MORAL BLINDSIGHT AND HUCKLEBERRY FINN[1]

RICHARD HOLTON, CAMBRIDGE

In his introductory 'Notice' to *Huckleberry Finn* Mark Twain cautions that 'Persons attempting to find a Motive in this narrative will be prosecuted; persons attempting to find a Moral in it will be banished' (Twain 1884, lv). It has not stopped philosophers from trying. In the fifty years since Jonathan Bennett introduced it to the philosophical literature, the example of how Huck, who is escaping an abusive home life, protects his friend Jim, who is escaping a life of slavery, has become a staple in discussions of moral motivation and of moral worth.

But have we been finding the right motives and morals in it? Central to philosophical discussions has been the episode in Ch. XVI when, some way into their flight, Huck becomes convinced that he should turn Jim in. Intent on the task, Huck paddles off in a canoe, leaving Jim on the raft on which they have been making their escape. But on meeting a pair of slave hunters, Huck finds himself quite unable to betray Jim, telling them instead that the man on the raft is white. He then keeps them from investigating by constructing an elaborate and beautifully delivered lie to the effect that the man is his father, who is suffering from small-pox.

Bennett saw the case as one of a conflict between ethics and sympathy. Others since have seen it as involving what Nomy Arpaly and Timothy Schroeder (1999) call 'inverse akrasia': akratically doing the right thing while believing one is doing the wrong. In fact though, what they attribute to Huck is more than that. They think that he shows something that I shall call *moral blindsight*: acting for a moral reason (and not simply in accord with it) while being quite unaware that that is what one is doing.

I am sceptical that any of these authors give the correct reading. Some ten years after writing *Huckleberry Finn*, Twain, short of money, gave a world tour in which he read excerpts from the book. Starting in the US in 1895, he read the relevant passages from Chapter XVI under the title 'Small-pox and a lie save Jim'. It proved to be the most popular of his readings, and on the three-week trip to Australia for the next leg of the tour he wrote a short introduction to it. It begins:

> Next I should exploit the proposition that in a crucial moral emergency a sound heart is a safer guide than an ill-trained conscience. I sh'd support this doctrine

[1] It took Mark Twain seven year to write *Huckleberry Finn*. I have comfortably outdone him here. The original version of this paper was written for a 2009 conference on Inverse Akrasia in Tübingen, and has been on my webpage ever since under the title 'Inverse Akrasia and Weakness of Will'. The current version, while keeping to the same main themes, has been completely rewritten. Thanks to those at the Tübingen conference, to those in the Good Life Workshop at Cambridge, and especially to Rae Langton, Hanna Pickard, Ian Phillips and two referees for comments on the written version that resulted in many much-needed changes.

> with a chapter from a book of mine where a sound heart & a deformed conscience come into collision, & conscience suffers defeat. (Twain 1895, 806)

That does not sound like a case of a person doing something in the conviction that it is simply wrong. Twain's portrayal of Huck was, I think, very different: that of someone caught between the voice of conscience, and a heartfelt motivation that is already recognizably moral, even to Huck himself. It is this latter that Twain wanted to privilege as the source of moral insight. In fact conscience turns out not to be a source of insight at all, but an indiscriminate monitor of one's own behaviour, as prone to rebuke for doing the right thing as for doing the wrong.[2]

Does this matter? Arpaly thinks not. She writes: 'Whether I (and Bennett) understand Twain's text correctly was not material to my argument' (Arpaly 2015, 142).[3] I'm not so sure. I have general methodological concerns about how we should use fictional examples. More specifically, I think that our failure to come up with realistic cases of moral blindsight says something important about how moral motivation works. I think that the alternative account that emerges from the reading I want to defend is in many ways more interesting than those defended by Bennett and by Arpaly and Schroeder: more plausible, more common, and more nuanced in the different routes to moral knowledge (or mistake) that it acknowledges. And it makes better sense of our own reactions to Huck's case.

Let me start with some reflections on the methodological issues.

THE USES OF FICTION

I aim to stand up for the real Huck Finn. But the real Huck Finn is a fiction. What can be learned from a fiction? Given that the use of fiction for moral insight has been the subject of considerable controversy, I'm not going to try for a general account here, but will focus just on Huck.[4] Twain is not the kind of writer to insert a bit of explicit theory into his fiction, a practice that Proust complained was like leaving on the price tag[5]; besides, *Huckleberry Finn* is written entirely in the voice of Huck, so there is no narrator to do the theorising. If there are lessons to be learned from the novel then, they will be shown rather than told.

---

[2] A similar dilemma arises for Huck in Ch. xxxi. Twain never revisited that discussion, but I think it should be understood in much the same way.

[3] See also Markovits (2010): 'Of course, nothing really turns on the correct interpretation of Huck's case—what matters is that a case of the kind I am imagining, in which someone acts rightly while really believing he is acting wrongly, is psychologically imaginable.' p. 208.

[4] For a useful summary of the controversies see (Currie et al. 2024).

[5] 'D'où la grossière tentation pour l'écrivain d'écrire des œuvres intellectuelles. Grande indélicatesse. Une œuvre où il y a des théories est comme un objet sur lequel on laisse la marque du prix.' (Proust 1954, 882). There is an interesting question whether this comment violates its own restriction.

Clearly one cannot look to fiction for straightforward ontological claims: no one would argue for the existence of talking horses by citing *Gulliver's Travels*. So how might *Huckleberry Finn* be used to help us to understand some phenomena in moral psychology? I see three potential uses, each of which have parallels elsewhere in metaphysics:

> (i) *the exemplary use*: to provide a clear and striking example of a phenomenon that we already know exists;
>
> (ii) *the possibility use*: to provide an argument for the possible existence of something which may not actually exist, typically using the premise that conceivability entails possibility;
>
> (iii) *the reactive use*: to provide a case against which we can judge our own reactions, whether or not the case is actual, or even strictly possible.

In metaphysics, the possibility and reactive uses predominate. When David Chambers imagines zombies in order to make a point about consciousness, and other authors cite a film like *Blade Runner* to reinforce it, there is no suggestion that zombies actually exist. The move is one from a conceivable story to a possible one. That is a possibility use. In contrast, when Putnam asks us whether we would call XYZ 'water', he is not really asking us to believe that XYZ is possible; indeed he doesn't say enough for it to be clear on what that possibility would be. He just wants us to have a good enough grip on it to ask about our reactions—much as I might ask: 'If it turned out that Hesperus was not Phosphorous, what would you say?'

The use of fictitious examples in ethics and moral psychology has been similar, but here there needs to be an important difference of focus. In moral psychology the mere demonstration of metaphysical possibility is not terribly interesting. Of course people could be configured in all kinds of ways that differ from how they are actually configured. What we want to know is what moral behaviour is possible—or better, what would be characteristic—when we hold how they are actually configured. In so far as the possibility use of fiction is helpful, it is thus within tightly constrained boundaries: as Rousseau famously put it, men must be kept as they are, even if we imagine their circumstances as they might be. Then it inevitably fades into the exemplar use. The fictional case simply provides a nice clear example of something that we are already confident actually exists. Bennett seems to see Huck in this way: as providing a striking example of a phenomenon with which we are already familiar, to be placed alongside the real examples of Heinrich Himmler and Jonathan Edwards. Similarly Arpaly and Schroeder seem concerned to use their reading of Huck (whether or not it is Twain's) to provide a good example of inverse akrasia, to be placed besides other examples: the fictional example of Neoptolemus, from Sophocles' *Philoctetes*; the real example of Oskar Schindler, though as he is portrayed in Spielberg's somewhat fictionalized film *Schindler's List* (we shall return to both of these cases later); and, in Arpaly's subsequent work, the imaginary example, not sketched in any detail, of the young Ayn Rand fan, who

believes in the rightness of selfishness, but is not herself selfish (Arpaly 2015, p, 143).

