally good reflective processes like conscious deliberation can take us astray when our initial premises are wrong—there is no sharp epistemological contrast between intuition and reflection here.

We typically use reflective processes to answer novel or complex questions that intuition is not poised to answer, although it is possible to deliberately generate a reflective solution to a question already answered intuitively, for example in deliberate double-checking of what seems right. Consciousness, on this view, plays a key role in distinguishing intuition from reflection, but the role it plays has to do with the generation of these attitudes rather than whether they are available to consciousness after being formed, for use in deliberation. This approach yields a somewhat different account of the phenomena motivating Gendler's paper. As Gendler describes the CV study, for example, scientists who knew the equality of men and women were moved by factors outside the realm of consciousness to undervalue the female candidates, and their knowledge was powerless to stop them. I see the study as showing that those with some reflectively generated commitment to gender equality could also have a reflectively generated attitude that the female candidate is less competent, where this latter attitude is consciously

available and can support a pattern of action that is inconsistent with the general reflective commitment. Understood this way, the CV study does not pose a threat to the view that knowledge commands right action.

The CV study was a between-subjects design: an otherwise identical CV was randomly assigned the name 'Jennifer' for half the participants and 'John' for the other half. To increase the potential for informative variation in ratings, this undergraduate CV showed 'high but slightly ambiguous competence', including a co-authored publication. Given the impression that they were delivering feedback to help a real student with career development, participants rated their own likelihood of hiring the student, after rating the student's competence and likeability. Gendler sees the study as marking a hidden gap 'between our evaluations on the one hand, and our reflective commitments on the other,' where there are 'associations that implicitly guide' the problematic evaluations. But what is the psychological contrast between the participants' presumed reflective commitment to gender equality and their distorted evaluation of the female applicant? Both of these judgments should count as reflective because formed through deliberation on consciously available considerations. The ambiguous CV is a novel and complex stimulus, and the judgments made about it do show the mediation of consciously available factors. Gendler emphasizes that the evaluations of hireability were not driven by conscious feelings of dislike—likeability scores failed to predict hireability—but, as she herself also notes, the same mediation analysis shows that hireability evaluations were predicted by conscious judgments concerning competence. It is a presumption of the study that those evaluations of competence were generated reflectively: reading a CV to calculate a numerical competence score is not like reading a facial expression for emotion—it requires conscious

weighing of factors. This is not a case in which a reflective commitment is in tension with an intuitive assessment; this is a case in which one reflective judgment (about the particular fictional woman candidate) is in tension with another reflective judgment (about women in general). To assess the upshot for knowledge, we need to examine the epistemic status of these two commitments.

On the general attitude, Corinne Moss-Racusin and colleagues (2012) did not directly question participants on their reflective attitude to the equality of women, but they did end their survey with the eight-question 'Modern Sexism Scale' (Swim, Janet K, Aikin, Kathryn J et al. 1995), which asks participants to rate their level of agreement to statements such as 'It is rare to see women treated in a sexist manner on television', and 'On average, people in our society treat husbands and wives equally'. The MSS is a more sensitive measure of sexism than more blatant scales like Spence's (1973) Attitude to Women Scale (AWS), with items like 'Women should worry less about their rights and more about becoming good wives and mothers' and 'Women should be given equal opportunity with men for apprenticeships in the various trades.' It is hard to attribute knowledge of women's equality to those who give sexist answers on the AWS. However, there is significant correlation between the old and new scales: those who show strong bias on the MSS also show it on the AWS (Swim, Janet K and Cohen, Laurie L 1997). High scores on either scale correlate with the possession of false beliefs about topics such as the percentage of men and women in male-dominated industries (Swim, Janet K, Aikin, Kathryn J et al. 1995). It may be true that most of those now employed as research scientists would explicitly express some commitment to the equal treatment of men and

women, but it is not obvious that the study participants who exhibited higher bias on the MSS can be credited with knowledge as opposed to lip service here.

