DOI: 10.1111/phpr.13065

Philosophy and Phenomenological Research

ORIGINAL ARTICLE

Internalizing rules

Spencer Paulson

Northwestern University, Evanston, Illinois, USA

Correspondence
Northwestern University Evanston, Illinois, USA.
Email: spencerpaulson2023@u.northwestern.edu

Abstract

The aim of this paper is to give an account of what it is to internalize a rule. I claim that internalization is the process of redistributing the burden of instruction from the teacher to the student. The process is complete when instruction is no longer needed, and the rule has reshaped perceptual classification of the circumstances in which it applies. Teaching a rule is the initiation of this process. We internalize rules by simulating instruction coming from someone else. Running these simulations enables us to toggle between the perspective of the instructor and our own perspective. By doing this we coordinate our perspectives with that of the teacher. The account given here provides a deeper explanation of why internalizing a rule involves the dispositions and reactive attitudes proponents of Rule Consequentialism often say it does, why moral reflection is variably demanding, how intergenerational moral progress is made possible by our cognitive architectures, and why the adoption of a rule should be understood in terms of teaching that rule.

KEYWORDS

Adoption, Cultural Evolution, Developmental Psychology, Internalization, Metacognition, Rule Consequentialism, Social Cognition

> "Her grey, sun-strained eyes stared straight ahead, but she had deliberately shifted our relations, and for a moment I thought I loved her. But I am slow-thinking and full of interior rules that act as brakes on my desires." – F. Scott Fitzgerald, *The Great Gatsby*

Rule consequentialists typically begin with the idea that the authoritative moral rules are the ones we ought to collectively adopt.[1] Which rules we ought to collectively adopt is determined by the consequences of adopting them. The debate then turns to, among other things, the question of how we should understand adoption. Is adopting the rules merely complying with them? Is it internalizing them into our motivational structures? Is it teaching them?

The discussion of internalization itself, however, has been comparatively sketchy. The same is true of teaching, even by the admission of one of its most prominent advocates (D. Miller, 2021, p. 129). In both cases, the author tends to gesture at the sort of thing they have in mind rather than giving an account of its nature. Dale Miller, for example, tells us, "An agent who has internalized a moral code has a psychological disposition of some sort that gives her a motive to obey its rules." (D. Miller, 2014, p. 150). Of what sort? He says,

> "Internalizing a moral code is commonly taken to involve being disposed to feel compunction prospectively when one considers violating its rules and to feel guilt after the fact when one knows one has done so. Brandt, for example, says that adopting a moral code usually means having intrinsic motivation to obey it, feeling guilt when one violates the code oneself and disapproving of others when they do so, believing that acting in accordance with the code is important, esteeming others who are motivated to comply with the code to an unusual degree, using special terminology like 'morally ought' in connection with the code, and believing that these motivations, feelings of guilt, feelings of approval or esteem, are justified." (Ibid)

This seems right so far as it goes, both as a description of internalization and as an overview of the literature, but it doesn't provide much of an explanation.[2] It tells us what kinds of mental states are usually involved, but it doesn't even try to give us the nature of internalization. It gives us, at best, a list of diagnostic criteria but no theoretical account of how they hang together. Brad Hooker, for example, tells us that a moral code has been internalized when one has a moral conscience "of a certain shape" (Hooker, 2000, p. 91; cf. Brandt, 1979, p. 164–76). This may give us the nature of internalization at a certain level of abstraction (it consists in conscience-shaping), but the details are not forthcoming. What does it mean to shape one's conscience with a rule? Much the same concern applies to discussions of teaching. Dale Miller tells us that moral teaching takes place in families, schools, churches, and the media (D. Miller, 2021, p. 129), and that it involves a variety of techniques (Ibid), but we are not told how whatever happens in those places by way of these (unspecified) techniques sculpts the psychological propensities of those educated or why that kind of shaping is the relevant kind. Timothy Miller tells us that "teaching a moral code is to get it internalized and acted upon" (2021, p. 208), but this doesn't get us very far unless we already know what internalization is.

[1] See Hooker (2000, p. 1) for an influential statement of this approach. See Blackburn (1998, p. 281); Copp (1995, p. 112); Gert (1998, p. 9); and Mackie (1977, p. 87; 152) for variations on this idea.

[2] Hooker (2000, p. 76) says the same thing, Holly Smith (2010) appropriates his formulation noting its prevalence in the literature, Kevin Tobia (2018) seems to be doing this as well. Michael Ridge says that a rule is internalized when it plays a role in one's motivational economy (Ridge, 2006, p. 243), though he doesn't say which role.

This is unfortunate, in no small part because much of the appeal of Consequentialism generally comes from the precision and articulacy it allows us to bring to bear on moral philosophy. Consequentialism is motivated in large part by the desire to give a unified explanation of our considered moral judgments.[3] If our considered judgments are explained in terms of the consequences of adopting moral rules and adoption is identified with a heap of loosely related dispositions, judgments, and reactive attitudes, then Consequentialism hasn't yet realized its potential.

I will remedy the situation by giving an account of internalization: the process by which the instructions of a teacher are gradually replaced by self-instruction (initially through overt speech and later through simulations of it in inner speech) and culminating in perceptual learning that non-inferentially classifies situations in compliance with a rule. Since, as Timothy Miller says, teaching someone a rule is getting them to internalize it, it follows that teaching a rule is the initiation of the process just described.

To make my case, I will draw on developmental psychology and neuroscience. The resulting account will be empirically constrained without being merely empirical. Internalization is (essentially) the process of replacing external instruction by its internal counterpart. That is a philosophical claim. I supplement it with empirical claims to demonstrate that this is not only something humans routinely do; it is crucial to our cognitive development.

Several significant results follow from my proposal. The first is that the process of internalization, as I understand it, gives a unified account of how the medley of reactive attitudes and dispositions typically invoked to fix the reference of "internalize" hang together to form a single, scientifically significant process. The way they hang together helps explain why moral deliberation is characteristically reflective while also, sometimes, automatic, and reflex-like. This ensures that the account neither over- nor under-intellectualizes the phenomenon. The second is that since internalization consists in teaching oneself (except for the limiting endpoint of the process when instruction is complete), we ought to favor the teaching-centered construal of Rule Consequentialism that has emerged recently (though with some caveats mentioned below). Finally, my account of internalization provides a deeper explanation of how intergenerational incremental moral progress in the form of gradually revised moral codes can be realized in creatures with our psychological architecture. This helps us better understand the "moral spiral" (Skorupski, 1989) at which rule consequentialists have long been gesturing.

In section (I) I give the psychological underpinnings of my account. In section (II) I explain the relevance of the reactive attitudes and how our simulation of moral instruction makes moral deliberation variably demanding. In section (III) I explain how internalization is conceptually related to teaching and why this supports the teaching-centered formulation of Rule Consequentialism (subject to a caveat I discuss below). In section (IV) I explain how my account provides psychological mechanisms capable of implementing Skorupski's moral spiral.

I | INTERNALIZATION: A DEVELOPMENTAL PERSPECTIVE

Teaching is, paradigmatically, an interpersonal activity. There are two roles: the teacher and the student. In the paradigmatic case, those roles are played by different people. Internalization is the process by which the student takes on the task of the teacher by toggling between the roles.[4] As

[3] Cf. Hooker (2000, Chapter 1).

[4] This use of the term originates, to the best of my knowledge, with Vygotsky (1978). In recent years it has figured prominently in the influential research program of Michael Tomasello (1999; 2014; 2021).

the process unfolds, the amount of cognitive effort required for self-instruction decreases. Overt self-instruction is replaced first by inner speech and, eventually, by perceptual re-classification of situations in which the rule being taught applies.