Arpaly and Schroder clearly make use of the reactive use too, deploying the example to test our moral reactions to Huck, both our reactions to his actions and to his character:

> One does not only think helping Jim was a good thing for Huckleberry to do, but regards Huckleberry's action as saying something important about Huckleberry's self. The reader of Twain's novel tends to see Huckleberry's action not as an accidental good deed done by a bad boy, but as indicative of the fact that Huckleberry is, in an important sense, a good boy, a boy with his heart in the right place. (Arpaly and Schroder 1999, 163)

The use of fictitious examples in metaphysics has been controversial. Can we really imagine what at first glance we think we can? And even if we can, are our reactions to them genuine, or are they over-influenced by the theory that we have accepted?[6] In the case of ethics, given the need for realism, the problems are all the more pressing. Take the exemplary use. Can we be confident that the fictitious cases of moral failings are examples of genuine phenomena, rather than distorted renditions of those phenomena, influenced by a theory of it that may well be wrong? Roy Baumeister, in his study of evil, writes:

> I have, mostly, resisted the constant temptation to illustrate key points with episodes from literature, movies, and other fictional entertainments. Many books on evil are filled with such illustrations and some do not even seem to regard as meaningful the difference between, say, something said by Macbeth versus something said by Ted Bundy. I concluded fairly early in the project, however, that because our cultural ideas about evil are so powerful and so discrepant from the actual psychology of the perpetrators, fictional illustrations were often worse than useless. (Baumeister 1996, x–xi)

That is primarily a worry about the possibility use, but it leads to a concern too about the reactive use, about how to understand our moral reactions to fictional cases. If the fictional description involves a departure from reality in the way that Baumeister worries it will, what are we reacting to? Is it really to the fictional case? Or, given that our reactions have been trained in the real world, do we somehow look past the fiction to the closest real cases, and react as we would react to them?

Unlike Baumeister, I do not think that these worries are so great that we should avoid fictional examples altogether. I do think though that we need to be very careful. Bennett is surely right when he says that 'Twain wrote the episode with such unerring precision', but precision in itself gives no assurance of authenticity. What does, I think, give some assurance in this case is the fact that, although Huck is obviously fictional, there is much in the character that is autobiographical.

---

[6] See, for characteristic criticisms, Dennett 1995 and Wilkes 1985.

Twain did not grow up in the Deep South. He grew up in Hannibal, Missouri, the model for the fictional town in which he sets Huck's upbringing. Missouri was a slave state, and Twain accepted the almost universal belief in the rightness of slavery:

> In my schoolboy days I had no aversion to slavery. I was not aware that there was anything wrong about it. No one arraigned it in my hearing; the local papers said nothing against it; the local pulpit taught us that God approved it, that it was a holy thing, and that the doubter need only look in the Bible if he wished to settle his mind—and then the texts were read aloud to us to make the matter sure; if the slaves themselves had an aversion to slavery they were wise and said nothing. (Twain 1906, 212)

Nevertheless, the institution of slavery was different to that in the much-feared plantations of the Deep South. Hannibal was just across the river from Illinois, which was a (partially) free state. And, more importantly, the domestic nature of much of the slavery meant that contact was very different:

> All the negroes were friends of ours, and with those of our own age we were in effect comrades. I say in effect, using the phrase as a modification. We were comrades, and yet not comrades; color and condition interposed a subtle line which both parties were conscious of, and which rendered complete fusion impossible. (*Ibid*., 211)

Jim was partly based on a real character, Uncle Dan'l, 'a faithful and affectionate good friend, ally and adviser … whose sympathies were wide and warm' (*Ibid*.) As a great deal of recent work has shown, Twain grew up with ambivalence, and struggled with reconciling the effects of his own racist upbringing with his growing friendship with a number of black people.[7] The struggles that Huck faces are particularly dramatic, but the attitudes that they illustrate, including those of a child's inability to articulate quite what they are witnessing and thinking, had been experienced by Twain. Another telling passage:

> When I was ten years old I saw a man fling a lump of iron-ore at a slaveman in anger, for merely doing something awkwardly—as if that were a crime. It bounded from the man's skull, and the man fell and never spoke again. He was dead in an hour. I knew the man had a right to kill his slave if he wanted to, and yet it seemed a pitiful thing and somehow wrong, though why wrong I was not deep enough to explain if I had been asked to do it. Nobody in the village approved of that murder, but of course no one said much about it. (Twain 1897, 352)

So we have good reason to think that the psychological picture painted in *Huckleberry Finn* is a realistic one, or at least as accurate and realistic as the mature Twain, looking back on his own experiences, and refining his understanding over the years, could make it. It is a complex picture. Arpaly and Schroeder are surely right that we judge Huck well when he lies to the slave traders; certainly much

[7] See (Fishkin 2025) for a very helpful survey and discussion.

better than we would if he were to have turned Jim in. But we form that judgment on the basis of having read many pages about Huck and Jim and about the relationship between the two. It is hard to be sure of what it is in that story that is motivating our judgment. If we wish to rest anything on it, we cannot simply amend the story to suit our purposes, and assume that our reactions will remain the same. We need to leave the story as it is.

But how should we understand the story? Let us start by looking more carefully at what has been said by philosophers so far.

BENNETT

Bennett sees Huck's position as involving a battle between morality and sympathy. Morality tells him to turn Jim in; sympathy tells him not to. In saying this Bennett wants to draw a radical separation between sympathy and any kind of moral judgment:

> As for 'sympathy': I use this term to cover every sort of fellow-feeling, as when one feels pity over someone's loneliness, or horrified compassion over his pain, or when one feels a shrinking reluctance to act in a way which will bring misfortune to someone else. These feelings must not be confused with moral judgments. My sympathy for someone in distress may lead me to help him, or even to think that I ought to help him; but in itself it is not a judgment about what I ought to do but just a feeling for him in his plight. (Bennett 1974, 124)

It is rather surprising that Bennett says so little to defend this separation, given that there is such a powerful body of work that sees sympathy as central to morality, an approach that continues to be hugely influential, not just in philosophy, but also in psychology. Bennett does, however, offer a real example, that of Himmler, as someone who succeeded in making the separation in his own attitudes; so let me make a brief digression to examine that.

After quoting from Himmler's notorious Posen speech of 1943, a speech that, *inter alia*, stressed the need to remain decent even while committing genocide, Bennett writes:

> Himmler saw his policies as being hard to implement while still retaining one's human sympathies—while still remaining a 'decent fellow'. He is saying that only the weak take the easy way out and just squelch their sympathies, and is praising the stronger and more glorious course of retaining one's sympathies while acting in violation of them … Although his policies ran against the human grain to a horrible degree, he did not sandpaper down his emotional surfaces so that there was no grain there, allowing his actions to slide along smoothly and easily. (Bennett 1974, 128)

He goes on to say that Himmler paid a price for this, in terms of psychosomatic stomach pains that resulted from overruling his sympathies.