In the CV study, bias as measured by the MSS significantly predicted low evaluations of the female CV: the greater the bias, the lower the perceived competence (β=-0.36, $P<0.01$) and hireability (β=-0.39, $P<0.01$) of the female job applicant. Higher bias had the strongest impact on the amount of mentoring participants were willing to offer (β=-0.53, $P<0.001$). Bias was unrelated to responses to the male CV, except for a statistically marginal positive correlation in mentoring (more biased participants were keener to mentor a man than less biased ones, $P=0.09$). Setting aside the last (marginal) correlation, I take these results to indicate that the general disparity in responses marks a distorting effect of sexism specifically on the perception of women: thanks to sexism, there is a tendency to under-attribute competence to Jennifers. One might have thought that sexism also works to inflate assessments of men, or to distort perceptions of both groups equally but in opposite directions; this particular study suggests that the distortion compromises evaluations of women in particular. The variance in responses to the female CV indicates that not everyone was guilty of a distorted evaluation, and the correlation between the distortion and scoring on the Modern Sexism Scale suggests that weaker grasp of general truths about men and women correlates with more distorted evaluations of particular women. Far from showing that general knowledge of equality is powerless to stop unfair assessment of an individual, the study showed that unfair assessment was most likely in those with the most skewed general attitudes.

That last claim is an inference from an empirically discovered correlation. One might wonder whether a stronger claim could be made about the overall relationship between knowledge of the general and the particular. Is it possible to 'know the difference of good and evil' as far as the equality of men and women is concerned, while still tending to undervalue the competence of women? If the fact that a candidate is female leads you to rate her lower, you are failing in the application of the general principle to a particular case. But such failures of application arguably undermine a claim to full knowledge of the general principle. To count as knowing that equally trained and accomplished men and women are equally competent, it is not enough to say that they are; one should tend to recognize particular equally trained men and women as equally competent. Skewed ratings of competence are not signs of an invisible horse leading people away from knowledgeable reflective judgments; they are better read as reflective judgments which are in tension with other reflective judgments, where the existence of this tension is one sign that both sets of judgments fall short of knowledge.

Gendler is right that that the tension here is not as introspectively evident as the conflict we feel in classic cases of weakness of the will, but it may be overstating things to call it invisible, or to suggest that it is beyond the range of introspective probing. Intuitive attitudes about female competence are formed without awareness of a process of deliberation, but this does not entail that these attitudes are introspectively inaccessible.[5]

[5] By 'introspectively accessible' I just mean accessible by means of attending to inner speech and so forth; I do not mean to suggest that we have a special channel for detecting our own propositional attitudes which is radically different in kind from the channel we use in interpreting the attitudes of others. Peter Carruthers (2011) may well be right to argue that our own attitudes are also detected by the same mindreading capacity we apply to others, so that even our inner speech is interpreted rather than understood through direct access to something deeper. What is crucial here is whether we can have access to our own attitudes, not whether the access is of a distinctive type.

Implicit attitudes are those attitudes revealed by implicit tests such as the IAT, rather than by explicit self-report, but the fact that an attitude has been revealed by an indirect measure does not entail that it is unavailable for direct discovery (for a review of literature on this point, see Gawronski, Bertram, Hofmann, Wilhelm et al. 2006). Intuitive attitudes about women and racial minorities can differ sharply from reported attitudes when experimental participants have incentives to present themselves well, but this seems to say more about the unreliability of reported attitudes under these conditions than about the inaccessibility of the intuitive. People who are asked to predict their implicit associations (rather than endorse them) can do so quite well (Hahn, Adam and Gawronski, Bertram 2014). Among participants who are told that the implicit association test (IAT) functions as a lie detector revealing true attitudes, explicit attitude reports concerning race show a 'fairly strong positive relation' with IAT results (Nier, Jason A 2005). Similarly, if people are asked about their 'gut reactions' as opposed to their 'actual feelings' about gay people, their responses are a closer fit to their responses on the IAT (Ranganath, Kate A, Smith, Colin Tucker et al. 2008). Gut feelings, even if not endorsed, are available to introspection, and a suitably pensive person could reasonably find himself wondering, on the basis of what is introspectively available, about the relationship between his gut feelings of inequality and his contrary explicit pronouncements.

