Archimedes in the lab: Can science identify good moral reasoning?

Regina Rini and Tommaso Bruni

Abstract: Some ethicists try to settle moral disagreement by ruling out particular types of moral reasoning on the basis of cognitive scientific evidence. We argue that the cognitive science of reasoning is not well-suited to this Archimedean role. Through discussion of several influential research programs, we show that such attempts tend to either fail to be Archimedean (by assuming controversial moral views) or fail to settle disagreement (by getting caught up in unsettled debates about rationality). We speculate that these outcomes reflect a fundamental sort of normative disagreement, which can be reshuffled to the domains of morality or rationality, but cannot be avoided.

1. Cognitive science and the quest for an Archimedean cleaver

Moral disagreement is commonplace. In political communities, even within families, ordinary people disagree about the right thing to do. And in philosophy departments, trained ethicists disagree over which moral theory to endorse. Is the right action always whichever leads to the best consequences? Or are there moral restrictions on which sorts of actions may be employed to get those good consequences? Some moral disagreements have persisted for centuries. One explanation for this may be that moral arguments appeal only to those who already accept controversial background views; whether you'll accept a certain weighing of consequences against other considerations probably depends upon your overall fondness for consequentialism. So, if there's to be much hope of resolving these very old disagreements, it will probably have to come from somewhere outside of moral theory – somewhere neutral, somewhere unencumbered by moral commitments. Could cognitive science do this?

This isn't an obviously silly idea. Moral argumentation is a form of moral *reasoning*.[i] There is already a very well-established body of work in cognitive science apparently showing that people tend to make predictable mistakes in certain reasoning domains (we'll discuss this literature in the next section). Perhaps these findings could be adapted to the moral domain. If we could show that some patterns of moral reasoning resemble defective *non-moral* reasoning, then we might have identified bad moral reasoning. Whatever's left would then be (barring further information) good moral reasoning, identified by science.

What we're considering here is allowing cognitive science to play an *Archimedean* role in ethics. The Greek polymath Archimedes is purported to have boasted that he could move the earth itself, if he had only a place to stand outside it. In theoretical inquiry, an Archimedean point is somewhere outside the target domain, a neutral perspective allowing resolution of disagreement. In this chapter we will speak of an *Archimedean cleaver*: some criterion, expressed in neutral cognitive scientific terms, that allows us to divide the good moral reasoning from the bad.[ii]

We will look at various ways in which the cognitive science of reasoning might be thought to provide ethics with an Archimedean cleaver. It will be important to keep in mind the two essential features embedded in the term. (1) To be *Archimedean*, the criterion really must be

morally neutral; it cannot assume or rely upon any moral beliefs for its validity or application. (2) To be a *cleaver*, the criterion must demonstrate clearly that some forms of moral reasoning are good, and some are bad. It cannot waffle.

To put our cards on the table: we doubt that cognitive science can offer any Archimedean cleaver to ethics. But we think it will be illuminating to consider some proposals and to show why they are ultimately unsuccessful. We think that in the end all extant proposals either fail to be Archimedean, because they rely on hidden moral assumptions, or fail to be cleavers because they don't actually resolve moral debate. In fact, we suspect that there is a tension between the Archimedean and cleaving components, such that no cognitive scientific criterion could meet both at once.

The structure of this chapter works as follows. In section 2 we survey the cognitive science of *general* reasoning. In section 3 we argue that some proposed Archimedean cleavers fail to remain morally neutral, and so are not really Archimedean. In section 4 we argue that other proposed Archimedean cleavers rely upon unresolved debates in cognitive science, and so are not decisive cleavers. In section 5 we consider whether Archimedean aspirations may be fundamentally in tension with the way cognitive science can divide reasoning.

2. The cognitive science of reasoning

Our provisional, working definition of reasoning is the following: reasoning is the capacity to reach a conclusion through inferences from previously accepted propositions.. Reasoning is often prompted by the practical need of solving a problem. In practical reasoning, we manipulate information in linguistic form to decide what course of action is best. Moral reasoning is practical reasoning concerning morally relevant issues. In this section, we will briefly deal with the domain-general cognitive science of reasoning.

Cognitive science and other social sciences have investigated reasoning in multifarious ways. This investigation has not been only descriptive; most analyses of reasoning contain a normative component regarding the distinction between good and bad forms of reasoning. Good forms of reasoning use valid inferences and hence lead to true conclusions (provided that the premises of the inference are true). Bad forms of reasoning are prone to various errors. Which forms of reasoning yield reliable results and which are fallacious? This normative component is embodied in the concept of rationality.

