Is morality constitutive of rational human agency?

'Take an action allow'd to be vicious...Wilful murder, for instance. Examine it in all lights, and see if you can find that matter of fact, or real existence, which you call vice. In whichever way you take it, you find only certain passions, motives, volitions and thoughts. There is no other matter of fact in the case. The vice entirely escapes you, as long as you consider the object...You never can find it, till you turn your reflexion into your own breast, and find a sentiment of disapprobation, which arises in you, towards this action. Here is a matter of fact; but 'tis the object of feeling, not of reason. It lies in yourself, not in the object.'

David Hume (*A Treatise of Human Nature*, 3.1.1.26, SBN pp. 468—469)

In this essay I have examined the case for constituivist constructivism (CC), a theory founded on the belief that morality is constituted by what we are as agents, not by what can be ascertained by probing the external world. Morality's origins sit within the reasoned authority of our rationality, which guides us; moral truths are not examinable by laying problems down metaphysically *in front* of us. I have paid close attention to recent and notable publications on the subject written by Christine Korsgaard. By parsing the details of Korsgaard's project I have argued that her Kantian-born CC is flawed: (i) because it falsely assumes that agents are inescapably bound to morality through rationality; and (ii) because her theory optimistically relies on the idea of an agent whose practical reasoning powers are packed into a single unified entity, for such an agent does not exist.

1. Constituivist constructivism

Constituivist constructivism (CC) abandons the notion that the origins of morality can be found rooted in the external world, as detectable tokens of rightness and wrongness. It, instead, appeals to an internal, mind-dependent version of morality. When approaching moral dilemmas to determine what we think we ought to permit we should begin with first-order questions about *us* (hence 'constituivist') since morality's true source is hidden within us, separable from the world we experience with our senses.

The tactic of CC is as follows: demarcate moral agents; such agents can construct justifiable moral claims rationally. It is a theory which postulates that moral claims can be objective if they are rooted in what we are. It *ipso facto*, however, avoids committing itself to making metaphysically objective claims. This is a dialectically opposite position from 'realist' metaethics views, which express ethical sentences as propositions about objective features of the world, though we may respond to them differently. For example, take 'robust non-natural moral realism', through which at least some claims, morally speaking, just are irreducible 'brute facts'—non-descriptive ethical features which are independent from our cognitive states and cross possible worlds. Naturalistic moral realism holds that moral claims supervene on physical properties because they are reducible to or constituted by scientific facts, be they *a priori* or *a posteriori*.

In stark contrast to moral realism, CC attempts to provide a foundation for defining morality in terms of *what we are*, where we can all obtain substantive, *world-guiding* moral facts with certain agential properties. The central challenge to CC, as explored in this essay, is: how? Exactly what is our cognitive formula? In the following section I have described the CC view which claims morality is constituted by *rationality*—a position I have gone on to contest.

2. A contemporary Kantian account

The question of how morality is constituted in agents divides CC theorists. I have discussed one side of the current schism, Kantianism, which can be contrasted to the other, Humeanism. Here I focus on the former through the views of Christine Korsgaard (1996; 2009).

Kant's deontological theory (Kant 1998) lends itself to the metaethical discussion of CC because his axioms entail a version of morality which constitutively begins in us through the reason of our rationality.

Veering away from morality which is conditional on events that transpire, Kant argued that humans should be held as the necessary ends of our moral assessments; and, to reach these ends, we should employ our common tools of rationality. It is this part of *our* shared nature which distinguishes us from non-human animals as we can provide reasons for our beliefs; and if we direct this practical reasoning power towards humans as ends, the result will be good for the human race.

Under the 'categorical imperative' (CI) such a moral, law-abiding agent constructs maxims through rational assessments. In self-regulatory fashion the agent knows how they ought to treat people under these maxims—laws of moral conduct which they will to become universal laws. The CI, then, holds unconditional authority over rational agents in terms of their normative judgements because *they ought to* conform to it. It, therefore, provides a transcendental account of morality because our rationality is thought to govern morality noumenally.