But all this is to completely misrepresent Himmler. A little earlier in the very same Posen speech that Bennett quotes, Himmler had made it clear that he thought the group to whom real decency should be extended was tightly resitricted. It was, he said, necessary 'to be honest, decent, loyal, and comradely … to those of our own blood *and to no one else*' [my emphasis]. He went on to say that, in so far as decency could be applied to the victims of his policies, it was simply of the kind one might apply to animals, and should be extended absolutely no further:

> We Germans, who alone on this earth have a decent attitude to animals, will of course adopt a decent attitude to these human animals, but it is a crime against our own blood to worry about them and to apply ideals to them. (Longerich 2011, 309)

So Himmler thought that it would have been a crime to show normal human sympathy to those he aimed to kill; there is absolutely no evidence that he ever did. Contra Bennett, he had no need to 'sandpaper down his emotional surfaces'. When it came to those he persecuted, he had none.

As for the stomach pains: Himmler did indeed have them, and he thought them psychosomatic; he had discussed them in correspondence with his fiancée way back in 1928, long before he was actually murdering anyone. (We only have her side of the correspondence, but much can be inferred from that.) He didn't, however, think that the pains resulted from suppressing his sympathies. On the contrary, as his biographer Peter Longerich puts it, Himmler ascribed his stomach pains 'to his constant attempts to discipline and control himself—to be "good and decent", as he put it. He remarked to her that he would like once in a while to be allowed "not to be good and decent", as a kind of therapy.' (Longerich 2011, 105).

So Himmler provides no example of someone who separated the sympathetic and the moral. I remain sceptical that such a clear separation as Bennett posits is possible; I'll return to this in the conclusion. But even if we can conceptualize sympathy in Bennett's way—if we understand it as *mere* sympathy—then it is certainly inadequate to describe Huck's motivations. As we have seen, Arpaly and Schroder think that this episode is fundamental to our seeing Huck as 'a good boy'; yet if he acts of the basis of 'just a feeling' it is hard to see why we would think that.[8] What more is going on?

---

[8] It is not uncontroversial, even in the philosophical discussions, that we should think so well of Huck. (Hursthouse 2001) raises qualms; (Gehrman 2018) and (Clarke 2020) go much further, seeing him as a deeply flawed individual, largely because they put a great deal of weight on the so-called 'Evasion' chapters towards the end of book. Here Huck is, to say the least, complicit in Tom Sawyer's callous and unnecessary manipulation of Jim for his own vainglorious ends. I don't know what to make of these chapters. The literary discussion has tended to treat them as an extended allegory for the Reconstruction era; see (Clarke 2020) for a nice summary. The chapters certainly seem to abandon any careful characterization of Huck's psychology. For most of this paper I will in effect follow Hemingway's advice to stop reading the book before getting to them.

Arpaly and Schroeder frame their account in terms of *inverse akrasia*. In fact though, that is only one of a number of features that they think is distinctive of Huck and of the other cases they discuss. Let me spell out these features, starting with akrasia itself.

Standard contemporary accounts of akrasia are subjective—or internalist—in the sense that make reference to the attitudes of the agent, not to whether those attitudes are right. (Whether Aristotle himself understood akrasia in this way is an issue to which we shall return.) In defining *inverse* akrasia, Arpaly and Schroeder take this standard subjective approach, but then add an objective or external element. Here is how they introduce the notion:

> To act akratically is to believe one course of action to be, all things considered, the right one, yet to take some other course of action instead. The most commonly discussed examples of akrasia center around wrongdoing of one sort or another … In the cases of interest to us, however, akrasia results in what, for lack of a better word, might be called *rightdoing* of one sort or another. That is, the akratic course of action is superior to the course of action recommended by the agent's best judgment. Because these cases reverse our usual expectations from akratic action, we call them cases of inverse akrasia. (Arpaly and Schroeder 1999, 162)

Arpaly and Schroeder thus start with the normal subjective notion of akrasia, defined in terms of an agent's beliefs about what is best; but then they partition the resulting cases using an objective measure of wrongdoing or rightdoing. Inverse akrasia comes about when an akratic action results in an outcome that is *in fact right*, even though the agent believes it to be wrong. To be clear then: inverse akratic action is akratic in the standard sense; we simply *add* the further requirement that the action is in fact right.

Inverse akrasia is nothing very exotic in itself. Because it is simply distinguished from other forms of akrasia by the nature of the outcome, whenever someone ends up akratically doing the right thing for whatever reason—fear, habitual inattention, a brute tendency to social conformity, or whatever—they will display it. As a result there is not going to be any particular unity in the category, and it is not going to be theoretically very interesting. Huck becomes interesting because Arpaly and Schroeder see him as instantiating two further features, and these are not just about the outcome, buit are much more subjective. First, Huck acts for the *right reasons*, more specifically, the right *moral* reasons. It is not just that he acts *in accord* with those reasons; it is that they move him. There will thus be, at the very least, a counterfactual dependence between the moral reasons and Huck's behaviour. If the reasons had been different, Huck would have acted differently.

Second, this process of tracking the right moral reasons is quite opaque to Huck. As Arpaly puts it in a later piece:

> [M]y point is not simply that Huckleberry Finn does not have the belief that his action is moral on his mind while he acts, but that he does not have the belief that what he does is right anywhere in his head—this moral insight is exactly what eludes him. (Arpaly 2003, 77)

These points are independent. One could (just about) imagine an agent who exhibited inverse akrasia, who moreover acted akratically for the right moral reasons, but who fully understood what was going on. Such an agent would judge that it would be best to act immorally, would instead do the moral thing, because it was the moral thing, in the knowledge that it was the moral thing, even while judging that so acting was, all things considered, a mistake.[9] That is not Huck on Arpaly and Schroeder's characterization. They think that Huck has no conception that what he is doing is right.

On their view then, Huck is someone possessing what I will call *moral blindsight*. Textbook *visual* blindsight is the condition in which agents can functionally see, but apparently have no awareness that they are seeing, with the result that they can unwittingly avoid the physical pitfalls of the world. In parallel, *moral* blindsight is the condition in which an agent is able to see the moral facts with no awareness of what they are seeing, with the result that they can unwittingly avoid the moral pitfalls of the world. It is this feature, rather than simple inverse akrasia, that makes Huck so distinctive and so theoretically important on Arpaly and Schroeder's account.

I speak of textbook visual blindsight; in fact there has been some discussion of whether the characterization I have offered is really what is going on in the clinical cases that have been studied, or whether in fact agents do have some visual awareness, but are unsure of whether they are genuinely seeing or not (Phillips 2021). In similar vein, I want to challenge the idea that what Huck displays is moral blindsight. Further, I deny that Huck's actions are even straightforwardly akratic. I will argue that he is morally uncertain and confused—unsure of whether he is morally insightful or not—but that is very different.

AKRASIA AND WEAKNESS OF WILL

Before starting on the substantial argument, let us first get a potential equivocation out of that way. Both Bennett, and Arpaly and Schroder, discuss the phenomenon at times as one of weakness of will, and this is indeed a common translation of the Greek term 'akrasia'. But in ordinary English, 'weakness of will' is often used just to mean irresoluteness: an agent shows weakness of will when they abandon a resolution without good reason, that is, without a good reason relative to the standards governing resolution maintenance (Holton 1999). An agent can show such irresolution without being akratic.

---

[9] This assumes, following (Austin 1956) and (Davidson 1970), that akrasia itself is not essentially a moral failing:, the agent's judgement of the best course of action need not be that it is the most moral course of action.