Consider an adult teaching a small child to perform a task according to a short series of instructions. We can imagine a child who has seen a fair amount of baseball on television who is trying to imitate the swing of a bat. The teacher might say things such as, "First you grip the handle with your hands together, then you get in your stance…". In early childhood the child doesn't typically understand all the instructions at first, so the role of the teacher is to demonstrate what each of them means by adjusting the child's body to conform with the instructions while they are spoken aloud. The teacher might, for example, say, "First you grip the handle with your hands together, oops, not quite they've got to be right next to each other, there we go good job!" while sliding the child's hands together into the correct position. This helps the child associate the verbal instructions with the corresponding bodily position. By jointly attending to things in the world (in this case, position of the body relative to the bat), the child infers the intentions of the instructor.[5]The inference is made possible by the interaction of several relatively low-level abilities and attentional biases. These include a preference for attending to face-like stimuli (Johnson et al., 1991; Johnson & Morton, 1991), a predisposition to distinguish self-propelled motion from other kinds of motion (Massey & Gelman, 1988; Premack, 1990), and a predisposition to associate pointing with line of gaze (Butterworth, 1991; Baron-Cohen, 1991).[6]

What happens when the activity is performed in the absence of the instructor? Two- and three-year-old children are prone to mimicking adult instructions while performing an activity, but they nonetheless behave in a way that disregards those instructions (Luria, 1961). They associate the instructions with the activity because those are the instructions they've received in the past when they were trying to perform that activity. So, they repeat those instructions at the time of the activity. However, they haven't yet learned to treat those instructions as instructions.

At 4–5 years-old, however, they not only mimic the instructions (overtly at first) but also coordinate their behavior with the instructions (Tomasello 1999, 192). Now they not only associate the verbal performance of the instructions with the activity, but they are also beginning to associate those instructions with the role of the instructor even when the instructor is not present. The key ontogenetic development at this stage is the ability to toggle between the role of an instructor and the role of the student.[7] This is sometimes called "role reversal imitation".[8] At this point the child's mindreading abilities are sophisticated enough to discern the communicative intentions of the instructor. By attending to elements of the non-linguistic environment with the instructor, the child came to recognize that the instructions were themselves a goal-oriented action meant to guide the child's performance. This is not to say that the child has a discursive understanding of the instructor's communicative intentions, but rather that they have registered enough low-level information to simulate the mental states of the instructor. In so doing they imaginatively project themselves into the role of the instructor as they give themselves instructions. They then switch back into the role of the student and respond to the instructions they just simulated giving

[5] Cf. Bakeman & Adamson (1984); Bates (1979); Corkum & Moore (1995); Tomasello (2014, Chapter 2).

[6] See Karmiloff-Smith (1992) for a helpful overview. Note that the abilities mentioned are low-level relative to social cognition, but not relative to perception. They are much higher-level than detection of edges, vertices, etc.

[7] Cf. Bakhtin (1981); Fernyhough (1996); Wertsch (1991).

[8] See Meltzoff & Moore (1977); Meltzoff (1995); Tomasello (2003, p. 25–8); see Tomasello, Kruger & Ratner (1993) for further discussion.

themselves. I will say a bit more about how simulations are realized in mammalian brains once the basic developmental story is in place, since that will be relevant in the following section.

There is independent evidence that the ability to toggle between complementary roles comes online at this point in ontogeny. For instance, it is at around this time that children engaged in a joint activity with another child will stop and help their counterpart when difficulties are encountered.[9] This happens when they are playing complementary roles in the shared activity, suggesting that the child who stops to help has a "bird's eye view" of the activity (Hobson, 2004; Tomasello, 2014; p. 41) that enables her to see both roles as distinct components of a shared undertaking.

At this point, the instructions are beginning to be internalized. Not only does the child conform to the instructions, but she also recognizes them as authoritative by simulating the mental states of someone issuing imperatives. It is because of this simulation that she begins to reliably conform to the instructions (unlike her 2-year-old counterpart). At around the same age, (5-7 years), children regulate the behavior of others not long after having learned the instructions themselves, thus more fully occupying the teaching role (Ratner & Hill, 1991; Foley & Ratner, 1997). This makes sense: self-regulation is, in a sense, teaching oneself (Ashley & Tomasello, 1998; Tomasello, 1999, p. 193). After teaching oneself it is a small (but non-trivial) step to teach others.

Here is Tomasello drawing out the consequences,

> Children thus show relatively clear evidence of internalizing adults' regulating speech, rules and instructions as they are reaching the later stages of the early childhood period. What is internalized is, as Vygotsky emphasized, a dialogue. In the learning interaction the child comprehends the adult instruction (simulates the adult's regulating activity), but she does so in relation to her own understanding—which requires a coordinating of two perspectives. (Tomasello, 1999, p. 193).

Crucially, at a certain point in ontogeny the child acquires the ability to "reflect on her own behavior and cognition *as if* they were another person looking at it" (Tomasello, 1999, p. 196, his italics). They then coordinate their own cognition (and, derivatively, behavior) with that of the perspective they've just simulated.

The simulation of instruction is easily observable in early childhood because it involves overt speech. As ontogeny unfolds the same basic pattern continues, except the overt speech is often stifled and replaced with inner speech. When children rehearse instructions to themselves out loud, it is usually because they were learning to do something difficult (Goodman, 1984). The same is true when the instructions have been relocated to inner speech. What happens when repeated self-instruction has made the task easy? At this point perceptual learning has taken place: the process of self-instruction has re-shaped your immediate, non-inferential perceptual classification of the situations in which the rules you've taught yourself are applicable.[10] As Tomasello & Call put it, "once an organism has 'thought through' a problem, its future encounters with the same or similar problems may show insight and foresight on the immediate perceptual level" (1997, p. 11).

Let us take stock of the key takeaways that have emerged so far. The first is a big picture consideration: some of our higher or, as it sometimes put, "executive" mental functions take the form of self-instruction. We regulate ourselves by stepping back and reconsidering prospective actions from the perspective of an instructor. Instructors give us instructions or rules of the form "When in

[9] cf. Hamann et al. (2012).

[10] This is not to be confused with cognitive penetration (cf. Siegel, 2012). Cognitive penetration is synchronic, perceptual learning is diachronic. See Jenkin (2020) for further discussion.

C, do A". Our cognitive maturation involves retraining our response to C by internalizing relevant instructions. This process has several stages. The first is mere imitation of the instructions. This is a necessary precondition for internalization, although not sufficient for it. Internalization begins in earnest when the subject becomes capable of toggling between the (simulated) perspective of the instructor and her own perspective, adjusting for discrepancies by changing her actions. At first this involves overt speech mimicking the instructor while simulating the instructor's psychological states and then switching back to the role of the student and responding to those instructions. Later, it takes the form of inner speech that simulates instruction without overtly reproducing it. In both cases internalization of the rules has already taken place to some extent, but the process of internalization is incomplete. The rules have been to some extent internalized since the subject has taken on the role of the instructor. However, it is incomplete because instruction is still necessary: she still needs to remind herself of what to do in C. The rule has been fully internalized when instruction is no longer necessary and her automatic response to C is to do A. At this point, the rule has shaped her perceptual classification of C, and she does A without needing to deliberate.

Although this process is typically studied in the context of early childhood development, it persists into adulthood. In childhood I internalized a series of instructions that enabled me to tie my shoes. Those instructions are fully internalized; I can tie my shoes now without deliberating. However, there are other rules that I have only internalized partially. For instance, I am capable of reading music, but I am incapable of sightreading. If given sheet music, I can play it. However, I often need to rehearse the rules and associated mnemonic devices to remember which lines of the treble clef correspond to which notes. Someone who can sightread, on the other hand, can read the treble clef without inference just as they read printed words. They have fully internalized the rules. Something similar applies to chess (cf. Jenkin, 2020). Novice players need to rehearse strategic rules while playing, whereas grandmasters immediately perceive their applicability.

This is important because internalization, even in adulthood, is an ongoing affair. For many mathematical rules (e.g., order of operations) cognitively mature adults need to rehearse a rule in inner speech to apply the rule. Only in the limiting case in which teaching is no longer necessary has perceptual classification completely replaced deliberation.