Christine Korsgaard outlines a form of CC with Kantianism at its foundation, named '*procedural* realism'. According to this view, we are capable of behaving as moral-truth-makers through rationality. Like any belief, moral judgements can be true of false, where true beliefs are composed by the procedures of rational justification. Thus if moral knowledge is metaphysically constituted in what we are, *what we are* is rational agents.

The principle of universality commits us to morality when followed. If I operate outside of it, I act outside my moral agency. This renders the notion of agency essential to Korsgaard's view:

'*[R]eason's own principle just is the principle of acting in a way that constitutes you as a single unified agent…When you deliberate in accordance with these principles, you pull yourself together and place yourself, so to speak, behind your movement, rendering it an action that can be ascribed to you as a whole.*' (Korsgaard 2009, p. 179)

This is the 'volitional unity' of a self-reflective agent whose rational, deliberative reasoning powers alleviate from them the ingredients of their usual, amoral agency—motives, beliefs, desires, spirit, hunger—whatever they are—and transform the person into an autonomous moral agent who is capable of rationally devising new moral recipes.

In expounding her view, Korsgaard pre-empts questions of why such an 'inwardly just' way of living—the virtue of the soul with unity and harmony—should be conducive to morality. Korsgaard contends that, if morality starts within each of us, if we treat others as we would treat ourselves, internal justice will be wedded to outward justice in the world: 'Legislating for oneself, and legislating for the Kingdom of Ends, are one and the same thing'.

Justice is not conditional, Korsgaard says. We should find harmony from good intentions through Kant's formulae of humanity and universality.

This begins with the rational faculties of human agency.

3. Rationality fails to bind us inescapably to morality

A Kantian, like Korsgaard, claims that morality is *necessarily* found in rationality. This is in contrast to the position of Humeans, whose versions of CC describe morality as *contingent* on one's desires. Rationality, according to Korsgaard, releases us from these inclinations: we are inescapably bound to each other through the rational norms of morality but granted the autonomy to rationally construct universal rules.

I counter: (1) rationality cannot be truly invoked independently and transcendentally outside the sphere of irrational evaluation; and (2) Korsgaard is forced to naively presuppose that value in humanity is built into the procedures of rationality alone; when seeking to unite the two she fails on recursive terms.

These criticisms of Korsgaard's theory (and, indeed, any deployment of rationality in CC) rest in the fact that it fails to account for the emotive components of life. In contrast, Humean CC theorists, such as those of Sharon Street (2012), permit moral agents to use practical reasoning; however, their moral language directly expresses their *evaluative* attitudes towards the various objects in their lives, exposing where Korsgaard's project falters.

3. 1 Rationality is not transcendentally authoritative

A successful CC theory relies on the notion of moral agents, where rationality is but one proclaimed mode of internally representing how the external world ought to be.

Even when rational agents claim to have rationally obtained moral truths, their claims might be no more than the development of sentimental content—causally determined rational representations—explaining why moral claims can be absurd but nonetheless viable (Street 2009). Like any thought process, rational judgements could start as desire-born impressions. I argue this makes rationality morally escapable through the notion of evaluative attitudes, meaning it holds no *transcendental* authority, for it could just be an instrument which ultimately answers to irrationality

Moreover, rational moral agents would be prepared to forsake their self-concerned values (i.e. value in people they most-care about) for rationality to bind us to universal morality. However, the level of rationality necessary to abstract us away from this is not plausibly achievable for human agents, who are imperfect, desire-driven beings with evaluative ties to countless objects in their lives (such as family). So transcendence to a 'noumenal' realm of rational thought is idealised (Kant 2015).

So it is ambitious to suggest that: (a) rationality transcendentally lays down a framework for moral autonomy because rationality is either wired to human desire, else an empty legalism; and (b) rationality underpins a morally efficacious ethical theory because agents would have to be *extremely* rational to meet Kant's high bar of practical reasoning power. For now, though, let us continue to focus on bringing rationality together with the moral content we care about evaluatively because this is something Korsgaard tries to achieve—with rationality in the driving seat.