Section 3: Self-regulation and the Socratic view

Even if implicitly tested attitudes are available to consciousness when we focus on them, they are not always present to mind. Gendler is doubtless right to note that these attitudes

can diverge from our reflective commitments, and that they can lead us away from those commitments without warning. She proposes that implicitly tested attitudes consequently call for special forms of self-regulation: 'traditional strategies for self-regulation cannot be straightforwardly applied to their management' (p.2). This may be too strong, at least as long as we count the 'management' of implicitly tested attitudes as including taking steps to ensure that they are not what governs action. Perhaps this counts as suppression rather than management, but in cases of conflict between gut feelings and reflective commitments, there are some familiar steps we can take to stick to our commitments. Sometimes all that is needed is the chance to think twice before acting: the shooter bias effect that Gendler discusses emerges only under time pressure (exactly to exclude deliberate strategies, participants had to respond within 630 ms to avoid a 'TOO SLOW' error message). According to the study authors, 'there is little doubt that people can control their behaviors when they have motivation and opportunity to think carefully' (Stewart, B.D. and Payne, B.K. 2008, 1333). A recent review concludes that 'implicit measures will primarily predict behaviour under conditions of low opportunity or motivation to control, or when individuals rely on automatic processes to guide their behaviour for any other reason' (Friese, Malte, Hofmann, Wilhelm et al. 2009). It's true that the silence of our automatic processes may leave us unaware that rapid action will run contrary to our reasoned commitments, in those cases where they will be out of line; however, this failure to get advance warning for the need for control is not a special problem arising from the implicit. We can be equally unaware of lurking danger in conscious deliberation, where for example the easiest way to frame a problem will lead to a fallacious reflective judgment (see e.g. the discussion of the Levesque task in Stanovich, Keith 2011). No subjectively apparent signal

marks that we are falling short of knowledge: even in conscious deliberation many aspects of our thinking are introspectively inaccessible (Carruthers, P. 2011).

There are, however, cases in which extra time is not an option, and in any event, there is no reason to think that we should always act only on reflective commitments and not on intuition. Insofar as our intuitive judgments constitute knowledge, they make a fine basis for action, but what options do we have for self-regulation in those conditions where we have no choice but to act intuitively? Gendler draws inspiration from Stewart and Payne's 2008 shooter bias paper, which explores the problem of acting rapidly under false impressions. Although American Blacks are only half as likely as American Whites to own guns (Pew_Research_Survey 2013), distorted media impressions associate Blacks much more strongly with guns, generating interference for American participants charged with the task of very rapidly sorting guns from other objects when primed with Black and White faces. As Gendler explains, instructing participants to attend to accuracy was not effective in reducing this interference. Gendler highlights the benefits of Stewart and Payne's less conventional remedies; here I'd like to highlight their residual drawbacks.

The first experiment probed a manipulation designed to offset the anti-Black bias ('whenever I see a Black face on the screen, I will think the word 'safe''). In the authors' view, the reason why substituting 'safe' with 'accurate' (or 'quick') had no effect on the bias was that the task already demanded quickness and accuracy; those instructions were simply redundant, making them good controls. As Gendler observes, the 'safe' manipulation reduced but did not eliminate stereotypes; false negatives (failing to shoot in the gun condition) were always significantly rarer for Whites than Blacks, although less so in the