Work in cognitive and social sciences has generated a range of conceptions of rationality. These differ along two distinct dimensions. First, what are the normative standards that we should adopt to evaluate human choices? Secondly, how well does actual human behavior live up to these standards?

The first conception that we will examine is the microeconomics view. According to rational choice theory (Von Neumann & Morgenstern, 1972), a consumer aims to maximize her utility

given a budget and a range of goods and services that she can buy. To do so, the consumer must have a set of preferences that allows her to put options into an ordinal ranking.

If it is also possible to assign a real number to the value of each preference, then there exists a consumer's utility function that can be maximized. According to rational choice theory, consumers should choose options in such a way as to maximize their utility. Hence, rational choice theory is both a normative standard and a descriptive claim. It purports that people actually maximize their utility while acting in the market and that they should do so to be rational.

Although this view is still the dominant view on human choice, many criticisms have been raised. We will examine alternative views, from the less heterodox to views that greatly distance themselves from rational choice theory.

The heuristic and biases approach by Kahneman and Tversky allegedly showed that actual people deviate from the predictions of rational choice theory in many ways. For instance, humans tend to desire the avoidance of a loss of magnitude M more than gaining something new that is worth M. Humans value avoiding losses more than equivalent gains, other things equal. As a result, framing two identical choice scenarios in terms of losses or gains heavily influences the participants' decisions.

The typical case of a framing effect involving losses and gains is the so-called Asian Disease scenario (Tversky & Kahneman, 1981, p. 453). The experiment has two conditions. Participants are presented with a scenario involving a deadly epidemic and options for intervention. Participants in Group 1 must choose either a risky or a riskless option, and the outcomes of these options are phrased as "lives saved." Participants in Group 2 must pick between options whose outcomes are logically equivalent to those seen by Group 1, but that are phrased in terms of "deaths" instead. Experimental results showed that Group 1 preferred the riskless option to the risky option, whereas Group 2 preferred the risky option to the riskless option.

The most straightforward explanation of these results concerns a difference in how the baseline is identified. The framing leads Group 1 to think that the potential victims of the disease are doomed, so much so that they must be "saved." In this case, each lost life is a no-gain, because people are already seen as almost dead. By contrast, Group 2 is led to believe that the potential victims are in perfect health. Under this framing, each death is a loss. The results of this experiment can be taken as evidence that people are more prone to take risks in order to avoid losses than to secure gains.

Kahneman and Tversky created a new normative theory of choice, prospect theory (Kahneman & Tversky, 1979), which still relies on utility maximization, in order to accommodate the differential treatment of losses and gains. The picture of human reasoning that emerges from this research is that we should maximize prospects, but in many cases we do not do so. Human irrationality is rampant.

Some theorists, however, have parted ways with maximization as a normative standard and have opted for satisficing instead. According to satisficing, a choice or behavior is rational when it solves a given problem, not necessarily when it solves the problem in the best possible way. Given that choice is cognitively and metabolically effortful, that information-gathering is both time consuming and costly, and that time constraints are virtually ubiquitous in real-world choices, maximization is often implausible or unfeasible. Gerd Gigerenzer (2005) has developed theories of human reasoning that do without maximization. According to Gigerenzer, the human mind reasons through fast-and-frugal heuristics that serve us well most of the time. Errors due to heuristics are inevitable side-effects of the correct working of the human heuristics toolbox. The only normative criterion for a cognitive strategy that Gigerenzer seems to acknowledge is its success in the real world, where success is to be interpreted mainly in evolutionary terms: reasoning is successful if it allows a given human being to survive, thrive, and eventually reproduce in a given environment.

Gigerenzer is deeply unsatisfied with the picture of mankind depicted by Kahneman and Tversky. In his view, humans are normally rational, and Kahneman and Tversky's judgment of irrationality relies, among other things, on the application of the wrong norms and on neglect of environmental features. For instance, Gigerenzer (2015) claims that the Asian disease case is due to the fact that the riskless options (A and C) are incomplete. The scenario does not specify that 400 people will not be saved in Program A and that 200 people will not die in Program C. In Gigerenzer's view, the incomplete phrasing of these options would lead participants to interpret the information as an implicit recommendation. Furthermore, if experimental participants are given complete descriptions, the framing effect disappears (Mandel, 2001).

Cognitive science has not yet yielded a consistent picture of human rationality. Both how humans actually reason and how humans ought to reason are still hotly contested topics (Kahneman & Tversky, 1996; Gigerenzer, 1996; Stanovich & West, 2000; Krueger & Funder, 2004).