Before moving on to the next section I will say a bit more about how simulations work. The basic idea is that one system simulates another just in case the former mimics the latter. In a computer simulation of a weather event there will be representations of weather-states and their relations. If the simulation is veridical, then the syntactic roles of the representations are isomorphic to the causal relations between the states.[11]

In the case of a computer simulation, the representations are amodal. That is, they have been transduced from the environment (either via keyboard input or connection to measuring instruments) and their current representational format is unrelated to their origins. Computers store and manipulate information in a digital code that differs from the representational format of the input (e.g., the analogue representational format of meters). A common view is that something similar applies to human cognition.[12]On this view, perception has its own code with analogue format whereas the higher cognitive functions traffic in a different code with a digital format (cf. Dretske, 1981). Images registered in analogue format must be converted to something else (e.g., feature lists stored in binary code) before central processing takes place.

[11] The kind of isomorphism I have in mind is a structure-preserving bijection.

[12] Noteworthy proponents include Burge (2010); Carey (2009); Fodor (1975); Newell & Simon (1972); Pylyshyn (1984); Dretske (1981).

The things I have said so far are consistent with this position. I said that certain simulations take place at certain stages in development and that they enable us to perform various tasks, but I didn't say how. However, there is another way of thinking about the simulations according to which they are housed in the sensory systems of the brain. On this view, sensory information is not converted into a cognitive lingua franca before being brought to bear on the higher cognitive functions. Rather, these cognitive functions are realized in "perceptual symbol systems" (Barsalou, 1999). The symbols manipulated in the simulations are realized in the neural populations of sensory systems in the brain and the higher cognitive functions are performed by integrating them with domain-specific knowledge to achieve competent simulations (Ibid). The higher cognitive functions work by partially reactivating the neural populations that encode sensory representations and manipulating them internally to draw inferences. These representations are multi-modal, integrating aspects of experiences encoded by all five senses as well as proprioception and introspection. They are also schematic: they don't contain all the information from the original experiences but rather condense them into informationally significant fragments to store in long-term memory. These fragments are later retrieved and treated by the cognitive system as a symbol. Their sensory character does not preclude their behaving in many respects like amodal symbols in a formal language. Recursive operations are defined over them in virtue of their symbolic character (cf. Langacker, 1986; Barsalou et al., 1999).

Although it does not entail the perceptual symbol system architecture, much of the work in developmental psychology is done by researchers who are sympathetic to it.[13]The reasons are never, to my knowledge, discussed explicitly, but I believe they have much to do with the apparently situated and embodied character of rule internalization. It is probably no accident that amodal representations first became popular in research programs trying to produce computers that can play chess and perform other similarly dispassionate tasks. There is perhaps an abstract and dispassionate way of cognizing rules generally, but it doesn't appear to be how we do it. It is of course possible to set up interfaces between an amodal symbol system and an affective system, but a simpler way to forge the link with the affects is to do everything by repurposing parts of the brain that are already trafficking in them.

Similarly, there is considerable evidence that our ability to recognize patterns in the behavior of others recruits our ability to mirror their emotions and our ability to mimic their behavior recruits our ability to mimic their affective states.[14] Biologically, this is what we should expect. Some of the most evolutionarily ancient mental capacities are dedicated to conative cognition (cf. Schulz, 2022) and sensory representation. Evolution is an incremental process that implements new traits by tweaking what is already there, so we should accord a high prior probability to the hypothesis that human-specific rule-based cognition is implemented by integrating these ancient capacities.[15]Relatedly, neural reuse is an important organizational principle of the brain.[16]New cognitive functions are often realized by re-purposing neural populations already performing other tasks. This is just another case of that general pattern.

Furthermore, it has long been known that damage to a sensory system decreases performance on categorization tasks for categories primarily perceived through that sensory system (Damasio, 1989; Kosslyn, 1994; Pulvermüller, 1999). For example, birds are typically (by us) perceived visually.

[13] For example, Tomasello (2003; 2014).

[14] See Barsalou (2003) for a useful overview.

[15] This style of argument is common in cognitive science. See for example Brooks (1991); Carruthers (2006); Clark (1989); Tooby & Cosmides (1992).

[16] See Anderson (2010).

When our visual system is damaged, our ability to think conceptually about birds suffers (Barsalou, 1999, p. 579). This would be a surprising coincidence if the conceptual system were amodal and, consequently, separate from the damaged sensory system. This gives us (compelling but not conclusive) evidence that at least some important aspects of our higher cognitive functioning are realized in perceptual systems.

Here is a sketch of how I think it works in the cases that matter for rule-based instruction. It is generally well-known that category instances have features that statistically covary.[17] As a result, different perceptually encountered instances of the category have similar features and perceptions of them activate similar "feature maps" (Farah & McClelland, 1991; McRae & Cree, 2002). Modality specific neural populations code for features detected by that modality: auditory neural populations have receptive fields or "tuning curves" that resonate to features of auditory stimulation such as pitch, *mutatis mutandis* for neural populations in the visual cortex.[18] Each modality has "feature neurons" that resonate to different modality-specific features, but the brain also has "convergence zones" (Damasio, 1989; Man, et al., 2013; Walsh & Oakley, 2022) populated by "conjunctive neurons" (Manohar et al., 2019). Conjunctive neurons are highly plastic and can be molded through training to resonate to combinations of activity in the feature neurons.[19] Since instances of a category have properties that statistically covary, conjunctive neurons are trained to covary to category instances by responding to the conjunction of their features. Neural activity in convergence zones integrates category-relevant information from various modalities (including introspection and proprioception) into a single representation (Barsalou, 2013). By integrating affective responses registered by introspection with the detection of low-level features relevant to mindreading mentioned above, the perceptual symbols in convergence zones enable the subject to mirror the affective states of the instructor and associate them with performance of the task being learned. Simulations are run by manipulating those perceptual symbols.

It is important not to confuse my account with a related one, according to which we determine the mental states of others by simulating what we would do if we were in their situation with their beliefs and desires (cf. Harris, 1991; Goldman, 2006). Rather, we figure out their mental states with an interacting set of low-level abilities. The information is then stored in perceptual symbols which facilitates its use in simulations. These simulations are used to regulate our behavior.

As I said above, perceptual symbols are schematic. They retain compressed representations of their referents. For instance, a representation of a penny in memory might retain information pertaining to the color, texture, approximate size, and which President is depicted, without containing information pertaining to the year it was minted or the direction the President is facing (cf. Kosslyn, 1994).

Similarly, when I simulate rule-based instruction, I do so schematically. I might not include the location of instruction, the time of day, or in some cases even the identity of a particular instructor. Nonetheless, I include enough to draw certain important inferences (e.g., if I were to violate the rule, the instructor would be upset). When the developmental process is nearing completion, this omission becomes important, particularly in the context of moral rules. Suppose I am considering telling a lie to get out of mowing the lawn. I can simulate moral instruction forbidding that act. I might simulate it coming from a particular authority (e.g., my mother), but I might not. Part of what one learns to appreciate when the developmental process is in its more advanced stages

[17] See Rosch & Mervis (1975) for an early discussion that has received much subsequent attention.

[18] Cf. Kosslyn (1994).

[19] Hebbian learning alone may be sufficient but see Damasio (1989) for an alternative view.

is that the moral rules are available for anyone to appeal to in forming a criticism.[20] So, even if my simulation involves a perceptual representation of my mother instructing me about a rule prohibiting lying, I could be in effect using her as a stand-in for any rational agent. Compare: I can use a picture of the Empire State Building to represent large buildings in general (Barsalou, 1999, p. 584).

I am going to draw on the perceptual symbol system hypothesis in some of what follows. As with nearly any other empirical theory, there are reasonable and well-informed people who doubt it (see fn. 12). This means I am not offering a knock-down argument. I am comfortable with this. Knock-down arguments are, by my estimation, quite rare in Philosophy. They are even less common in empirical disciplines and, sooner or later, consequentialists need to make empirical claims. If the empirical work of the future refutes the empirical work I invoke here, then the account here needs to be replaced by one that marches in step with the correct empirical theory. This paper might still be a success so long as it sets a precedent for future work on rule internalization that is constrained by the best empirical work available and clear about the philosophical implications of that work.