safe condition than in the quick and accurate conditions. If the differences between responses to Black and White primes shifted in the direction of equality, however, it is somewhat disturbing that it didn't do so by increasing the accuracy on the Black primes, but by decreasing the accuracy on the White primes (for cases where there was a tool); and on the gun trials, by decreasing accuracy for the Black primes. The second experiment, which expanded to include previously unseen Black and White faces, did see decreased error on trials where a Black face appeared before a tool, but increased error when a Black face appeared before a gun. It is not obvious that this is a 'significant reduction in error rates' as opposed to a case of one kind of error being traded for another. If implementation intentions work by strengthening the association between the stimulus and the desired response, one lesson of the study is that associating safety with the Black prime led participants to expect safe tools rather than guns across the board (whether or not these expectations were accurate). On its own, this manipulation does something attractive (by reducing the Black-White discrepancy), but it is not epistemically unproblematic (even if we are very clear that the association between Blackness and a failure of safety is both empirically false and morally troubling). If this particular manipulation still leaves us with a weak capacity to act knowledgeably—to produce an accurate sorting of the guns and tools—then it is not an unmixed blessing.

It is not easy to undo intuitive impressions formed by long experience, even when one is aware that this experience has been misleading. We can, as Gendler observes, work to structure our external environment so that our experiences will, over time, be more accurate: for example, by campaigning against distorted media presentations of demographic groups. We can also take steps to reduce reliance in decision-making on the

types of judgment we know to be most compromised. I see both types of measure as well motivated by the Socratic view: the first aims to increase the extent of our knowledge, and the second aims to shift the basis of our action away from judgments made in ignorance. Advocating both of these methods is entirely compatible with the view that knowledge is a noble and commanding thing, which cannot be overcome.[6]

[6] For feedback on the ideas in this paper, I am grateful to Tamar Gendler, Sergio Tenenbaum, and participants in the 2012 Chapel Hill Colloquium in Philosophy. I am also grateful to the Social Sciences and Humanities Research Council of Canada for funding my work.

References:

Baddeley, Alan 2007: *Working memory, thought, and action*: Oxford University Press, 2007.

Carruthers, P. 2011: *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. New York: Oxford University Press, 2011.

Carruthers, Peter 2009: 'An architecture for dual reasoning', in *In Two Minds: dual process and beyond*. Edited by Keith Frankish and Jonathan Evans. Oxford: Oxford University Press, 2009, pp. 109-27.

Carruthers, Peter 2013: 'The fragmentation of reasoning', in *La coevolución de mente y lenguaje: Ontogénesis y filogénesis. Retrieved from* http://www*. philosophy. umd. edu/Faculty/pcarruthers/TheFragmentationofReasoning. pdf*. Edited by P. Quintanilla. Lima: Fondo Editorial de la Pontificia Universidad Católica del Perú, 2013.

De Neys, Wim 2006: 'Automatic–heuristic and executive–analytic processing during reasoning: Chronometric and dual-task considerations'. *The Quarterly Journal of Experimental Psychology*, 59, pp. 1070-100.

Evans, Jonathan 2008: 'Dual-processing accounts of reasoning, judgment, and social cognition'. *Annual Review of Psychology*, 59, pp. 255-78.

Evans, Jonathan 2010: 'Intuition and Reasoning: A Dual-Process Perspective'. *Psychological Inquiry*, 21, pp. 313-26.

Evans, Jonathan and Stanovich, Keith 2013: 'Dual-Process Theories of Higher Cognition Advancing the Debate'. *Perspectives on Psychological Science*, 8, pp. 223-41.

Friedman, Jane 2013: 'Question-directed attitudes'. *Philosophical Perspectives*, 27, pp. 145-74.

Friese, Malte, Hofmann, Wilhelm and Schmitt, Manfred 2009: 'When and why do implicit measures predict behaviour? Empirical evidence for the moderating role of opportunity, motivation, and process reliance'. *European Review of Social Psychology*, 19, pp. 285-338.

Gallistel, Charles R and King, Adam Philip 2009: *Memory and the computational brain: why cognitive science will transform neuroscience*: John Wiley & Sons, 2009.

Gawronski, Bertram, Hofmann, Wilhelm and Wilbur, Christopher J 2006: 'Are "implicit" attitudes unconscious?'. *Consciousness and Cognition*, 15, pp. 485-99.