3. Non-Archimedean Cleavers

If the cognitive science of reasoning were able to provide an Archimedean cleaver, what would it be like? It would need to provide some criteria, derived from cognitive scientific inquiry, by which we can label some types of moral reasoning 'good' and others 'bad' – *without* relying upon any moral assumptions. In this section we will look at proposals that fail to be Archimedean because of the latter clause. That is, they might be able to determinately sort good from bad moral reasoning, but they only accomplish this by relying upon moral assumptions. They are cleavers, but not Archimedean.

Here is an illustration. Suppose that you think consequentialism is the one true moral theory. That is, you think that the consequences of actions are what matters morally, and nothing else matters morally. Then you have an easy cleaver: forms of moral reasoning that depend upon counting something other than consequences as mattering will be 'bad' forms of moral reasoning, and forms of reasoning that only count consequences will (or at least might) be 'good'

forms of moral reasoning. This is not Archimedean, since it relies upon assuming the truth of consequentialism, a particular moral theory.

Now suppose that you use your consequentialist cleaver to sort through data from the cognitive science of moral reasoning. You look at studies of how people reason about punishment, or distribution of resources, or the right way to resolve dilemmas. The results show that certain patterns of moral reasoning are insensitive to consequences. People assign greater punishment to 'outrageous' violations even when lesser punishment might lead to better behavior (Sunstein 2005). People are also more likely to donate aid to individually identifiable victims of disaster than to larger numbers of anonymized victims (Small and Loewenstein 2003). Perhaps these people don't even realize that they aren't attending to consequences – but tracking their pattern of judgment will allow us to see it. So cognitive science can detect *the fact that people aren't making consistently consequentialist judgments*.

But, of course, cognitive science can't tell us that it is *bad* to attend to something other than consequences. That is a moral assumption. And that is the assumption made by social theorist Cass Sunstein in his paper "Moral Heuristics" (2005). Building on the Kahneman and Tversky research program discussed above, Sunstein argues that many moral judgments appear to resemble defective economic reasoning in failing to maximize good outcomes. Hence, Sunstein says, these forms of reasoning are "moral heuristics" – they track good consequences most of the time, but predictably fail in certain cases, and therefore are not reliably good reasoning.

Sunstein is explicit in relying upon what he calls "weak consequentialism" as a moral standard. He defines weak consequentialism like this: "the social consequences of the legal system are relevant, other things being equal, to what law ought to be doing" (2005, 534). (Though the title of his paper uses the word "moral", Sunstein prefers legal examples.) Sunstein thinks this is a very ecumenical criterion for detecting good moral reasoning; he doubts anyone but extreme deontologists might disagree.

There are problems with weak consequentialism. One is the idea that consequences are "relevant, other things being equal". The problem is that almost everyone accepts that consequences *are* relevant; what is distinctive about an anti-consequentialist perspective is that it says other things, in addition to consequences, are *also* relevant (Scheffler 1982). In fact, even Sunstein's purported heuristics needn't be interpreted as opposed to weak consequentialism. For instance, Sunstein considers the morality of emissions trading: should companies that pollute be able to pay a fee in order to avoid legal punishment? Philosopher Michael Sandel (1997) thinks not. He believes that emissions trading turns immoral excess pollution into just another cost of doing business rather than an occasion for serious moral censure. Here is what Sunstein says about this:

> "At least some level of pollution is a byproduct of desirable social activities and products, including automobiles and power plants. … When Sandel objects to emissions trading, he is treating pollution as equivalent to a crime in a way that overgeneralizes a moral intuition that makes sense in other contexts. There is no moral problem with

emissions trading as such. The insistent objection to emissions trading systems stems from a moral heuristic." (Sunstein 2005, 537)

However, notice that Sandel's view is not opposed to weak consequentialism. Sandel can agree that consequences "are relevant, all else being equal" – he can accept that consequences matter. But he doesn't think that all else *is* equal when it comes to pollution. So weak consequentialism isn't a sufficient basis to declare Sandel's moral reasoning defective, since Sandel can *agree with* what weak consequentialism says.

What Sunstein needs is a much stronger form of consequentialism. He needs to say that consequences are *all* that matters to moral reasoning. If we hold *that* view, then of course it turns out that Sandel is reasoning poorly. Sandel thinks that excess pollution is wrong even when permitting it for a fee might lead to positive social consequences. But this is bad reasoning only if we assume that consequences are *all* that matters.