II | UNIFYING THE REACTIVE ATTITUDES

When discussing rule internalization, proponents of Rule Consequentialism tend to indicate what they have in mind by enumeration, as was seen above. That is, they give a list of reactive attitudes, judgments, etc. people tend to have when they have internalized a rule. Our understanding of internalization could be improved with an account that includes not only the components of internalization, but an account of the state itself that explains why those components together compose a non-gerrymandered process-type.

We now have the necessary materials for such an account. Since different authors give different lists of reactive attitudes, judgments, etc. and the lists are sometimes rather long, I will not be able to cover all of them. Instead, I will cover a few that are frequently mentioned. I will explain how they relate to simulated instruction. This will give a sense of how the list could be extended to cover further cases.

Dale Miller says, "Internalizing a moral code is commonly taken to involve being disposed to feel compunction prospectively when one considers violating its rules and to feel guilt after the fact when one knows one has done so." This seems right and we can now say a bit more about why.

The ability to simulate the evaluations of oneself coming from someone else is the foundation of a moral conscience (Tomasello, 2021, p. 281; cf. Hardy & Carlo, 2005). This ability underpins the internalization of rules generally, as was argued in the last section. Moral rules are a special case. Moral evaluations involve reactive attitudes that aren't common to rule based instruction in general.[21] This is because moral decisions involve more than just a "me-concern" (Tomasello, 2021, p. 288). If I break a grammatical rule, the consequences don't concern others (at least not generally). If I break a moral rule, perhaps one forbidding embezzlement, then things are different, and others have a stake in the matter. Since others are affected, their evaluations involve reactive attitudes. When one is disrespected, for example, one evaluates the agent of the disrespectful act with resentment and indignation (Ibid). When those reactive attitudes are internalized through

[20] Tomasello (2014, p. Chapter 3) develops this point at length.

[21] Thanks to Dale Miller for encouraging me to be clear about this point.

simulation, one feels guilty for an act one has already performed or compunction for an act one is considering prospectively (Tomasello, 2021, p. 281ff.).[22]

The reactive attitudes Tomasello mentions are sometimes thought to amount to blame, in which case one feels guilt or compunction because one has internalized blame (cf. Wallace, 1994, p. 75).[23] The reactive attitudes one is simulating may or may not be those of the victim. The criticism that one has broken the rules is one that is available to anyone in the moral community that abides by those rules. So, anyone in the moral community can play the role of the instructor and act as emissary of the group (cf. Tomasello, 2014, Chapter 3).

To see how this works, let us consider the disposition to feel compunction prospectively when one considers violating a moral rule. Suppose I have partially, but not completely, internalized a rule forbidding embezzlement. I haven't internalized it well enough so that the possibility of embezzlement no longer occurs to me, but I have internalized it enough so that I am prone to simulating self-instruction when it does. I am not yet virtuous, but enkratic. Imagine an occasion on which I feel inclined to embezzle funds from a charitable organization of which I am the treasurer. Upon feeling the inclination to appropriate the funds for my own use, I simulate a stern lecture from a moral instructor. I am simulating the kind of lecture I would have received in childhood. The lecture is coming from an authoritative source.[24] Furthermore, the authoritative figure is disappointed in me. The typical instance in which someone instructs someone else with a moral rule is when that rule is violated, and this is the kind of case I am simulating. As we saw earlier, the internalization of rules works by simulating not the mere utterance of words, but rather a dialogue. I am not just recalling the words of the instructor but toggling between my perspective and that of the instructor. I am simulating the mental states of the instructor as I simulate the utterance of the rules, in effect imagining viewing my own (prospective in this case) action from the perspective of the instructor, which includes the indignation that accompanies the lecture. Indeed, the convergence zones that integrate the sensory information needed to recreate the behavior of the instructor also mirror their affective responses (including reactive attitudes).[25]Mill thought that discomfort should be imposed on those who transgress the rules, "if not by law, by opinion of his fellow creatures; if not by opinion, by the reproaches of his own conscience" (1861/1998, Chapter 5, para. 14). Internalization is the process by which the opinion of his fellow creatures is replaced with the reproaches of his own conscience. The reproach of conscience is fit to replace the opinion of others because it works by mimicking it.

Crucially, the opinion of his fellow creatures (at least in the context of instruction) is taken to be authoritative. It is no coincidence that people frequently simulate verbal instruction from their mothers, as the authoritative source of the instruction is crucial to the development of metacognition (Kontos, 1983). I feel compunction because I have internalized an authoritative condemnation that involves resentment and indignation. I have translated the reactive attitude-involving condemnation of an authoritative critic into how I regard myself by representing my own performance and the critical response to it in the same convergence zone. The same

[22] Tomasello explicitly considers guilt only, but a similar treatment applies to compunction since it is just like guilt but experienced in advance of performing an action.

[23] This might be the case if to blame someone (as opposed to merely judge that they are blameworthy) is to target them with certain reactive attitudes (see Wallace, 1994; 2011, cf. Darwall, 2006). However, it might also be the case if blame itself is characterized by a certain role and the reactive attitudes are (contingently, perhaps) capable of playing that role (see McGeer, 2014; Smith, 2014).

[24] Recall that it there needn't be any particular authoritative source such that it is coming from them, however.

[25] Barsalou (2003; 2013).

applies to why I feel guilty after the fact. The only difference is that I am simulating an assessment of an action that I have already performed rather than one I might perform but haven't yet.

Richard Brandt (1979, p.164-76; cf. Hooker 2000, p. 91) says we feel intrinsic motivation to obey rules we have internalized. We can now see why that is the case. This is because of the role that self-regulation by the internalization of rules plays in executive functioning. The paradigmatic cases of executive functioning are those in which we step back and reconsider our inclinations. The ability to delay gratification, for example, is a clear case of executive functioning.

In the cases of interest, we step back and reconsider our inclinations by checking their compatibility with a rule. By running simulations of rule-based instruction and using them as the benchmark for self-assessment, we take for granted the to-be-done-ness of the rule. The rule is the yardstick by which we measure our actions, so it is itself beyond question. This is not to say that we can't reconsider the rules we have internalized. Our ability to do so will be the topic of the final section. However, insofar as the act of reconsidering things in general takes the form of stepping back and assessing them in light of rules, we will simply have to take for granted the to-be-done-ness of some other rule(s) to do so. Some rule or other, then, will always be the highest court of appeals. Our motivation to comply with that rule will be intrinsic, at least provisionally.

Finally, I will consider why we feel esteem for others who are motivated to comply with the rules we have internalized. As we saw in the previous section, once you have internalized a rule, you have learned to play the role of the instructor. Consequently, it is a short step to instructing others and, as we would expect, the instruction of others falls hard on the heels of self-instruction in the developmental sequence. Consider the role of positive reinforcement in early rule-based instruction. The instructor doesn't just make note of deviations from the rule, but they congratulate conformity to it. Once one has stepped into the role of the instructor, one has acquired the reactive attitudes that one was experiencing derivatively when one was simulating instruction coming from someone else. So, when you commend others for their conformity to the rule, you experience positive reactive attitudes. Similarly, when you imagine (i.e., simulate) yourself commending them, you experience those same reactive attitudes.

Before moving on it is worth noting how the episodes described so far can involve varying degrees of conscious effort. This is important in part because it helps show how the account proposed here can avoid either under-intellectualizing or over-intellectualizing self-regulation with internalized rules. It is important not to under-intellectualize the phenomenon since it is our capacity to reflectively apply the rules that makes us responsible for our violations of them. At the same time, we don't engage in explicit deliberation every time we apply them. Much of the time we apply them automatically. We are nonetheless responsible for our conduct at these times.