3. 2 Korsgaard is unsuccessful at building value into rationality

According to Korsgaard, rationality inescapably traps us into valuing humans as ends in themselves. I argue that, while rationality provides us with an intellectual form of autonomous moral agency, it cannot, by itself, necessarily illuminate what a person cares about and what they are motivated to do.

We have to want to be rational—assuming rationality is something we ought to be. Korsgaard, however, maintains that rational moral agents demonstrate value in humanity through Kant's principle of universality, not through these inclinations. But, while rationally born constitutive moral facts can exist in academic veneers of who we think we ought to be, they might not compel us into action. That is, an agent can be aware of what they ought to do, what positions they ought to espouse, and what social movements they should support but, simultaneously, not care to act on this awareness.

Consider, by way of analogy, a person who wishes, idealistically, that they valued the rights of poor people despite not really doing so. They can provide *rationale* for valuing their rights but actually caring is just not part of what they are, for caring is not a purely rational stance. Rational moral agents must possess an original sense of value in humanity to begin with.

But precisely why is this a problem if we are still capable of defining objective moral facts? An agent might rationally concede that killing non-human animals for pleasure is wrong or that sexual abuse is wrong but what does it say about a stance on morality—especially a supposed form of it that starts within us—if the agent does not have to care about the objects of torment and be motivated into action in either case (assuming non-human animals can be treated as ends at all)? Korsgaard's CC potentially leaves a vast space to be occupied by injustice if we, as members of a species who are capable of normative assessments, are not constitutively intertwined with moral fact *and* the very inclinations Kant tries freeing us from. If ethics is the art of living, it seems odd that agents could be licenced to build maxims which were bereft of the agent's *desires* to do good. Metaethically speaking, what is the point in moral facts if our normative claims are just rational expressions of our mental states? What can such an inescapable moral position achieve—what can we construct—when an agent can rationally assert that some facts are true and do nothing about it? Stuck within the language of their cognitive economy, we might not be any closer to justice in the real world.

However, to define moral fact is our metaethical task here: to assess moral-truth-claims; and *it is* metaphysically possible to dichotomise claims to one side and our personal values to the other. Just because someone *feels* indifferent to suffering in the examples above it does not necessarily undermine the fact that they *know* it is wrong. To undermine my own criticisms laid out above, then, justificatory questions of what makes a moral truth are quite different from the question of human constitution, which happens to demonstrably reveal us to be frequently indisposed to 'do good'.

Yet a theory that neglects sentiment unduly misses a piece of the moral puzzle, for sentiment is *deeply constituted in us anyway*, even if passions, desires, attitudes, and so forth do not appreciably manifest in ways we ourselves can recognise rationally. Evaluation plays a significant part in how we see each other through our worldviews existentially and might be tied to 'ought to' (§ 3.1).

Korsgaard, however, does not want to ignore evaluation. She claims rational agents are inescapably bound to morality because the instruments of practical reason commit us to valuing humanity. More specifically, she posits that we can *use* rationality to assess what we value retrospectively (Korsgaard 1996; Korsgaard 2009). Korsgaard opines that it is through our 'practical identities' that we can reason out our emotional obligations to objects in our lives.

Through these identities we can sublimate our inclinations into a mode of finding outward justice: we can rationally choose to value the ends our identities entail. For example, one practical identity could be: 'I am committed to fighting for my country'. This has arisen because of the sources of value to my reasons—patriotism, pride, sense of honour, and so forth—which I have rationally accessed in accordance with my self-perceived identities. A myriad of other roles in my life provide me with other pools of special bonds to reflect on; and through reason I can rationally endorse certain moral positions which fulfil these bonds—so long as they adhere to the principle of universality.

However, take Korsgaard's own example of war. Assuming war is something undesirable, she describes a constitutional democracy whose citizens do not want to wage it. The citizens can rationally conceptualise inward justice: they would not wish war to be waged against themselves and so would not wage it against others. But there is an underlying premise here: that the citizens relate to the other country's citizens. Her thought experiment, then, acts to concede that humans *are* irrationally evaluative creatures who need to be orientated towards others, even when they ostensibly act rationally. The 'altruistic' person running the cancer charity run, for example, is personally chained to the cause, either directly or indirectly: her reasons for 'doing good' are functions of constitutive evaluative starting points.