One might, of course, disagree with Sandel for reasons other than accepting consequentialism. But the point here isn't about whether philosophers can provide counter-arguments to Sandel's view. The point is about whether Sunstein is in a position to discard Sandel's reasoning as *defective* on scientific grounds. And he can only do that if he assumes the truth of consequentialism. But then he could hardly claim to be Archimedean. It isn't possible to settle moral disagreement by assuming the truth of one side of the disagreement.

We think that this pattern repeats across the literature on moral reasoning. Theorists claim to detect bad moral reasoning on the basis of cognitive scientific evidence, but this detection only works if we assume a philosophically controversial account of what matters morally. Sunstein is unusual in being transparent about it; many other theorists do not openly acknowledge that they are assuming something morally controversial. Rather, they claim to build an argument from neutral, scientific bases, and *derive* from this a conclusion that cleaves good from bad moral reasoning. We will consider one more example.

The developmental psychologist Lawrence Kohlberg claimed to accomplish the greatest of philosophical tricks: he could derive an 'ought' from an 'is'. More precisely: he could use the tools of psychology to show, objectively, what the best form of moral reasoning must look like. As he put it: "an ultimately adequate *psychological* theory as to why a child does move from stage to stage, and ultimately adequate *philosophical* explanation as to why a higher stage is more adequate than a lower stage are one and the same theory extended in different directions" (1971, 104; emphases in original). Kohlberg intended to solve moral philosophical problems.

Kohlberg studied children's moral reasoning for decades and collected an enormous amount of data on developmental patterns. Young children tend to justify moral decisions by mentioning punishment or reward, while older children tend to provide increasingly abstract and universal justifications. Most adults, Kohlberg claimed, only reach a level of moral development concerned with upholding conventional law and order, but a few progress beyond this to thinking in terms of a social contract or universal human rights.

One reason Kohlberg saw the latter stages as morally superior was his assumption that "the cognitively and ethically higher or more adequate must come later than the less adequate" (132). His assumption was that *better* moral reasoning must build upon and then replace *worse* moral reasoning. This suggests a deceptively easy cleaver: good moral reasoning is what older people do; bad moral reasoning is what younger people do. But of course Kohlberg acknowledged that some older people could be bad reasoners, since some never developed to the highest level and others might regress. Though age correlates with higher moral reasoning, Kohlberg needed some age-independent criterion to pick out the best sort.

What he provided is the claim that higher moral reasoning does a better job of "handling" moral dilemmas because it makes use of sophisticated logical operations that are unavailable at lower levels. These logical operations are differentiation (drawing relevant distinctions between apparently similar cases) and integration (synthesizing relevant commonalities among apparently dissimilar cases). Kohlberg claimed that the most complete operation of differentiation and integration came only in the form of absolutely universal justice: moral reasoning that tries to construct principles governing the actions of all people, regardless of their social position or personal preferences. Particular moral judgments could only be justified as derived from this universal perspective. According to Kohlberg, John Rawls' (1971) contractarianism is the philosophical manifestation of the highest sort of reasoning.

Kohlberg clearly intended these observations to function as a moral cleaver. He argued that *both* Kantian normative ethics and standard utilitarian ethics reflect a limited sort of moral reasoning, inferior to Rawlsian contractarianism. If his views were accepted, then we would have a cognitive scientific basis for trusting the results of Rawlsian reasoning more than those of Kantian or utilitarian reasoning. And if Kohlberg were right that he reached these conclusions without making any moral assumptions, then this would be an Archimedean cleaver.

Unfortunately, Kohlberg *did* make moral assumptions. This comes out most clearly in the long-running debate between Kohlbergians and psychologist Carol Gilligan. Gilligan argued that Kohlberg failed to take account of gender differences; on her interpretation, many girls and young women approached moral differences from a *relational*, rather than universal, perspective (Gilligan 1982). For instance, girls first consider whether a given choice would maintain a particular family connection and only then look at it from a universal perspective. Kohlberg was forced to interpret this as a sign of moral immaturity in girls, but according to Gilligan it is a sign merely of difference.

There has been much dispute about Gilligan and Kohlberg's empirical claims, and we can leave to the side whether either has shown anything about gender differences. For our purposes, the important point is this: in order to draw his cognitive scientific conclusion about the superiority of late-stage moral reasoning, Kohlberg had to assume that universalistic reasoning is the *better* type. On Kohlberg's view, the moral value of particular concrete relationships is only a derivative value, one that can be justified from abstract principles. It is not of intrinsic moral value itself. Yet Gilligan articulated a rival view, according to which relationships are both intrinsically valuable *and* the root justification for moral reasoning.