The account on offer here explains why rule-based self-regulation is a characteristically reflective capacity that can nonetheless be employed unreflectively. This secures the best of both worlds. In short, the reflective and unreflective moments of our moral lives are distinct but continuous moments of a single process.

At the beginning of the process, explicit instruction is coming from someone else. The process of internalizing that instruction comes when one undertakes the pedagogical burden oneself by acting out that verbal instruction oneself. At this stage one needs to recreate as much of the context of instruction as possible to properly respond to it, this is why one needs to recreate the auditory stimuli. The developmental process continues when overtly acting the instructions out is replaced by running offline simulations of them. As we saw above, this involves schematic representations that leave out a great deal of information and, consequently, require less conscious attention. Instructing oneself is a reflective act. However, the amount of conscious effort

necessary decreases as the task becomes more familiar.[26] The process culminates in a state in which no conscious effort is necessary, and the rule is manifested in one's immediate perceptual classification of the situations in which the rule applies.[27] If the rule is "In C, do A", then one immediately perceives situations where C obtains as situations as where one does A. The rule is still guiding your behavior, the difference is that its influence has now spread to perception itself as opposed to being confined to downstream cognitive processing.[28] The upshot is that one is still performing a characteristically reflective task: applying a rule. One is just doing it in an unreflective way.

We can think of it in the following way. The earlier stages of rule internalization require one to recreate the circumstances of instruction by audibly recreating the instruction itself. One then responds to the instruction one has recreated as if it were issued by an authoritative teacher because one is simulating the reactive attitudes of an authoritative teacher while one does this. As we become more comfortable applying the rule, we no longer need to recreate canonical instances of learning by instruction in their full force and vivacity. Rather, we can get by with an interior, schematic, and somewhat muted recreation of the canonical instances. We are still reflecting. We are still consciously attending to our prospective actions and thinking about whether any adjustments are necessary to bring them into line with the rule. The act is reflective, but we do it with less conscious effort which makes our performance of it, in a sense, less reflective (at least insofar as "reflective" is to be contrasted with "automatic").

At the final stage of the process our application of the rule has been crammed into our perceptual classification of the circumstances in which the rule applies. We can think of our perception as an abridged version of the simulations we used to run to reach the same conclusion (i.e., we need to do A). We are no longer reflecting, but we are nonetheless engaged in a sparse recreation of lessons learned by reflection. Think of the process like a baroque musical motif that is being played repeatedly. Each time the motif is condensed a bit; some of the ornamentation is left out and greater emphasis is placed on the important notes.[29] The song ends with a two-note refrain that only includes the first and the last note of the motif. Hearing those two notes in this context, we hear the rest of the baroque motif latent in them. Much in the same way, overt self-instruction gives a schematic re-creation of the original context of instruction, then simulations give an even more schematic representation of the same process, until, finally, all you are left with is the input and output to the original process: a perception that C obtains and a representation that A is to be done. You have been re-programmed by the rule, and you did much of the re-programming yourself.

III | THE PROBLEM OF ADOPTION

Rule consequentialists believe that the evaluation of actions is determined by their conformity with authoritative rules. What makes the authoritative rules authoritative is that their adoption has at least as good of consequences as any alternative set of rules. Rule Consequentialists must then give us an account of adoption.

[26] See Baars (1988); Shiffrin (1988).

[27] Perceptual representations are not by definition conscious, not are simulations run by partially reactivating them. They often will involve conscious experiences, however.

[28] The downstream cognitive processing still involves perception, but it is not identical to it.

[29] In the event it helps, think of Louis Armstrong later in his career.

The obvious way to go is to identify adoption of a rule with conformity to it. This approach faces several difficulties. The first is that it might be extensionally equivalent to Act Consequentialism.[30] Many Rule Consequentialists would like to avoid this consequence since one of the selling points of Rule Consequentialism is that it seems to have the resources available to avoid some of the counterintuitive implications of Act Consequentialism.[31]

Even if this is not the case, there are nonetheless costs associated with being motivated by rules that are not costs of complying with them.[32] If moral rules are supposed to be action-guiding, then we need to assess the rules not just as evaluative criteria but as fixtures of our psychologies that themselves have consequences.[33] Some consequentialists would like to drive a wedge between the rules that guide action and the rules that serve as criteria of evaluation (e.g., Eggleston, 2013; 2014). Rule consequentialists are not among them, and I will take their position for granted in what follows. My purpose here is not to argue for Rule Consequentialism, but rather to argue that the account of rule internalization I recommend has implications for which kind of rule consequentialist one should be if one is to be a rule consequentialist at all.

At any rate, internalizing rules has consequences beyond complying with them. Those need to be factored into the evaluation of the rules themselves. So, we might think that the best formulation of Rule Consequentialism is one according to which adopting a rule is internalizing it.[34]

There has been a recent surge of interest, however, in the possibility that we should identify adopting a rule with a commitment to teaching it rather than with having internalized it. One argument for this is that teaching a rule is causally upstream from its internalization. Furthermore, the effect underdetermines the cause. That is, the state of internalization could have been brought about by different teaching methods, each with different cost-benefit profiles. So, we can't work back from the effect to the cause. We should instead begin our evaluation of rules as far upstream as possible (T. Miller, 2021).[35]

These disputes are too complicated to fully retrace the dialectic here. Instead, I propose we move the discussion along by considering the nature of internalization in more detail. Several parties to the debate so far have proposed very precise accounts of how the consequences of adopting a rule are to be calculated, but they have been much less precise about what teaching and internalization themselves are. Both parties contrast the state of internalization with the process of teaching that (in the good case) brings about the state. Then they argue about where the focus of evaluation should be. However, I have argued that internalization is itself a process, it is a continuation of teaching except the burden of the teacher has been relocated to the student, who now toggles between two roles. The endpoint of the process is a state in which teaching is complete and perceptual classifications are molded by the rule internalized. At this point there is no further teaching to be done. However, as stated above, rules that are hard to apply are not fully internalized. It takes effort to apply them, and that effort involves simulated self-instruction (think of me reading the treble clef). It quite often takes effort to apply moral rules, even for mature adults. Most of us, I conjecture, are enkratic rather than virtuous, at least

[30] Smart (1961, p. 10–12); Brandt (1965).

[31] See for example Harsanyi (1982); Hooker (2000; b); Smith (2010).

[32] See Lyons (1965: 137–9) for an argument that Rule Consequentialism does not collapse into Act Consequentialism.

[33] See Brandt (1979, p. 271–77); Hooker (2000, p. 91); D. Miller (2014); Parfit (1984, p. 26–7); Ridge (2006, p. 243).

[34] See for example Brandt (1979); Hooker (2000); D. Miller (2014) and Parfit (2011). There is some debate as to whether Mill (1861/1998) did as well.

[35] See D. Miller (2021) for a separate argument centered around the problem of partial acceptance.

in many situations (think of the epigraph). The burdens of instruction are not fully discharged; the balance has been transferred from our childhood teachers to us.

Once we come to understand what internalization really is, we see it really isn't as separate from teaching as causal observation might make it seem. If internalization is just the later stages of teaching in which the burden of instruction has been re-located, then it is unclear why those stages specifically are the ones relevant to the formulation of Rule Consequentialism. After all, it is not merely that the later stages are caused by the earlier stages: they are simulations (or schematic perceptual renderings) of them. So, we don't even understand the content of the later stages independently of what is going on in the earlier stages.

This is not a knock-down argument, of course. Rule consequentialists who focus the evaluation of rules on their internalization still incorporate teaching costs into their evaluation. Maybe reasons can be given for thinking that this is enough. I have at least shifted the burden of proof. If we can't even (notionally) make sense of internalization apart from the way in which it is a continuation of the teaching process, then it is unclear to me what principled reason there could be for separating the later stages from the earlier ones for the statement of Rule Consequentialism. Looking at the process by which we re-program our psychological states to accord with rules, it appears to be a mistake to focus primarily on the rarely achieved state of a completely re-programmed computer. Adult humans are partially re-programmed by the rules we've been taught, trying to use the completed fragment of the program (mixing metaphors) to fix the ship while it is at sea. The latter stages are not merely dominos that fall after the earlier ones, but an attempt to schematically recreate them. So, focusing primarily on the completed state of internalization is not only to focus on an arbitrary segment of the process, but to misunderstand that segment by trying to understand it independently of its relation to the previous stages.