Hence we should question whether rationality can orientate our value in humanity by itself: why *should* a purely rational agent value humanity? An agent *could* be a rational amoralist, for example, as Bernard Williams (1973) put it when he considered the 'ethical egoist' in the context of altruism. He argued that a rational ethical egoist can exclusively pursue their own interests whilst happily being treated with the same fundamental disregard. As such, the moral agent's rationality does not commit them to morality, which entails valuing other people's interests as a 'basic and universal function of morality'. Being amoral, they cannot constitutively obtain necessary moral facts, for they have escaped this version of morality.[1]

Korsgaard's position necessitates that rational agents value valuing ends in the first place, which *presupposes* value in humanity. This creates a regression of conditions because this is exactly what she wants to delineate. If 'humanity' was replaced by 'being a cool guy' in my value system, consider that I would have to value value 'being a cool guy' in the first place. Both are recursive. Her 'practical identities' are insufficient.

We might, instead, look towards sentimental moral agents, whose moral claims rest on contingent desires. While this does not lead us to necessary moral facts, value in humanity can already be baked into their constitutions. The 'altruistic' agent does not actually act selflessly: they respond to what they already value.

4. Disunified human agency

A cogently described moral agent is a prerequisite of Korsgaard's theory (see quote on page 2). Yet through her CC it is extremely difficult to conceptually collapse the various rationales of a person into a single, so-called-autonomous agent. She, therefore, fails to meet her own challenge. Morality *might* be constituted in what we are but we cannot be rational moral-truth-makers if we are divisible within ourselves; and if there are no such agents, there are no single entities in which morality is constituted.

Korsgaard contradicts the basic notions of human psychology by positing transcendental authority over agency. In the usual psychological realm general agency derives from a sense of self-governance. In this familiar view we are seen as the sums of innate predispositions (e.g. intelligence) which interact as bundles of beliefs and desires with their environments (e.g. culture). Arguably, then, any one of our moral claims is discoverable naturalistically by studying causal chains outside of us or is indiscoverable by virtue of indeterminate randomness. Next, Korsgaard is forced to presuppose the autonomy of rational agents when, in the real world we might not have free will at all (Fischer *et al.* 2007), in which case autonomous rational agency and thus morality cannot be constituted in what we are. But let us assume that psychology as a field is reductionist and that we can be autonomous for the sake of the discussion.

[1]Perhaps, though, we are all egoists and Williams' altruism is a false idealism, which would threaten the integrity of unimported universal moral codes.

Korsgaard struggles to explain how agents are unified at all. We commonly rationalise a plethora of motivations in our heads prior to cementing a moral position (e.g. I want to buy a Starbucks coffee because I will derive pleasure from it; but I also want to not buy a Starbucks coffee to boycott capitalism and tax avoidance). Rifts divide my agency. I assign different weights to the options I present myself with; however, it is not clear how they reduce to one true claim. I argue that a motivation needs to be rooted to a clear-headed, rationally instituted individual who can reliably construct moral knowledge about the world as a single entity without conflict.

These divisions against oneself can go further, too, when we consider repression, compartmentalisation, something akin to Freud's 'unconscious', and any other parts of our being which can drive our decision-making without being called rational. As examples, consider the agent who loves the planet and wishes it safe from environmental catastrophes but who chooses to stay alive and damage it; the philosopher who proselytises pessimism whilst encouraging people to desire their value system; and the family man who really wants to have a few drinks with his friends tonight instead of looking after his children but knows he ought not to. We witness contradictions to unified human agency every day.