It is plausible that Huck shows irresolution in this way. Indeed, he plausibly shows it twice over. On first meeting the fleeing Jim, Huck promises not to betray him: 'I said I wouldn't and I'll stick to it. Honest injun I will. People would call me a low down ablitionist and despise me for keeping mum—but that don't make no difference. I ain't agoing to tell'. So in deciding to break his resolution not to tell—a resolution of which Jim reminds Huck as he paddles off to reveal Jim's presence to the slave hunters—we might well think that he has shown weakness of will. That, of course, is controversial, depending on whether we think that the pangs of conscience make it reasonable to revise his resolution, but a good case can be made for it.

Plausibly Huck shows weakness of will a second time when he fails to go through with this second resolution, and fails to turn Jim in. Here is Huck's own account of what happens:

> I didn't answer up prompt. I tried to, but the words wouldn't come. I tried, for a second or two, to brace up and out with it, but I warn't man enough—hadn't the spunk of a rabbit. I see I was weakening; so I just give up trying, and up and says: 'He's white.'

Huck thus thinks he has shown weakness of will, and I think we might well agree with him. Certainly before he came to act he had judged that turning Jim in was right; resolving to act on that belief gave him a rare period of equanimity. 'I felt easy and happy, and light as a feather, right off. All my troubles was gone'. And Twain adds, in the reading version: 'O! it was a blessed thought! I never can tell how good it made me feel—cuz I knowed I was doing right now'. Huck's resolve does not, however, survive his attempt to act upon it. Admirable though his hasty revision of his plan is, we might doubt that it stems from any new information or insight of the kind that should rationally have him reconsider. If that's right, we can conclude it is a revision that, by the standards of resolution maintenance, he should not have made. The standards of decency and humanity say otherwise; but they are not what are in play when we are assessing weakness of will.

But if Huck is weak-willed, is he akratic? Is this a case of moral blindsight? That depends on whether, at the time he acted—at the time, that is, he lied to the slave hunters—he believed that turning Jim in was the right thing to do. The answer to that is far less obvious. In trying to assess it, I am going to start with an argument from Zoë Johnson King (2020) to the effect that Arpaly and Schroeder's account can't be right, and—in my terms—that moral blindsight is impossible. While there is much that I agree with in what she says (especially her view on what is really happening in Huck's psychology, p. 196), I don't think that her argument quite works, but discussing it will set the stage.

The crux of Johnson King's argument is that, if Huck is acting without an awareness of what he is doing, then his act must be accidental; and that accidental acts cannot reflect well on the agent in the way that Arpaly and Schroder say that Huck's reflects well on him. Such acts will lack 'moral worth'. In our terminology then, moral blindsight is impossible, since acts that are blind are not moral.

Crucial to Johnson King's argument is a premise that I'll call *The JK Principle*:

> For all types of acts *A*, someone accidentally *A*s if she has no idea that she is performing an act of type *A* when she does so. (Johnson King 2020, 188)

That principle, together with Arpaly and Schroeder's contention that Huck has no idea what he is doing, does bring the conclusion that Huck's act lacks moral worth, given the further plausible premise that accidental actions have no moral worth.

I think, though, that the JK Principle is too strong. How does one steer a bicycle? Most cyclists think that they turn the handlebars in the direction in which they intend to go. But while that is a correct account for a tricycle it is wrong for a bicycle. On a bicycle, simply turning the handlebars in the direction one intends to go will cause the machine to skid. To avoid the skid one needs to lean into the direction of the proposed turn, and the normal way of achieving that, highly counterintuitively, is to briefly turn the handlebars in the *opposite direction to the way one intends to turn*—countersteering as it is called. Children learn to do this when they learn to ride, without realizing that this is what they are doing. One result of this, as Ryle (1940) noted, is that experienced cyclists who mount a tricycle often initially find it very hard to steer, since they intuitively countersteer while trying to lean the machine, which of course they cannot do. The result is that the machine goes, disconcertingly, in the opposite direction to that intended.[10] Countersteering is surely an action, one that most cyclists perform without any idea that they are doing it. Yet it is not accidental. It is learned; it is reliably performed when needed. So we have a counterexample to the JK Principle.

Might this provide a model for what is happening with Huck? Might his unwittingly virtuous actions be parallel to those of the cyclist who is unwittingly countersteering? That doesn't seem very plausible. Cyclists don't notice themselves countersteering; it happens too fast. Ryle characterizes such actions as reflexes.

---

[10] Another is that, once a bicycle is ridden closely parallel to a kerb, it is very hard to steer away from it, with the result that the rider feels that some malign force is keeping them in the gutter. The basic facts of all this have been known for a long time: the Wright brothers were well aware of it, and used the idea of rolling into the turn when designing the control systems for their aircraft. But knowledge of the details, involving the interaction of centrifugal and gravitational forces acting on the bicycle, together with those applied by the rider to the handlebars, came surprisingly late. The classic and amusing empirical studies were done by David E H Jones (Jones 1970), later to gain fame as the author of the Daedalus columns in *The New Scientist* and *Nature*; for a bit more theory, see (Frajans 2000).

There may be some morally significant actions that are like that: microaggressions —or, on the positive side, and rather less discussed, microaffirmations—may be a case in point. These seem like genuine actions that may be brought into awareness, but much of the time they go unnoticed even by the agents themselves; in that way they are like countersteering. In contrast, Huck certainly notices when he plans to turn Huck in, and then notices when he decides against it. That is why his conscience can get to work.

More broadly: if Huck exhibits, as I think he does, a consistent pattern of behaviour towards Jim that shows he treats him as a moral equal, someone who deserves respect and consideration, someone who cannot be bought or sold or owned, a pattern that, importantly, is reciprocated by Jim; and if Huck also thinks that this behaviour is simply unjustified, is grounded on a moral mistake; then there would be something quite baffling, baffling to Huck himself, about what he is doing. How can he explain the relationship and how he behaves? He could think that he is simply moved by the kind of mere sympathy that Bennett talks of; that is, he could take Bennett's view of his own actions. But as we have seen, if Huck is genuinely responding to moral reasons, that won't be good enough explanation; and that too is something that, by reflecting on his actions, or even on how he would act, will be available to Huck. He has a relationship of trust with Jim, and Jim with him, and he knows it.[11] That should give him grounds for thinking that there is something good about protecting him.

Might Huck fail to realize what he is doing because he is self-deceived? There is plenty of evidence that people do go in for quite elaborate self-deception, and that this is something that regularly happens in the moral sphere. But the documented cases involve agents performing acts of dubious morality, while convincing themselves that what they are doing is morally fine. We can readily see why that happens. Most agents, in fact pretty much all who fall short of outright psychopathy, want to maintain a picture of themselves as moral, a picture that will be presented to others, and, just as importantly, to themselves. (See Holton 2023 for a range of cases and discussion of the likely mechanisms involved.) With moral blindsight though the suggestion is that we have the opposite phenomenon: the agent is acting well, but fails, despite the evidence, to see that they are. Could that be explained by self-deception? We certainly need some account of why Huck might be motivated to make himself look worse.