Before concluding this section, I will briefly mention another way one might understand the upshot of this section. I have argued that an improved understanding of internalization gives us reason to think it is not sufficiently separate from teaching for adoption to be understood as what we internalize rather than what we teach. The conclusion I have drawn from this is that we should understand adoption in terms of teaching and internalization as a component of the teaching process. I have in effect argued that the correct account of internalization requires us to expand the scope of teaching to include some things that transpire *in foro interno*. This strikes me as the best way to interpret the account. However, you could perhaps read the lesson in the other direction. Maybe this account comes to the rescue of internalization by expanding it to include teaching. A further possibility is that the distinction between teaching and internalization isn't quite what anyone expected, and the result is embarrassing for everyone. The lesson might be that the debate between proponents of internalization and proponents of teaching is less important for the future of Rule Consequentialism than is sometimes thought. Maybe we have a single process which could be referred to by emphasizing either its earlier or later stages. We need to understand adoption in terms of that process, and it doesn't really matter what we call it so long as we are clear what we mean.

IV | THE MORAL SPIRAL

One of the ideas animating Rule Consequentialism is that moral philosophy ought to concern itself with rules that can guide our action and enable us to coordinate mutual expectations.[36] The

[36] Cf. Hooker (2000); D. Miller (2021).

authoritative rules are the ones with the best consequences. The scope of that claim is controversial. Are they authoritative for everyone existing at any time? Are they relativized to generations of people or by geographical factors? Looked at one way, the rules are constantly changing. Looked at another way, the rules stay the same and we change our understanding of them. I won't try to settle this dispute here. Regardless of which view proves correct, rule consequentialists are concerned with the possibility of moral progress (Cf. Skorupski, 1989).[37] Depending on how we settle the above questions, that might consist in different rules becoming authoritative at different times. On the other hand, it might consist in gradual improvements in the rules people internalize so that over time they better approximate eternally authoritative rules. The important point is that Rule Consequentialism offers a framework within which we can reflectively evaluate the rules we have internalized with an eye toward improving them. Incremental improvements can accrue over generations. Rule Consequentialism provides the framework within which we can evaluate changes to determine if they amount to progress.

The account of rule internalization given in this paper provides the psychological underpinnings that facilitate incremental intergenerational improvements. In recent decades there has been a great deal of empirical work done on cumulative cultural evolution.[38] Several species have regional variations in behavior. Macaques in some places wash potatoes differently than Macaques in other places, for example.[39] However, Macaques don't inherit a goal-directed practice, improve upon it, and then pass that practice down to the next generation for them to employ and eventually improve upon. There is no species other than humans that does this.

We can do this because we are unusually good at inferring the intentions of conspecifics. Even without explicit instruction, human children are typically quite adept at figuring out what adults are trying to do. Due to our unusually long childhood and adolescence, we have ample opportunity to experiment with tasks and roles we see others pursuing.[40] Crucial to our development is that we gradually come to occupy these roles ourselves by imaginatively undertaking these roles and performing these tasks ourselves (Sterelny 2012, Chapter 2). The period of experimentation involves figuring out which parts of the behavior we witness are done for the sake of which ends. In imaginatively projecting ourselves into the roles we see around us we not only adopt the ends characteristic of those roles (doctor, lawyer, mother, athlete, etc.) but we also acquire a sense of how one goes about achieving them. By inheriting not only the ends but the means to those ends, we spare ourselves the need to re-invent the wheel. We start out wherever our ancestors ended up, regarding the pursuits we share. This is what makes iterated improvement possible.[41] Generation n inherits means and ends from generation n-1 as part of their maturation. As a result, they have the rest of their lives to improve upon the means to those ends. They might recognize, for example, that a carriage could better achieve the aim of convenient long-distance transportation if the horses were replaced with a steam engine. Generation n then develops those means and passes them along to generation n+1, who then in turn improves upon them. N+1 might recognize, for instance, that the aim of convenient long-distance transportation could be better

[37] This is especially clear in its early 19th century variants, e.g., Bentham (1996) and Mill (1967, p. 740), though see also D. Miller (2021).

[38] See for example Boyd & Richerson (1996); Danchin & Luc-Alain (2004); Laland & Hoppitt (2003); Reader & Laland (2002); Sterelny (2012); Tomasello (1999).

[39] Avital & Jablonka (2000).

[40] Cf. Sterelny (2012, p. 32ff.).

[41] Tomasello (1999) calls this the "ratchet effect".

served by internal combustion engines than steam engines. In this way, we "accumulate cognitive capital" (Sterelny 2012, Chapter 2) over generations.

Most children can figure out that the purpose of automobiles is efficient transportation without anyone explicitly telling them that. The rationale behind rules of conduct tends to figure more prominently in explicit instruction. We don't simply see people living by the rules, we hear them justify themselves and criticize one another by invoking them. Children are themselves told which rules they have violated and, with varying degrees of articulacy, why those rules matter. Explicit instruction then plays a more prominent role in the inheritance of moral rules than many other cultural practices.

Explicit reflection also plays a more significant role in moral progress than it does in many other culturally inherited domains. Application of the rules involves the simulation of (more or less) explicit instruction. As was shown earlier, we can step back and reconsider prospective actions by simulating critical assessment of them from the perspective of a moral instructor acting as emissary of the group.[42] This process itself admits of iterated application of the sort common to incremental intergenerational progress more generally. That is, we can simulate criticism of our application of the rules we have inherited. Part of our instruction with the rules includes an account of their purpose. When we lie as children, we are told why it is important that we refrain from lying. This means we can imagine criticism of our application of the rule in cases where it doesn't fulfill its purpose. Furthermore, we can imagine criticism of the entire practice of applying the rule if it can be shown to not achieve its purpose. This will involve an appeal to a more encompassing rule, perhaps one stating a general principle of instrumental rationality.[43]

The important point here is that applying a rule is itself an act and consequently the sort of thing of which we can simulate critical assessment. Critical assessment involves applying further rules which itself makes possible yet another (higher order) act of critical assessment, and so on. This has the iterative structure necessary for intergenerational incremental improvement (recall that perceptual symbol systems support recursion). Within a single generation, iterative self-criticism is necessary to put the rules that generation has inherited in jeopardy in order to improve upon them (perhaps iteratively). Since the improvements are transmitted to the next generation, the process proceeds across generations in much the same way it does within them (cf. Mill 1967, p. 740).[44]

None of this should be taken to imply that all change is progress or that things are always getting better. They are not. I am trying to explain how iterative improvements to the rules are possible. I want to develop a psychologically realistic theory of how rules are internalized that vindicates the ambitions of Rule Consequentialism. Rule Consequentialism has long been seen by its proponents as a framework for proposing and evaluating experiments in living. Rule consequentialists should be precise about what they mean when they talk about internalizing rules. I hope to have shown that there is something empirically and philosophically defensible one might mean by it. If that is what one means, then there is a deeper explanation available of why the usual grab-bag of reactive attitudes are involved, why it makes sense to understand adoption in terms of teaching (subject to the caveats discussed above), and how the reformist ambitions of consequentialists such as Bentham and Mill make sense in light of our cognitive architecture.

[42] Cf Tomasello (2014, Chapter 3).

[43] This rule may not have ever been taught explicitly but is surely implicit in much of our explicit instruction, much the way that modus ponens is implicit in domain specific inference rules we are explicitly taught.

[44] See Tomasello (1999) for a detailed account.