Reflect on it like this. Agent A holds multiple motivations, X_n. However, for A to be entitled to autonomy in virtue of rationality and to be free of subjectivity, Korsgaard needs to clearly express why X_i rationally prevails over another X_j such that A would always commit to it. If the rug (coherence), so to speak, is pulled from under the very agent supposedly capable of rational moral-truth-making, there cannot be an agent who can coherently construct morality from within themselves in order to care for the people under the wings of their practical identities, in public spheres, and then across all possible worlds.

Though Korsgaard tackles these in 'Dealing with the Disunified' she is self-defeating in rehearsing the thoughts of various philosophers. Through Derek Parfit's arguments (Parfit 2007), for instance, she discusses the idea that we can be divided against ourselves, illustrated by the Russian nobleman who wants to commit his future self to his *present* values by ensuring his wealth distributed accordingly in his old age, even if against his will. The nobleman will maintain the capacity to reason—but his tools of reason may dip into different pools of motivations in accordance with his new practical identities. Time makes fools of us all: its continuum ruptures the ideas we have about the ourselves and others. As Parfit reasons, it is difficult to guarantee we can be the same person with the same identity over time; so how can there be a definitive moral agent straddling it all?

Time, additionally, threatens our commitment to humanity through our personal relationships. Korsgaard utilises Kant's arguments for marriage by deploying the 'unity of will': making someone your end so you belong wholly together, possibly as a metaphor for the commitment to the CI. But this is idealistic. The nobleman should recognise that his wife cannot, with integrity, guarantee to fulfil his promises by proxy, for both of their value sets, though procedurally born, can fracture later on. If, for example, someone asks us to oblige a deathbed wish, disunity would soon be at play again: where is the other person to commit values from? We could endorse their wish there and then but our commitment *to the agent* is metaphysically eradicated by their passing. Something similar happens when our relationships mutate such that we conceive new values or question the commitments we made to them in the first place—say a promise to a best friend in the remit the 'best friend' practical identity. Values, including those committed to humanity, change; and as an agent with multiple transient desires at all times, any desire to make a commitment would be at risk of being usurped by another future desire, creating instability for Korsgaard's autonomous, rational moral agent. The notion of a unified agent who can develop robust normative claims thus slowly erodes by itself.

Assertions about what is right and what is wrong cannot be traced to a single unified agent, while their claims might not survive along with 'them'. At best, then, an agent, in Korsgaard's sense, is only coherent for a single, static timepoint, representing a mere flash of autonomy.

In conclusion

There are just too many problems for Korsgaard's CC. Rationality *might* tie us inescapably to morality in academic terms; but rationality is too much of an escapable concept: it crumbles because we are unavoidably evaluative and usually divided against ourselves. Moreover, if 'legislating for oneself, and legislating for the Kingdom of Ends, are one and the same thing', Korsgaard's CC struggles to get off the ground: without a coherent view of inward justice we cannot have coherent view of outward justice.

Bibliography

Pereboom, D. 2007. Hard Incompatibilism, ed. JM Fischer, R Kane, D Pereboom, and M Vargas, *Four Views on Free Will*, Oxford: Blackwell Publishing.

Kant, I. 1998 [1785]. *Groundwork of the Metaphysics of Morals*, trans. M Gregor, Cambridge: Cambridge University Press.

Kant, I. 2015 [1788]. *Critique of Practical Reason*, trans. M Gregor, Cambridge: Cambridge University Press.

Korsgaard, C. 1996. *The Sources of Normativity*, ed. Onora O'Neill, Cambridge: Cambridge University Press.

Korsgaard, C. 2009. *Self-Constitution: Agency, Identity and Integrity*, Oxford: Oxford University Press.

Parfit, D. 1971. Personal Identity, *The Philosophical Review*, 80(1), 3–27.

Street, S. 2009. In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters, *Philosophical Issues*, 19, Metaethics, 273–298.

Street, S. 2012. Coming to Terms with Contingency: Humean Constructivism about Practical Reason, ed. J Lenman and J Shemmer, *Constructivism in Practical Philosophy*, Oxford: Oxford University Press.

Williams, B. 1973. *Problems of the Self*, Philosophical Papers, Cambridge: Cambridge University Press.