So, in the particular case of Huck, we have good *prima facie* reasons for doubting that he straightforwardly believed that in protecting Jim he was doing the wrong thing. Can we turn this into a deductive argument? I doubt that we can. I doubt that we can formulate a revised version of the JK Principle that would demonstrate that moral blindsight is *impossible*. It is rather that there are factors that would make moral blindsight a surprising phenomenon, demanding an explanation on a

[11] There has been a tendency in the philosophical literature to leave Jim out of it. As a literary corrective see (Everett 2024), though in the version of the episode given there (Ch. 13) Huck feels no obvious obligation to turn Jim in.

case-by-case basis. We lack any such explanation in the case of Huck. That should make us suspicious. Let us look then to see quite what Twain thought was happening.

TWAIN ON HUCK

As Huck sets out to tell on Jim, he is disturbed by Jim's cries of affection, and of Jim's confidence that Huck will do the right thing. In the published version of the novel Huck says:

> I was paddling off, all in a sweat to tell on him; but when he says this, it seemed to kind of take the tuck all out of me. I went along slow, then, and I warn't right down certain whether I was glad I started or whether I warn't.

In the revised reading copy, Twain makes clear that this is not just an issue of whether he was *glad*, but of what he thinks he *morally ought* to do, adding:

> It kind of all unsettled me, and I couldn't seem to tell whether I was doing right or doing wrong.

That is as clear a statement as one can get that Huck is not sure of what he ought to do. He is carried on, not by conviction, but by the pure force of his resolution. In the published version:

> Well I just felt sick. But I says, I *got* to do it—I can't get *out* of it.

Between those two sentences, Twain added in the reading copy:

> I just let go and give up.

So, on Twain's own account, this is not a case of akrasia. At the time he acts, Huck does not go against his judgment of what is best. He has no judgment of what is best; he is confused and conflicted, cannot tell whether in carrying out his plan he is 'doing right or doing wrong', and has 'let go' on trying to figure it out.[12]

What, then, is Huck's struggle a struggle between? Let's go back to the idea which Twain raised in his introduction to the reading passage, that it is a collision between 'a sound heart and a deformed conscience'. Here conscience does not give voice to the genuine demands of morality. Nor does it even give voice to what Huck *thinks* are the genuine demands of morality. Rather, as Huck himself comes to realize, it makes a claim to moral authority, but one that is highly fallible—most obviously because it makes inconsistent demands.

---

[12] See (Freedman, 1997) for a similar reading.

The issues here are delicate, since at this point in the novel Huck is still somewhat in thrall to his conscience, and his views are far from stable. Here is how, in the published version, he describes his return to Jim, having failed to turn him in:

> I got aboard the raft, feeling bad and low, because I knowed very well I had done wrong, and I see it warn't no use for me to try to learn to do right; a body that don't get started right when he's little, ain't got no show - when the pinch comes there ain't nothing to back him up and keep him to his work, and so he gets beat. Then I thought a minute, and says to myself, hold on—s'pose you'd a done right and give Jim up; would you feel better than what you do now? No, says I, I'd feel bad—I'd feel just the same way I do now. Well, then, says I, what's the use you learning to do right, when it's troublesome to do right and ain't no trouble to do wrong, and the wages is just the same? I was stuck. I couldn't answer that. So I reckoned I wouldn't bother no more about it, but after this always do whichever come handiest at the time.

Bennett interprets this as a rejection of morality altogether, and there is something of that hard-bitten cynicism in Huck (and indeed in Twain himself). But Huck's talk of right and wrong here can equally be seen as an inverted commas use, denoting right and wrong as conscience presents it. And when Huck talks of how he would *feel*, he might be talking about what his conscience would say were he to have turned Jim in. That this is the right interpretation of the passage is strongly suggested by a further passage written in May of 1896 for the South African leg of the reading tour.[13] Here Twain makes it clear that, on his return to the raft, Huck realizes his conscience is highly fallible exactly because it is so inconsistent:

> Old Jim was saved; & just as I was going to be glad, my con shoved in, just in the same old way & begun to skin me & scorch me & blister me because I'd done wrong, till I just couldn't *stand* it no more. So I just turned on it I says—Looky here I says, Things is gone jist a little too fur betwixt you & me, I says, & I'm gitting powerful tired of it. … I don't want no such thing around as a conscience… You ain't wanted, you ain't welcome, you ain't no use to me. I never see such a low-down troublesome cuss, I says. It don't make no difference what a person does, you ain't ever satisfied and you is as free as if you owned the whole layout. If I'd a give Jim up you'd a kep me awake a week mournin' about it; and now you're gittin' ready to try to keep me awake another week because I didn't give him up … I wouldn't be as ignorant as you for wages. You don't know right from wrong, you ain't got no judgment, you ain't got no sense about anything—you ain't no good but just to lazy around, find fault and keep a person in a sweat… I said them very words. I talked perfectly plain; I warn't in no humor to mince things. Them was my sentiments then, & they're my sentiments *yet*. (Twain, 1896)

This is not to abandon morality; it's to say that conscience is no guide to it: conscience 'don't know right from wrong'. And that idea, that conscience is unreliable since it will complain whatever one does, is made very clear in a later section of the published novel:

---

[13] It's unclear whether these are new reflections, or whether Huck is recording the things that he had been saying in the performances earlier in the tour.

> It don't make no difference whether you do right or wrong, a person's conscience ain't got no sense, and just goes for him anyway. If I had a yaller dog that didn't know no more than a person's conscience does, I would pison him. It takes up more room than all the rest of a person's insides, and yet ain't no good, nohow. (Ch xxxiii, 290)

Those sentiments are very close to Twain's own. 'Mine was a trained Presbyterian conscience', he wrote in his *Autobiography* 'and knew but one duty—to hunt and harry its slave upon all pretexts and on all occasions; particularly when there was no sense or reason in it.' (Twain 1906, 158) This idea of the conscience incorporating the result of bad training comes up again in the introductory passage to the reading tour. Talking of how slavery had seemed 'natural enough' to him in his boyhood, Twain wrote:

> It shows how that strange thing, the conscience—that unerring monitor—can be trained to approve any wild thing you want it to approve if you begin its education early and stick to it. (Twain 1895, 807)

So did Twain think that we should give up on conscience altogether? In his short story 'The Facts Concerning the Recent Carnival of Crime in Connecticut' the narrator complains:

> All consciences *I* ever heard of were nagging, badgering, fault-finding, execrable savages! Yes; and always in a sweat about some poor little insignificant trifle or other—destruction catch the lot of them *I* say! I would trade mine for the small-pox and seven kinds of consumption, and be glad of the chance. (Twain 1876, 646)

He succeeds in killing his conscience—personified as a 'shriveled, shabby dwarf'—but while he is happy with the outcome, it is scarcely a good one:

> Nothing in all the world could persuade me to have a conscience again. I settled all my old outstanding scores, and began the world anew. I killed thirty-eight persons during the first two weeks—all of them on account of ancient grudges. I burned a dwelling that interrupted my view. I swindled a widow and some orphans out of their last cow, which is a very good one, though not thoroughbred, I believe. I have also committed scores of crimes, of various kinds, and have enjoyed my work exceedingly, whereas it would formerly have broken my heart and turned my hair gray, I have no doubt. (Twain 1876, 650)

Consciences clearly have their uses; they are just highly fallible.

This might raise a worry though. Twain had described Huck's dilemma as involving a collision between conscience and the heart. Suppose that Huck had instead turned Jim in, and his conscience had still bitten him. How could we now understand the conflict?