V | CONCLUSION

The primary task of this paper was to explain what it is to internalize a rule and how that bears on a variety of topics of interest to rule consequentialists. These include the relevance of the reactive attitudes to internalization, the question of what it is to adopt a rule, and the psychological underpinnings of moral progress. In closing I should mention that the account developed here is not only of interest to rule consequentialists. Act consequentialists, for example, still invoke the same rules as their rule consequentialist counterparts. The difference is that they understand these rules as decision procedures rather than criteria for moral evaluation (cf. Eggleston, 2014). The account is also of interest to non-consequentialists of certain kinds. John Rawls, for instance, says that institutions are systems of rules (1971/1999, p. 49). Furthermore, for an institution to embody a rule is for its participants to have the common knowledge and mutual expectations that they would have if that rule were to be the result of explicit agreement. More work would need to be done to extend my account to cover the common knowledge and mutual expectations of interest to Rawls, but I am optimistic it can be done.

ORCID

Spencer Paulson https://orcid.org/0000-0001-5794-1387

REFERENCES

Anderson, M. (2010). Neural reuse: a fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33, 245–66. https://doi.org/10.1017/S0140525X10000853.

Ashley, J. & Tomasello, M. (1998). Cooperative problem-solving and teaching in pre-schoolers. *Social Development*, 17, 143–63. https://doi.org/10.1111/1467-9507.00059.

Avital, E. & Jablonka, E. (2000). *Animal Traditions: Behavioral Inheritance in Evolution*. Cambridge: Cambridge University Press.

Baars, B. (1988). *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Bakeman, R. & Adamson, L. (1984). Coordinating attention to people and objects in mother-infant and peer-infant interactions. *Child Development*, 55, 1278–89. https://doi.org/10.2307/1129997

Bakhtin, M. (1981). *The Dialogic Imagination*. Austin: University of Texas Press.

Baron-Cohen, S. (1991). Precursors to a theory of mind: understanding attention in others. In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development, and Simulation of Everyday Mindreading*, (pp. 233–51). Oxford: Basil Blackwell.

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 588–609. https://doi.org/10.1017/s0140525x99002149.

Barsalou, L.W. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes*, 18, 513–62. https://doi.org/10.1080/01690960344000026.

Barsalou, L.W. (2013). Mirroring as pattern completion inferences within situated conceptualizations. *Cortex* 49, 2951–3. https://doi.org/10.1016/j.cortex.2013.06.010.

Barsalou, L.W. & Solomon, K. & Wu, L. (1999). "Perceptual Simulation in Conceptual Tasks". *Cultural, Psychological, and Typological Issues in Cognitive Linguistics: Selected Papers of the Biannual ICLA Meeting in Albuquerque July 1995*, 209. https://doi.org/10.1075/cilt.152.15bar.

Bates, E. (1979). *The Emergence of Symbols: Cognition and Communication in Infancy*. New York: Academic Press.

Bentham, J. (1996). *An Introduction to the Principles of Morals and Legislation*. (eds.) J.H. Burns & H.L.A Hart. Oxford: Clarendon Press.

Blackburn, S. (1998). *Ruling Passions*. Clarendon Press: Oxford, U.K.

Boyd, R., & Richerson, P. J. (1996). Why culture is common, but cultural evolution is rare. In W. G. Runciman, J. M. Smith, & R. I. M. Dunbar (Eds.), Evolution of social behaviour patterns in primates and man (pp. 77–93). Oxford: Oxford University Press

Brandt, R. (1965). Toward a credible form of utilitarianism. In H.N. Castaneda & G. Nakkhnikian (Eds.), *Morality and Language of Conduct* (pp. 107–43). Detroit: Wayne State University Press.
Brandt, R. (1979). *A Theory of the Good and the Right*. Oxford: Clarendon Press.
Brooks, R. (1991). Intelligence without reason. *Proceedings of the 1991 International Joint Conference on Artificial Intelligence*, 47, 569–95.
Burge, T. (2010). *Origins of Objectivity*. Oxford: Oxford University Press.
Butterworth, G. (1991). The ontogeny and phylogeny of joint attention. In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development, and Simulation of Everyday Mindreading* (pp. 223–32). Oxford: Basil Blackwell.
Carey, S. (2009). *The Origin of Concepts*. Oxford: Oxford University Press.
Carruthers, P. (2006). *The Architecture of the Mind*. Oxford: Oxford University Press.
Clark, A. (1989). *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*. Cambridge: MIT Press.
Corkum, V. & Moore, C. (1995). Development of joint visual attention in infants. In C. Moore. & P.J. Dunham (Eds.), *Joint Attention: Its Origins and Role in Development* (pp.61–83). Hillsdale: Erlbaum.
Damasio, A. (1989). Time-locked multi-regional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. *Cognition*, 33, 25–62. https://doi.org/10.1016/0010-0277(89)90005-x.
Danchin, E. & Luc-Alain, G. (2004). Public information: from nosy neighbors to cultural evolution. *Science*, 305, 487–91. https://doi.org/10.1126/science.1098254.
Harris, P. (1991). The work of the imagination. In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development, and Simulation of Everyday Mindreading* (pp. 283–304). Basil Blackwell: Oxford U.K.
Copp, D. (1995). *Morality, Normativity, and Society*. Oxford: Oxford University Press.
Darwall, S. (2006). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge: Harvard University Press.
Dretske, F. (1981). *Knowledge and the Flow of Information*. Palo Alto: CSLI.
Eggleston, B. (2013). Rejecting the publicity condition: the inevitability of esoteric morality. *Philosophical Quarterly*, 63, 29–57. https://doi.org/10.1111/j.1467-9213.2012.00106.x
Eggleston, B. (2014). Act consequentialism. In B. Eggleston & D. Miller (Eds.), *The Cambridge Companion to Utilitarianism* (pp. 125–46). Cambridge University Press: Cambridge, U.K.
Farah, M.J. & McClelland, J.L. (1991). A computational model of semantic memory impairment: modality specificity and emergent category specificity. *Journal of Experimental Psychology*, 120, 339–57. https://doi.org/10.1037/0096-3445.120.4.339
Fernyhough, C. (1996). The dialogic mind: a dialogic approach to the higher-mental functions. *New Ideas in Psychology*, 14, 47–62. https://doi.org/10.1016/0732-118X(95)00024-B.
Fodor, J. (1975). *The Language of Thought*. Cambridge: Harvard University Press.
Foley, M. & Ratner, H. (1997). Children's encoding in memory for collaboration: a way of learning from others. *Cognitive Development*, 13, 91–108. https://doi.org/10.1016/S0885-2014(98)90022-3.
Gert, B. (1998). *Morality*. Oxford: Oxford University Press.
Goldman, A. (2006). *Simulating Minds*. Oxford: Oxford University Press.
Goodman, S. (1984). The integration of verbal and motor behavior in preschool children. *Child Development*, 52, 280–9.
Hamann, K., Warneken, F., & Tomasello, M. (2012). Children's developing commitment to joint goals. *Child Development*, 83, 137–45. https://doi.org/10.1111/j.1467-8624.2011.01695.x
Hardy, S.A., & Carlo, G. (2005). Identity as a source of moral motivation. *Human Development*, 48, 232–56. https://doi.org/10.1159/000086859.
Harsanyi, J. (1982). Some epistemological advantages of a utilitarian system in ethics. In P.A. French, T.E. Uehling., & H.K. Wettstein (Eds.) *Midwest Studies in Philosophy Volume III* (pp. 389–402). Minneapolis: University of Minnesota Press.
Hobson, P. (2004). *The Cradle of Thought: Exploring the Origins of Thinking*. London: Pan Books.
Hooker, B. (2000). *Ideal Code, Real World*. Oxford: Oxford University Press.
Jenkin, Z. (2020). Perceptual learning and reasons-responsiveness. *Noûs*, 57:481–508. https://doi.org/10.1111/nous.12425.
Johnson, M.H., Dzuriawec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of facelike stimuli and its subsequent decline. *Cognition*, 40, 1–19. https://doi.org/10.1016/0010-0277(91)90045-6.