7

Theorists can, of course, argue over whether moral reasoning ought ultimately be grounded in universal principles or maintenance of relationships (or something else). Our point here is only that cognitive science will not settle the argument. In fact, we will only be able to employ cognitive science as Kohlberg employed it if we *assume* certain goals for the highest sort of reasoning. Kohlberg assumed that the highest sort of moral reasoning would aim at abstract universal principles, but without this assumption his cognitive criteria of differentiation and integration lose their normative force. Making concrete situations more abstract is only valuable if abstraction is the ultimate goal – and that is exactly what is in dispute.

So Kohlberg's cleaver was not Archimedean. And this, we think, is the fate of many purported attempts to wield cognitive science to separate moral intuitions. Like Sunstein, Kohlberg could determinately answer that, on his view, certain sorts of moral reasoning are better than other sorts. But neither Sunstein nor Kohlberg could provide a morally neutral way of doing this, and so neither is in an Archimedean position to settle moral disagreement.

4. Archimedean Non-Cleavers

By contrast with the attempts described in the previous section, there are also ways to use cognitive science that do not rely on implicit or explicit assumptions about what matters morally. In particular, there exist attempts at using cognitive science to distinguish between good and bad forms of moral reasoning that are not based upon assumptions on what is or is not morally relevant. These attempts are Archimedean, at least as far as the moral domain is concerned. However, in this section we argue that they pay a price for their moral Archimedeanism. The price is that they fail to be cleavers.

We will examine two examples: an argument by Horowitz (1998) and another by Sinnott-Armstrong (2008).

Horowitz attacked an attempt by Quinn (1993) to use moral intuitions[iii] to buttress the Doctrine of Doing and Allowing (DDA). The DDA claims that it is more morally blameworthy to bring about the same harmful consequences through an action than through an omission. The moral intuitions Quinn is interested in are responses to these cases:

> "Rescue Dilemma 1: We can either save five people in danger of drowning in one place or a single person in danger of drowning somewhere else. We cannot save all five.
> Rescue Dilemma 2: We can save the five only by driving over and thereby killing someone who (for an unspecified reason) is trapped on the road. If we don't undertake the rescue, the trapped person can be later freed." (Horowitz 1998, p. 368)

Here the intuitions are that it is morally permissible to save the five in Dilemma 1 but not in Dilemma 2, in line with the DDA.

Quinn argued that these intuitions are caused by the DDA, and that the existence of these intuitions buttresses the DDA.

By contrast, Horowitz argued that these intuitions are uncoupled from the DDA. In her view, it is Kahneman and Tversky's prospect theory that actually explains our intuitions here.

In Dilemma 1, the baseline is that all people are close to death, so that all interventions are rescues, i.e. gains. So, the two options for Dilemma 1 are:
a) we save the five, which results in five gains and one no-gain,
b) we save the one, which results in one gain and five no-gains.
Since option a) yields more value than option b), it stands to reason to save the five.

Nonetheless, in Dilemma 2, the baseline is that the person trapped on the road is healthy, so that her death would be a loss. Hence, we face the following two options:
c) we save the five, which causes five gains and one loss;
d) we help the one, so that we get five no-gains, and the person trapped in the road can be freed.

In accordance with prospect theory, we attribute a big negative value to the loss in c). As a result, d) yields more overall positive value than c) and ought to be preferred to it. This means that, in Horowitz's opinion, the intuitions to which Quinn is appealing are not causally linked to the DDA. In this case, human choice is actually driven by prospect theory.

Horowitz does not argue against the DDA, but only against Quinn's argument in favor of it. Her argument is thus not directly an argument in normative ethics (in the sense that it does not argue for a normative ethical claim such as 'The DDA is false'), but an argument about the correct empirical explanation for the existence of some moral intuitions. Horowitz does not think that prospect theory is a good (or, for that matter, bad) guide for human moral reasoning. She is only interested in whether the DDA empirically and causally explains the existence of the aforementioned intuitions, and argues that prospect theory provides a better explanation than the DDA.

Horowitz's argument aims to distinguish good ways of reasoning and bad ways of reasoning. In her view, philosophical explanations of moral intuitions that do not take into account results in cognitive science are bad ways of reasoning in moral matters. By contrast, explanations of moral intuitions grounded in results in cognitive science are good forms of moral reasoning.

There are many perceptive replies to Horowitz's argument (Kamm, 1998; Van Roojen, 1999). We would like to add a point that has not been made so far: Horowitz's argument assumes the descriptive adequacy of prospect theory, and neglects the intense debate in cognitive science about human rationality. Since Horowitz provided no argument in favor of prospect theory as a descriptive theory of human reasoning, she will persuade only those who already uphold it.