The answer, I think, comes from taking seriously Twain's talk of conscience as a *monitor*. There are two ways that people have thought of conscience. One is as a source of moral knowledge: for instance, Cardinal Newman thought of it as some internalization or echo of God's voice—the 'aboriginal vicar of Christ'—speaking within us. (Newman 1975, 73). The other is as an instrument of self-policing, a faculty whose job is to monitor potential behaviour, judging it against a moral code that the agent independently possesses. (For a nice discussion of the distinction see Skorupski 2010.)

Twain's talk suggests the second: conscience monitors our behaviour against a moral code that has been independently established. What is distinctive about his view is that it is indiscriminate in the moral code it selects, ready to make use of any of the inconsistent codes that have somehow lodged in the speaker, whether they are true or not. When it rebukes Huck for protecting Jim, it uses the code that stems from a faulty education. Had he given Jim up, it would use the code that came from the findings of his sound heart. And, strikingly, this confirms the idea that the findings of his sound heart must be recognizably moral even to Huck himself. For it is only if this is so that they would be available for his conscience to exploit. It is not a superficial feature of the story that Huck feels himself morally conflicted; it is essential to the characterization of his conscience.

It is an interesting question whether the moral psychology that I have been arguing is there in *Huckleberry Finn* corresponds to Twain's own final considered view of things. As the large scholarly literature demonstrates, it is hard to pin him down. Much has been made of his familiarity with Edward Lecky's book *History of European Morals*. That is, at least from a modern perspective, a strange work, contrasting sentimentalism, not, as one might expect, with rationalism, but with utilitarianism. Lecky defends sentimentalism. Twain, in his marginal notes, is critical, pushing a more utilitarian, or at times egoist, position, something that comes out again in his later, more explicitly philosophical dialogue, *What is Man?* But as Gregg Camfield (1992, 1994) in particular has pointed out, there is much in Twain that is sympathetic to sentimentalism.[14] And there is plenty in Lecky that could have given the underpinnings of the kind of position that is implicit in *Huckleberry Finn*, while also preserving a place for something more utilitarian. Here, for instance, is Lecky's discussion of Hume:

> Hume maintained that the peculiar quality of virtue is its utility, but that our affections are purely disinterested, and that we arrive at our knowledge of what is virtuous by a moral sense implanted in our nature, which leads us instinctively to approve of all acts that are beneficial to others. Expanding a pregnant hint which had been thrown out by Butler, he laid the foundation for a union of the schools of Clarke and Shaftesbury, by urging that our moral decisions are not simple, but

[14] Caution is needed here, for sentimentalism as it was understood in the late 19th c., and indeed as it is still understood in many literary discussions, is rather different to the way it was understood by the English Moralists, and by most philosophers today. See the wide-ranging discussion in (Camfield 1994).

> complex, containing both a judgment of the reason, and an emotion of the heart. (Lecky 1865, 76–7)

MORAL BLINDSIGHT ELSEWHERE?

I have argued that Huck does not show moral blindsight when he saves Jim from the slave hunters, and that he is not even akratic. But that is just one case. If that fails, can't we just turn to a different one? What of the other cases that Arpaly and Schroder propose, Neoptolemus and Schindler? They do not say whether they take them to illustrate moral blindsight, or mere inverse akrasia; but since they are supposed to provide cases parallel to that of Huck, I assume that they are meant to demonstrate the former. Again I think that they fail, and that as the failures mount up, so our scepticism about the existence of moral blindsight should increase. I take them in turn.

Neoptolemus

Start with the story, or at least as much as is needed for our purposes. Neoptolemus, a young Greek warrior, has been persuaded—by Odysseus no less! —that he must deceive Philoctetes in order to obtain Philoctetes' famous bow. Neoptolemus is at first shocked by the idea of using trickery. He is a well brought up young man, son of Achilles, and knows how dishonourable such deception would be. But he is persuaded that the stakes are so high—victory against the Trojans, fame for himself—that, reluctantly, he agrees. He does lie to Philoctetes and obtains his bow, but then, ashamed at what he has done, confesses the truth, and ultimately gives him the bow back.

What are we to make of this? Picking up on the two brief discussions in the *Nicomachean Ethics* (7.2, 1146a 18–21; and 7.9, 1151b 18–22) Arpaly and Schroder take the point at which Neoptolemus admits his lie to Philoctetes as a case of inverse akrasia: he does the right thing, they say, believing that he is doing wrong.

Even if this is right, it is certainly not a case of moral blindsight. Neoptolemus is perfectly well aware that it is dishonourable to lie. He does so with the greatest reluctance, only because he thinks there is so much to be gained. 'All right, I'll do it' he says when he agrees to the trickery, 'I'll set aside my sense that it's wrong' (125). So this is not a case in which he finally does the morally right thing oblivious to the fact that that is why he is doing it. He knows that he admits his lie and returns the bow because that is the honourable thing to do.

Is it a case of inverse akrasia though, where we understand him to be akratic if he fails to do what he believes, all things considered, to be best, whether or not this is honourable? In fact Aristotle himself claims that Neoptolemus is not akratic (1151b 18). Partly this is for largely terminological reasons. I mentioned above that akrasia is now normally defined as a subjective state, characterized in terms of

whether the agent believed that what they did was wrong. It seems though that Aristotle's own conception was partly objective: an akratic act is a *bad* act that the agent believes is wrong (see the discussion in Cagnoli 2018). If that is right, then inverse akrasia is ruled impossible by definition.

That in itself is, of course, no rebuttal of Arpaly and Schroeder, given their now-standard, subjective, definition of akrasia. There is though a deeper reason why Aristotle would not have judged Neoptolemus to be akratic even had he accepted a purely subjective definition. Aristotle describes Neoptolemus as abandoning his belief that he should lie (1151b 18–20). That might make him weak willed in the sense discussed earlier (though perhaps not if his revision was justified); but it suggests he wasn't akratic, since he no longer believed it was right to lie. Sophocles' text brings this out in Neoptolemus's growing despair before he finally tells the truth. As Neoptolemus concludes: 'I don't know how to express my confusion … this is what I have been agonising over all this time' (911–32).

Like Huck, Neoptolemus was in a state of conflict. His conflict though was very different since it was between his commitment to honour, and an undertaking made to Odysseus on grounds of expediency. It was his fully intentional choice of the former that led Aristotle to see him as praiseworthy.[15] He was, indeed, in an uncomfortable state, but not one of akrasia, let alone of moral blindsight.

Schindler

Finally, the case of Oskar Schindler. Here is what Arpaly and Schroeder say:

> A more contemporary example of inverse akrasia can be found in the character of Oscar Schindler, as portrayed in Steven Spielberg's movie, *Schindler's List*. According to the movie, Schindler started his enterprise—giving employment to Jews in a Polish ghetto under Nazi rule—as a selfconfessed egoist, whose goal was to profit as much as possible from the war. Until rather late in the plot, Schindler's conscious convictions continue to be those of a cynical capitalist. Even as he finds himself doing extremely uneconomical things in order to save the lives of people for whom he supposedly does not care, he looks upon these actions as weakness of will, or rationalizes them after the fact the way one might the breaking of a diet. (pp. 162–3)

Perhaps that is one impression that can be gained from the film (though it wasn't mine). But let's examine the reality behind it more carefully. Spielberg's film is based on the 1982 book *Schindler's List*, by Thomas Keneally. That was marketed as a novel, but it is really a carefully researched piece of history, based on a series of