Hahn, Adam and Gawronski, Bertram 2014: 'Do implicit evaluations reflect unconscious attitudes?'. *Behavioral and Brain Sciences*, 37, pp. 28-29.

Koralus, Philipp 2014: 'The Erotetic Theory of Attention: Questions, Focus, and Distraction'. *Mind & Language*, 29, pp. 26-50.

Locke, John 1689: *An Essay Concerning Human Understanding*. London, 1689.

McDowell, John 1979: 'Virtue and reason'. *The Monist*, 62, pp. 331-50.

Mercier, Hugo and Sperber, Dan 2009: 'Intuitive and reflective inferences', in *In two minds: Dual processes and beyond. Oxford, UK: Oxford University Press*. Edited by Jonathan Evans and Keith Frankish. Oxford: Oxford University Press, 2009, pp. 149-70.

Moss-Racusin, Corinne A, Dovidio, John F, Brescoll, Victoria L, Graham, Mark J and Handelsman, Jo 2012: 'Science faculty's subtle gender biases favor male students'. *Proceedings of the National Academy of Sciences*, 109, pp. 16474-79.

Newell, Ben R and Shanks, David R 2014: 'Unconscious influences on decision making: A critical review'. *Behavioral and Brain Sciences*, 37, pp. 1-61.

Nier, Jason A 2005: 'How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach'. *Group Processes & Intergroup Relations*, 8, pp. 39-52.

Pew_Research_Survey 2013: 'Gun Ownership Trends and Demographics', 2013.

Plato 1976: *Protagoras*. Indianapolis: Bobbs-Merrill, 1976.

Ranganath, Kate A, Smith, Colin Tucker and Nosek, Brian A 2008: 'Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods'. *Journal of Experimental Social Psychology*, 44, pp. 386-96.

Reber, Arthur S 1967: 'Implicit learning of artificial grammars'. *Journal of verbal learning and verbal behavior*, 6, pp. 855-63.

Reber, Arthur S 1989: 'Implicit learning and tacit knowledge'. *Journal of Experimental Psychology: General*, 118, pp. 219.

Reber, Arthur S 1996: *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*: Oxford University Press, 1996.

Shah, Anuj and Oppenheimer, Daniel 2008: 'Heuristics made easy: An effort-reduction framework'. *Psychological Bulletin*, 134, pp. 207-22.
Sloman, Steven 1996: 'The empirical case for two systems of reasoning'. *Psychological Bulletin*, 119, pp. 3-22.
Spence, Janet T, Helmreich, Robert and Stapp, Joy 1973: 'A short version of the Attitudes toward Women Scale (AWS)'. *Bulletin of the Psychonomic Society*.
Stanovich, Keith 1999: *Who is rational?: Studies of individual differences in reasoning*. Mahwah, NJ: Lawrence Erlbaum, 1999.
Stanovich, Keith 2011: *Rationality and the reflective mind*. New York: Oxford Univ Pr, 2011.
Stanovich, Keith and West, Richard 1998: 'Individual differences in rational thought'. *Journal of Experimental Psychology: General*, 127, pp. 161.
Stewart, B.D. and Payne, B.K. 2008: 'Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control'. *Personality and Social Psychology Bulletin*, 34, pp. 1332-45.
Swim, Janet K, Aikin, Kathryn J, Hall, Wayne S and Hunter, Barbara A 1995: 'Sexism and racism: old-fashioned and modern prejudices'. *Journal of Personality and Social Psychology*, 68, pp. 199.
Swim, Janet K and Cohen, Laurie L 1997: 'Overt, Covert, And Subtle Sexism A Comparison Between the Attitudes Toward Women and Modern Sexism Scales'. *Psychology of Women Quarterly*, 21, pp. 103-18.
Tenenbaum, Sergio 1999: 'The judgment of a weak will'. *Philosophy and Phenomenological Research*, 59, pp. 875-911.
Tversky, Amos 1973: 'Availability: A heuristic for judging frequency and probability', in *Cognitive Psychology*: Netherlands: Elsevier Science, 1973, pp. 207.