It is important to notice that Horowitz's argument, unlike those by Sunstein and Kohlberg (see § 3), does not make any explicit moral assumption. Her argument against Quinn does assume a theory about human rationality, but makes no explicit assumption concerning morally relevant

factors. Hence, it may be Archimedean at the moral level, but is controversial at the level of rationality debates. As a result, the question about the validity of Quinn's reasoning is not solved but transferred from moral theory to the theory of rationality. Unless the quandary about rationality is solved, the argument will not be compelling. Those who deem prospect theory to be correct will accept Horowitz's explanation of rescue dilemmas, whereas those who are unconvinced by prospect theory will see her argument as moot. Therefore, the argument fails to be a cleaver.

Another argument of this type has been made by Sinnott-Armstrong (2008). In his opinion, results by Tversky, Kahneman, and other cognitive scientists showed that moral intuitions cannot be justified without inferential justification. His goal is to show that non-inferential justification of moral intuitions is impossible. As moral theory only admits of inferential and non-inferential justifications, the impossibility of non-inferential justification makes inferential justification the only available form of moral justification[iv].

The argument works as follows:

> "(1) If our moral intuitions are formed in circumstances where they are unreliable, and we ought to know this, then our intuitions are not justified without inferential confirmation.
> (2) If moral intuitions are subject to framing effects, then they are not reliable in those circumstances.
> (3) Moral intuitions are subject to framing effects in many circumstances.
> (4) We ought to know (3).
> (5) Therefore, our moral intuitions in those circumstances are not justified without inferential confirmation." (Sinnott-Armstrong 2008, p. 52)

The key premises are (2) and (3). Sinnott-Armstrong argues that moral intuitions are prone to many biases, including framing and order effects[v]. According to Sinnott-Armstrong, these effects are so pervasive in human moral cognition to warrant premises (3) and to make the mechanisms of moral belief formation unreliable in most circumstances.

The argument hinges on the extent of distorting influences on human moral reasoning. Are biases so widespread that humans are in need of inferential confirmation for their moral beliefs? This is an empirical question, which centers on cognitive science. However, as noted in § 2, there is a lively debate in cognitive science as to the nature and extent of biases.

Among other biases, Sinnott-Armstrong highlights framing effects like the Asian Disease case examined in § 2. But Gigerenzer and his co-workers claim that this framing effect is an artifact due to the incomplete presentation of the riskless options. It is thus unclear whether Sinnott-Armstrong can simply take for granted that framing effects are actual instances of widespread human irrationality. If they were artifacts due to the incompleteness of the options, they would not threaten the credibility of human moral cognition.

Hence, our take on this argument is that it relies on Kahneman and Tversky's conception of rationality. A philosopher can subscribe to whatever conception of rationality she wishes, but the reader must be given an argument as to why the conception of rationality endorsed by the philosopher is to be preferred to its competitors. It is likely that theorists sharing Gigerenzer's view would not consider framing effects like the Asian Disease case a problem for human moral intuitions.

It is unclear whether Gigerenzer's theory can explain away *all* of the biases that were identified by Tversky, Kahneman, and other investigators. Nonetheless, given that many biases are controversial in cognitive science, Sinnott-Armstrong would need to separate controversial from uncontroversial biases. It remains an open question whether his premises (2) and (3) can be supported by uncontroversial biases on which the vast majority of experimental psychologists agree.

Sinnott-Armstrong's argument aims to differentiate bad forms of moral reasoning (those that mistakenly assume non-inferential justification of moral intuitions) and good forms of moral reasoning (those recognizing the need for inferential justification).

Like Horowitz's, Sinnott-Armstrong's argument makes neither covert nor explicit assumptions about what morally matters, but does assume a controversial conception of rationality that many reasonable people do not share. In particular, supporters of Gigerenzer's views will not accept his argument, because they will reject premises (2) and (3). These moral theorists will (rationally) be able to think that non-inferential justification of moral intuitions is indeed an open option. As a result, Sinnott-Armstrong's argument fails to solve the controversy that it wants to address. Hence, it cannot be considered a cleaver. In our view, all that this argument achieves is that it transfers the quarrel about non-inferential justification from the domain of moral theory to the controversies surrounding rationality in cognitive science. Unfortunately, little is solved, and we still do not know which forms of moral reasoning are good and which are bad.

To sum up, Archimedean arguments that do not make assumptions in normative ethics are possible, but they fail to put an end to the problems that they aim to solve.