[15] Again see (Cagnoli 2018) for a very helpful discussion of Aristotle's thought. Cagnoli's contention is that akrasia should be characterized relative to an agent's underlying *prohaireseis*, and not relative to any more superficial judgments that they might make. So on this view, given that his *prohaireseis* committed him to honourable acts, Neoptolemus would not have been akratic, even if he had continued to judge that the best thing to do here would have been to lie. See also (Liberti 2024) for another very helpful discussion of the case.

interviews with many of those who had been involved in working for Schindler in his enamelware (for an account of the writing of the book see Keneally 2007). According to the book Schindler was indeed powerfully motivated to make money. He had joined the Nazi party, and would wholeheartedly fraternize with Nazis to achieve that. But did he really not care for those Jews to whom he gave employment? Here is what Keneally says of Schindler's witnessing of the SS evacuation of the Jewish quarter of Kazimierz in December 1939. At this point Schindler had recently arrived in Krakow, and had not yet purchased the enamelware factory that he subsequently ran there:

> The SS invasion of Kazimierz would, however, arouse in Oskar a fundamental disgust; not yet one which would impinge too visibly on the level at which he made his money, entertained women or dined with friends, but one that would, the clearer the intentions of the reigning power became, lead, obsess, imperil and exalt him. (Keneally 1982, 64)

The picture one gets is of someone who, throughout all of his time running the enamelware factory, was deeply committed to mitigating the antisemitic evils of the regime. There is no doubt that he was a hedonistic and self-centred individual, and doubtless that sometimes got the in way. But he never saw his actions to save his employees—and those who were not his employees but could be saved by being made so—as anything other than what they were: morally required actions in the face of horror. Again these are not cases of akrasia at all, and certainly not of moral blindsight.

So Huck, Neoptolemus, and Schindler should each be understood differently. But none of them exhibits inverse akrasia, let alone moral blindsight. We are left without any convincing examples. It is time to draw some conclusions.

conclusions

This paper has had two main themes. First, that cases of moral blindsight are much rarer than the philosophical discussions have made them seem, and that the examples we have been offered are actually of something else. Second that, properly understood, Huck still presents a very interesting case, but it is one of the unreliability of conscience as a guide to the moral.

On the first theme: to reiterate, I do not doubt that there are plenty of cases of mere inverse akrasia. People often akratically abandon mistaken beliefs about what is best, and end up doing the right thing. My focus has instead been on the much more interesting case of moral blindsight—cases where one does something right for the right reason but without any awareness that that is what one is doing. Even here, I doubt that there is any argument to show that moral blindsight is *impossible*. I think though that there are two sorts of consideration that should lead us to conclude that it will be rare, considerations that I have already alluded to, but that

deserve spelling out. The first come from contingent claims about what people want; the second from somewhat deeper claims about the nature of moral attitudes.

Start with the contingent claims. People have a general tendency to try to understand themselves in a way that puts a good gloss on what they do; they are engaged in 'self-presentation' as psychologists are apt to call it. If an agent is in fact tracking and being moved by moral reasons, then, in the absence of special factors obscuring them, those reasons will be to some degree visible to the agent; and the agent will be strongly motivated to tease them out in order to present them to themselves, and to others, as acting well. This is not to deny that there will be other conflicting considerations present; but the accurate moral picture is unlikely to be crowded out altogether. We should not expect genuine moral motivation to be completely hidden. There is also a further factor that is important here. If agents find themselves doing, or about to do, something that antecedently they would have judged undesirable, then there are two ways that they can regain harmony: they can change their behaviour, or they can change their judgment. The former might look like the only legitimate course, but the latter is in fact very common.[16] It is the tendency that underlies the phenomenon of cognitive dissonance. In many cases it is a force for corruption: Baumeister (1996) gives many examples of those who, engaged in committing the atrocities of the modern world, come to change their moral judgments to the view that what they are doing is right. But in the cases we are considering the force may be for the good. The pull of sympathy that changes an agent's behaviour may sway their moral views. That might lead to corruption; but, suitably constrained, it may be the beginning of moral insight.

This takes us to the second sort of consideration, which, is, I think, philosophically more interesting. It goes to the heart of the nature of our moral sensibility. Moral blindsight would require us to have certain attitudes that would keep us morally on track; in the cases we have been considering, these are attitudes of sympathy. For there to be genuine moral blindsight then, these must be quite independent of our explicit moral beliefs. But as I mentioned when discussing the case of Himmler, I am sceptical that we can make a clean break between an agent's moral sympathies, broadly understood, and their moral beliefs. Why might one think this? One route would be to give up on cognitivism altogether, denying that apparent moral beliefs really are beliefs at all, and instead identifying them with sympathies. I do not wish to take that path, for familiar reasons. However, even retaining cognitivism, it is possible to think that our sympathies are more intimately connected with our moral beliefs than just as sources of evidence. A parallel with visual perception may be helpful. Visual perception doesn't simply provide evidence upon which a subject forms beliefs about the world. It *coerces* belief, in a way that is typically immediate, direct and involuntary.[17] Standardly, seeing is believing. Of course it can be defeated. We can believe that the lines in a Muller-Lyre illusion are of the same length, because we have measured them. But we remain conflicted: they still look to

[16] I discuss this, under the heading of 'judgment shift', in (Holton 2009) Ch. 5.

[17] I take the terminology here from Mike Martin's very helpful discussion: see (Martin 2002, 389–92)

be different. I suggest that sympathy, and associated reactive attitudes, stand in much the same relation to our moral beliefs. Sympathetic feelings coerce us into the associated moral beliefs; the response is immediate, direct and involuntary. Again the response can be over-ridden. We can think that we suffer from moral illusions on a par with visual ones, and so do our best to discount the sympathy. Again though the conflict does not go away.

The analogy also helps us to see how Huck's case differs from visual blindsight, even when the latter is understood in the rather deflationary way that Phillips understands it. On Phillips' view, subjects with visual blindsight have such degraded visual perceptions that, even if the perceptions are accurate, they lack the confidence to trust them; the perceptions thus completely lack the normal coercive power. In contrast, Huck's inputs are not too meagre. They are too many—coming from sympathy and from earlier testimony—all with a great deal of power, and pulling in diverse directions. That is what leads to the conflict he feels.[18]

We might wonder whether there are other cases in which subjects do have degraded sympathetic input. Psychopathy presents an obvious possibility. So might that lead to genuine moral blindsight? It is an interesting possibility, and deserves some exploration, but I rather doubt it. Unlike in the case of visual blindsight, psychopaths have no lack of testimonial input. When they do show moral motivation the empirical literature suggests it is more likely to exhibit an over-rigid reliance on learned rules.

What, finally, of conscience? I think that there is much to be said for Twain's view that it is just one voice, though a highly salient one, in the cacophony of the interior: a voice that judges our own actual or potential conduct, is tightly connected to the emotions of guilt and regret, but is highly fallible in its moral deliverances.

That is a much more modest role than is assigned to conscience than by many of the philosophers who have written on it, who see it as authoritative. In its phenomenology it harks back to an old conception of the conscience as something that bites or pricks.[19] A similar idea is there in Kant, who though he does see it as authoratative, stresses its limited role as an assessor of one's own potential actions (see Hill 1998 for a concise discussion; and see also Ryle 1940 for a discussion of why the focus of conscience is on one's own acts). But the philosopher who perhaps comes closest to Twain in combining these features with an awareness of the fallibility of conscience is Mill. Conscience is, he writes,

---

[18] Liberti (2025) sees the action of both Huck and of Neoptolemus as stemming from an underlying moral self-identity. This strikes me as much more plausible for Neoptolemus than for Huck. One could easily imagine Huck's action as having gone the other way. What is not so easily imagined is the idea that Huck would not have been conflicted.