Johnson, M.H. & Morton, J. (1991). *Biology & Cognitive Development: The Case of Face Recognition*. Hoboken: Blackwell.
Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective*. Cambridge: MIT Press.
Kontos (1983). Adult-child interaction and the origins of metacognition. *Journal of Educational Research*, 77, 43–54. https://doi.org/10.1080/00220671.1983.10885494.
Kosslyn, S. (1994). *Image and Brain*. Cambridge: MIT Press.
Langacker, R.W. (1986). An introduction to cognitive grammar. *Cognitive Science*, 10, 10–40. https://doi.org/10.1207/s15516709cog1001_1.
Laland, K. & Hoppitt, W. (2003). Do animals have culture? *Evolutionary Anthropology*, 12, 150–9. https://doi.org/10.1002/evan.10111.
Luria, A. (1961). *The Role of Speech in the Regulation of Normal and Abnormal Behavior*. New York: Boni & Liveright.
Lyons, D. (1965). *Forms and Limits of Utilitarianism*. Oxford: Clarendon Press.
Mackie, J.L. (1977). *Inventing Right and Wrong*. Hammondsworth: Penguin.
Man, K., Kaplan, J. Damasio, H., & Damasio, A. (2013). Neural convergence and divergence in the mammalian cerebral cortex: from experimental neuroanatomy to functional neuroimaging. *Journal of Computational Neurology*, 521, 4097–4111. https://doi.org/10.1002/cne.23408.
Manohar, S.G., Zokaei, N., Fallon, S.J., Vogels, T.P., Husain, M. (2019). Neural mechanisms of attending to items in working memory. *Neuroscience Behavioral. Review*, 101, 1–12. https://doi.org/10.1016/j.neubiorev.2019.03.017.
Massey, C., & Gelman, R. (1988). Preschoolers' ability to decide whether pictured or unfamiliar objects can move themselves. *Developmental Psychology*, 24, 307–17. https://doi.org/10.1037/0012-1649.24.3.307
McGeer, V. (2014). Civilizing blame. In D.J. Coates & N. Tognazzini (Eds.) *Blame: Its Nature and Norms* (pp. 162–89). Oxford: Oxford University Press
McRae, K. & Cree, G.S. (2002). Factors underlying category-specific semantic deficits. In E. Forde & G. Humphries (Eds.), *Category-Specificity in Mind and Brain* (pp. 211–49). East Sussex: Psychology Press.
Meltzoff, A. (1995). Understanding the intentions of others: re-enactment of intended Acts by 18-month-old children. *Developmental Psychology*, 31, 838–50. https://doi.org/10.1037/0012-1649.31.5.838.
Meltzoff, A. & Moore, K. (1977). Imitation of facial and manual gestures by newborn infants. *Science*. 198, 75–8. https://doi.org/10.1126/science.897687
Mill, J.S. (1861/1998). *Utilitarianism*. (ed.) Roger Crisp. Oxford: Oxford University Press.
Mill, J.S. (1967). Chapters on socialism. In J. Robinson (Ed.) Collected Works of John Stuart Mill, *Volume V* (pp. 703–53). Toronto: University of Toronto Press.
Miller, D. (2014). Rule utilitarianism. In B. Eggleston & D. Millers (Eds.), *The Cambridge Companion to Utilitarianism* (pp. 146–66). Cambridge: Cambridge University Press.
Miller, D. (2021). Moral education and rule consequentialism. *Philosophical Quarterly*, 71(1), 120–40. https://doi.org/10.1093/pq/pqaa023.
Miller, T. (2021). From compliance, to acceptance, to teaching: relocating rule consequentialisms stipulations. *Utilitas*, 33, 204–220. https://doi.org/10.1017/S0953820820000369.
Newell, A. & Simon, H.A. (1972). *Human Problem Solving*. Hoboken: Prentice Hall
Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.
Parfit, D. (2011). *On What Matters*. Oxford: Oxford University Press.
Premack, D. (1990). Words: what are they, and do animals have them? *Cognition*, 37, 197–212. https://doi.org/10.1016/0010-0277(90)90045-L.
Pylyshyn, Z. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge: MIT Press.
Ratner, H. & Hill, L. (1991). "Regulation and Representation in the Development of Children's Memory". Paper Presented to the Society for Research in Child Development, Seattle.
Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences*, 22, 253–336. https://doi.org/10.1017/S0140525X9900182X
Rawls, J. (1971/1999). *A Theory of Justice*. Cambridge: Belknap Press.
Reader, S. & Laland, K. (2002). Social intelligence, innovation, and enhanced brain size in primates. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 4436–4441. https://doi.org/10.1073/pnas.062041299
Ridge, M. (2006). Introducing variable-rate utilitarianism. *Philosophical Quarterly*, 56, 242–53. https://doi.org/10.1111/j.1467-9213.2006.00440.x.

Rosch, E. & Mervis, Carolyn, B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology*, 7, 573–705. https://doi.org/10.1016/0010-0285(75)90024-9.
Schulz, A. (2022). *Efficient Cognition: The Evolution of Representational Decision Making*. Cambridge: MIT Press.
Shiffrin, R. M. (1988). Attention. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey & R. D. Luce (Eds.) *Stevens' handbook of experimental psychology: Vol. 2. Learning and cognition* (pp.740–805). Hoboken: Wiley.
Siegel, S. (2012). Cognitive penetrability & perceptual justification. *Noûs*, 46, 1–22.
Skorupski, J. (1989). *John Stuart Mill*. New York: Routledge.
Smart, J.J.C. (1961). *An Outline of a System of Utilitarian Ethics*. Melbourne: Melbourne University Press.
Smith, A. (2014). Moral blame and moral protest. In D.J. Coates & N. Tognazzini (Eds.), *Blame: Its Nature and Norms* (pp. 27–49). Oxford: Oxford University Press.
Smith, H. (2010). Measuring the consequences of rules. *Utilitas*, 22, 413–33. https://doi.org/10.1017/S0953820810000324.
Sterelny, K. (2012). *The Evolved Apprentice*. Cambridge: MIT Press.
Tobia, K. (2018). Rule consequentialism's assumptions. *Utilitas*, 33, 458–71. https://doi.org/10.1017/S0953820818000031.
Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press: Cambridge MA.
Tomasello, M. (2003). *Constructing a Language: A Usage Based Theory of Language Acquisition*. Cambridge: Harvard University Press.
Tomasello, M. (2014). *A Natural History of Human Thinking*. Cambridge: Harvard University Press.
Tomasello, M. (2021). *Becoming Human*. Cambridge: Belknap Press.
Tomasello, M. & Call, J. (1997). *Primate Cognition*. Oxford: Oxford University Press.
Tomasello, M. Kruger, A., & Ratner, H. (1993). Cultural Learning. *Behavioral and Brain Sciences*, 16, 495–552. https://doi.org/10.1017/S0140525X0003123X.
Tooby, J., & Cosmides, L. (1992). "Introduction: Evolutionary Psychology and Conceptual Integration". In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (pp. 19–137). Oxford: Oxford University Press.
Vygotsky, L. (1978). *Mind and Society: The Development of Higher Psychological Processes* (ed. Cole, M.). Harvard University Press: Cambridge MA.
Wallace, R.J. (1994). *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.
Wallace, R.J. (2011). Dispassionate opprobrium: on blame and the reactive sentiments. In R.J. Wallace, R. Kumar., & S. Freeman (Eds.), *Reasons and Recognition: Essays in Honor of T.M. Scanlon*. Oxford: Oxford University Press.
Walsh, E. & Oakley, D. (2022). Editing reality in the brain. *Neuroscience of Consciousness*, 2022, 1–12. https://doi.org/10.1093/nc/niac009
Wertsch, J. (1991). *Voices of the Mind: A Sociocultural Approach to Mediated Action*. Cambridge: Harvard University Press.

How to cite this article: Paulson, S. (2024) Internalizing rules. *Philosophy and Phenomenological Research*, 1–20. https://doi.org/10.1111/phpr.13065