5. Is Archimedean moral psychology possible?

We have discussed two distinct ways a purported Archimedean cleaver might fail: it might not be Archimedean at all (because it relies on some moral standard) or it might not decisively cleave (because it gets caught up in interminable debates about the psychology of reasoning). We have argued by example, showing that several seemingly promising research programs fall into one or the other of these problems. We have *not* shown that it is *impossible* to design some research program that will provide a genuine Archimedean cleaver. Perhaps someone will do so. But we are unconvinced that anyone has. And we suspect that it will not be done, for reasons explained in this section.

We think there is a relationship *between* the two mandatory elements of an Archimedean cleaver, such that achieving both at once is extremely difficulty, perhaps impossible. We can start to see this by returning to these elements, now refined by the discussions of the last two sections:

(1') To be *Archimedean*, a cognitive scientific criterion must not rely on any moral assumptions in either its scientific operationalization or in its analysis of data.

(2') To be a *cleaver*, a cognitive scientific criterion must decisively resolve moral disagreement by not relying on standards of domain-general reasoning that can be disputed among those disagreeing.

These are seemingly two distinct problems, in that one deals with *moral* assumptions and the other deals with domain-general *reasoning* assumptions. But it may be more helpful to see them instead as two ends of the same *normative* problem. We will now speculate (there is not space here to fully argue) that disagreement in the moral domain mirrors disagreement in other reasoning domains, in a way that complicates grounding Archimedean moral critique in other domains.

Disagreement in morality is not isolated from disagreement in domain-general reasoning. Consider disagreement between a consequentialist and an anti-consequentialist. They disagree about whether a moral agent ought to look only to the consequences of her actions to guide her decisions. What happens when these same people consider a different case, one that is not moral? Suppose they are now disagreeing about how an agent ought to spend her annual salary bonus. Suppose the agent has already donated a reasonable amount to charity this year and has decided to spend the bonus on herself. Her choice is therefore now a *prudential* one, rather than a moral one. Will our consequentialist and anti-consequentialist agree about the standards for assessing the rationality of her prudential choice?[vi]

Probably not. Let's imagine that this agent decides to buy a present for herself every day for one week. The shops have sales on various days, and there is clearly an optimal sequence to buy the presents in order to get the most for her money. She knows this, but she does not follow the optimal sequence. Instead, she decides to ignore the sales, and instead buy her least favorite present first, then the next-least, and onward until she buys her favorite preset on the last day. She does this knowing that it is financially sub-optimal and will result in her getting fewer presents than otherwise. But she likes the idea of having a rising 'shape' to her week of fun: it seems good to her that the week get better and better as she goes, even if this means the total amount of value she receives is less than it might be.[vii]

Is this character irrational? There's a good chance that our disagreeing theorists will disagree here as well. We suspect that the consequentialist is more likely than the anti-consequentialist to think that the character is behaving irrationally. This isn't a claim about logical commitment; we won't try to argue that the consequentialist *must* say anything in particular here. Rather, this is a speculative conjecture about the normative disposition of different types of people. We speculate that a theorist who is a consequentialist in the moral domain is more likely to endorse outcome-maximizing forms of reasoning in non-moral domains as well.

More generally, we can mark a distinction between broad ways of evaluating practical reasoning: a teleological way, which values reasoning likely to produce the most of some specified value (morally good states of affairs, personal utility); and a formal way, that values internal features of the reasoning (deontological constraints, concern for the 'shape' of a life). With this distinction, we can state our hypothesis like this: theorists tend to be consistent in favoring one way of reasoning – teleological vs. formal – across various practical reasoning domains.

We cannot establish this hypothesis here. But if it turns out to be right, it will support our suspicion that the two ways for a purported Archimedean cleaver to fail to be such are just different expressions of the same underlying disagreement about the nature of practical reasoning. This disagreement can appear in the moral domain, over whether consequentialist reasoning is the best sort of reasoning, and it can appear in the prudential domain, over whether e.g. economic decisions should be utility-maximizing. Wherever the disagreement appears, it will drive opposing sides' implementation and interpretation of psychological studies.

If our speculation is correct, then disagreements in morality may simply be one end of broader disagreements about reasoning. Since the cognitive science of reasoning relies upon normative assumptions, it cannot provide a neutral fulcrum on which to weigh moral arguments. Those seeking Archimedean resolution are not likely to find it in the psychology lab.

CITATIONS

Gigerenzer, G. (1991): How to Make Cognitive Illusions Disappear: Beyond "Heuristics and Biases." *European Review of Social Psychology, 2*(1), 83-115.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review, 103*(3), 592–596.