[19] The *ayenbite of inwyt*—the 'again biting of the inner wit'—in the colourful vernacular of Dan Michel of Northgate, a 14c. monk (Morris 1866).

> all encrusted over with collateral associations, derived from sympathy, from love, and still more from fear; from all the forms of religious feeling; from the recollections of childhood and of all our past life; from self-esteem, desire of the esteem of others, and occasionally even self-abasement. (Mill 1863 p.41)

His conclusion echoes, in rather wordier form, Twain's contention that conscience can be 'trained to approve any wild thing you want it to approve':

> Unhappily it is also susceptible, by a sufficient use of the external sanctions and of the force of early impressions, of being cultivated in almost any direction: so that there is hardly anything so absurd or so mischievous that it may not, by means of these influences, be made to act on the human mind with all the authority of conscience. (p. 44)

It is an approach that, I submit, is ripe for some more development.

BIBLIOGRAPHY

Arpaly, Nomy 2000: 'On Acting Rationally Against One's Best Judgment' *Ethics* 110, 488–513.

— 2003: *Unprincipled Virtue* (New York: Oxford University Press).

— 2015: 'Huckleberry Finn Revisited: Inverse Akrasia and Moral Ignorance'. In R. Clarke et al., *New Essays on Moral Responsibility*. (Oxford: Oxford University Press) 141–55.

Arpaly, Nomy, and Timothy Schroeder 1999: 'Praise, Blame and the Whole Self', *Philosophical Studies* 93, 161–88.

Austin, J.L. 1956: 'A Plea for Excuses' *Proceedings of the Aristotelian Society* 57, 1–30.

Baumeister, Roy 1996: *Evil* (New York: W. H. Freeman).

Bennett, Jonathan 1974: 'The Conscience of Huckleberry Finn' *Philosophy* 49, 123–34.

Camfield, Gregg 1992: '"I Wouldn't Be as Ignorant as You for Wages": Huck Talks Back to His Conscience', *Studies in American Fiction* 20, 169–75.

— 1994: *Sentimental Twain: Samuel Clemens in the Maze of Moral Philosophy* (Philadelphia: University of Pittsburgh Press).

Cagnoli Fiecconi, Elena 2018: 'Aristotle on the Structure of Akratic Action', *Phronesis* 63, 229–56.

Clarke, Steve 2020: 'Huckleberry Finn's Conscience: Reckoning with the Evasion' *The Journal of Ethics* 24, 485–508.

Currie, G., Ferguson, H., Frascaroli, J., Friend, S., Green, K., & Wimmer, L. (2024): 'Learning from Fiction'. In A. James, A. Kubo, & F. Lavocat (eds.) *The Routledge Handbook of Fiction and Belief* (New York & Abingdon: Routledge) 126–38.

Davidson, Donald 1970: 'How is Weakness of the Will Possible?' In his *Essays on Actions and Events* (Oxford: Clarendon Press) p. 21-42.

Dennett, Daniel 1995: 'The Unimagined Preposterousness of Zombies', *Journal of Consciousness Studies* 2, 322–6

Driver, Julia 2001: *Uneasy Virtue* (Cambridge; Cambridge University Press).

Everett, Percival 2024: *James* (New York: Doubleday).

Farjans, J. 2000: 'Steering in bicycles and motorcycles', *American Journal of Physics* 68, 654–9

Fishkin, Shelley Fisher 2025: *Jim: The Life and Afterlives of Huckleberry Finn's Comrade* (New Haven: Yale University Press).

Freedman, Carol 1997: 'The Morality of Huck Finn' *Philosophy and Literature* 21, 102–13.

Gehrman, Kristina 2018: 'The Character of Huckleberry Finn' *Philosophy and Literature* 42, 125–144

Hill, Thomas, 1998: 'Four Conceptions of Conscience' reprinted in *Human Welfare and Moral Worth* (Oxford: Oxford University Press, 2002) 277–309.

Hursthouse, Rosalind 2001: *On Virtue Ethics* (Oxford: Clarendon Press).

Holton, Richard 1999: 'Intention and Weakness of Will' *Journal of Philosophy* 96, 241–62.

— 2009: *Willing, Wanting, Waiting* (Oxford: Oxford University Press).

— 2022: 'Self-Deception and the Moral Self'. In Manuel Vargas and John Doris (eds.) *The Oxford Handbook of Moral Psychology* (Oxford: Oxford University Press) 262–84.

Johnson King, Zoë, 2020: 'Accidentally Doing the Right Thing' *Philosophy and Phenomenological Research* 100, 186–206.

Kaufman, Will 2006: 'Mark Twain's Deformed Conscience' *American Imago* 63, 463–78.

Keneally, Thomas 1982: *Schindler's Arc* (London: Hodder and Stoughton).

— 2007: *Searching for Schindler* (Sydney: Knopf).

Lecky, Edward (1869): *History of European Morals From Augustus to Charlemagne* Vol. 1. (London: Longmans).

Liberti, Matilde 2024: 'Neoptolemus and Huck Finn Reconsidered. Alleged Inverse Akrasia and the Case for Moral Incapacity' *Journal of Value Inquiry* (No volume number or page references yet available).

Longerich, Peter 2011: *Heinrich Himmler: A Life* (Oxford: Oxford University Press).

Markovits, Julia, 2010: 'Acting for the Right Reasons', The Philosophical Review, 119, 201-242.

Martin, M.G. F. 2002: 'The Transparency of Experience' *Mind & Language* 17, 376–425.

Mill, J. S. 1863: *Utilitarianism* (London: Parker, Son, and Bourn).

Morris, Richard 1866: *Dan Michel's Ayenbite of Inwyt* (London: The Early English Text Society).

Newman J. H. 1875: *A Letter Addressed to His Grace the Duke of Norfolk* (New York: The Catholic Publication Society).

Phillips, Ian 2021: Blindsight is qualitatively degraded conscious vision. *Psychological Review* 128, 558–84.

Proust, Marcel, 1954. *Le temps retrouvé*. In P. Clarac and A Ferré (eds.) *A la recherche du temps perdu* Vol III (Paris: Gallimard) 689–1048.

Skorupski, J. M. (2010). 'Conscience'. In J. Skorupski (ed.), *The Routledge Companion to Ethics* (London: Routledge) 550–61.

Sophocles, 2009: *Philoctetes* trans Judith Affleck (Cambridge: Cambridge University Press).

Twain, Mark 1876: 'The Facts Concerning the Recent Carnival of Crime in Connecticut' *The Atlantic Monthly* 37, 641–50.

— 1884: *Huckleberry Finn*. In Walter Blair and Victor Fisher (eds.) *The Works of Mark Twain*, Vol 8 (Berkeley and Los Angeles: University of California Press, 1988).

__ 1895 Notebook 35, Typescript, pp. 34–6. In (Twain 1884), 806–7.

— 1896 Notebook 38, Typescript, pp. 7–8. In (Camfield 1992), 171.

— 1897 *Following the Equator* (New York: Doubleday).

— 1906: *Autobiography of Mark Twain*, vol. 1, edited by Harriet Elinor Smith (Berkeley: University of California Press, 2010).

Wilkes, K. 1993: *Real People: Personal Identity Without Thought Experiments* (Oxford: Oxford University Press).