Gigerenzer, G. (2005). I think, therefore I err. *Social Research, 72*(1), 195-218

Gigerenzer, G. (2015). On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology, 6*, 361–383.

Gilligan, C. (1982). *In a Different Voice: Psychology Theory and Women's Development*. Cambridge, MA: Harvard University Press.

Glasgow, J. (2013). The shape of a life and the value of loss and gain. *Philosophical Studies, 162*(3), 665-682.

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review, 108*(4), 814-834.

Horowitz, T. (1998). Philosophical Intuitions and Psychological Theory. *Ethics, 108*(2), 367–385.

Kahneman, D., Fredrickson, D. L., Schreiber, C. A., Redelmeier, D. A. (1993) When more pain is preferred to less: Adding a better end. *Psychological Science*, *4*(6), 401–405.

Kahneman, D., & Tversky, A.. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, *103*(3), 582-591.

Kamm, F. M. (1998). Moral Intuitions, Cognitive Psychology, and the Harming-Versus-Not-Aiding Distinction. *Ethics*, *108*(3), 463-488.

Kohlberg, L. (1971). From 'Is' to 'Ought': How to Commit the Naturalistic Fallacy and Get Away with It in the Study of Moral Development. Reprinted in Kohlberg, L. *Cognitive Development and Epistemology*, ed. Theodore Mischel. New York: Academic Press.

Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences, 27*, 313-376.

Mandel, D.R. (2001). Gain-loss framing and choice: separating outcome formulations from descriptor formulations. *Organizational Behavior and Human Decision Processes, 85*(1), 56–76

Quinn, W. (1993). Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing. In Quinn, W., *Morality and Action*. Cambridge (UK): Cambridge University Press.

Rawls, J. (1971). *A Theory of Justice*. Cambridge MA: Harvard University Press.

Sandel, M. (1997, December 15). It's immoral to buy the right to pollute. *The New York Times*. p. A23.

Scheffler, S. (1982) *The Rejection of Consequentialism*. New York: Oxford University Press.

Singer, P. (1981). *The Expanding Circle: Ethics and Sociobiology*. Oxford: Oxford University Press.

Sinnott-Armstrong, W. (2008). Framing Moral Intuitions. In Sinnott-Armstrong, W. (Ed.), *Moral Psychology, Vol. 2*. Cambridge (MA): The MIT Press.

Small, D. A. & Loewenstein, G. (2003) Helping *a* Victim or Helping *the* Victim: Altruism and Identifiability. *The Journal of Risk and Uncertainty, 26*(1): 5-16.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences 23*, 645-726.

Sunstein, C. (2005). Moral Heuristics. *Behavioral and Brain Sciences, 28*, 531-573.

Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and The Psychology of Choice. *Science, 211*, 453-458

Van Roojen, M. (1999). Reflective Moral Equilibrium and Psychological Theory. *Ethics, 109*(4), 846-857.

Von Neumann, J., and Morgenstern, O. (1972). *Theory of Games and Economic Behavior*, Princeton, N.J.: Princeton University Press.

Williams, B. (1985) *Ethics and the Limits of Philosophy*. Cambridge MA: Harvard University Press.

---

[i] Some psychologists – most influentially Jonathan Haidt (2001) - doubt that moral judgment has much to do with *reasoning*. This is a controversial view, which we will leave to the side in this chapter.

[ii] A number of philosophers have expressed the hope, in the words of Peter Singer, that "precisely because science is outside ethics, the scientific study of our ethical judgment is a fulcrum on which we can rest our critical lever" (Singer 1981, 73). The idea of an 'Archimedean' approach to ethics (and criticism thereof) is even older. See Bernard Williams (1985) for an influential restatement.

[iii] Our working definition of a moral intuition is by Sinnott-Armstrong (2008): moral intuitions are strong and immediate moral beliefs. Strong beliefs are those that the believer does not easily give up. Immediate beliefs are those that do not stem from inference from other beliefs, be they moral or non-moral in nature.

[iv] This in turn would leave the justification of moral beliefs vulnerable to a regress argument (Sinnott-Armstrong, 2008, p. 49). We will not deal with this interesting argument in this article.

[v] Order effects arise when, all other things being equal, the order in which stimuli are presented to a participant influences her response to them.

[vi] Some philosophers may reject the moral/prudential distinction altogether. We set that view to the side, as it would only strengthen our claim that disagreements in moral and prudential reasoning are linked.

[vii] This example mimics Kahneman et al.'s (1993) finding that people will accept greater total pain so long as it ends on an upswing. For philosophical discussion of temporal 'shape', see Glasgow (2013).