Draft: Please do not quote without permission of the author; comments welcome at drkelly@purdue.edu

Forthcoming in *The Oxford Handbook of Moral Psychology*, ed. by Manuel Vargas & John Doris

Last Changes: 12/19/19 DRK
Word Count: main text 10,031

Two Ways to Adopt a Norm:

The (Moral?) Psychology of Internalization and Avowal

By Daniel Kelly[1]

Section I: Becoming Alice

Consider Alice. She is in her mid 20's, and WEIRD, i.e. lives in a modern culture that is predominantly western, educated, industrialized, rich, and democratic. She has made it through many of the stages of adolescence and young adulthood, figuring out who she is and taking steps towards becoming who she wants to be. She is, of course, still a work in progress (aren't we all). Nevertheless, by this point in her life, some of the guidelines she lives by, and even some of the more central elements of the identity she is constructing to help herself steer through the world, are *chosen*; they are self-selected and self-imposed. These sorts of voluntarily adopted rules and values can concern all manner of domains and behaviors, and what they have in common is neither scope nor subject matter, but rather that at some point Alice herself *decided* to adopt them; she explicitly formulated, consciously entertained, carefully deliberated over, and embraced them. She has also publicly endorsed some of them, maybe using social media to help along the indirect process of incorporating them more deeply into her habits, public personae, and self-conception.

For example, at various times Alice considered the pro's and con's of cutting meat out of her diet, of giving a larger percentage of her paycheck to Planned Parenthood, and of trading daily runs for daily yoga sessions. Once she, say, elected to go vegetarian, she adopted the rule *don't eat meat*, and committed herself to following it, allowing it to curtail her culinary options henceforth. Tempted by a juicy bratwurst at a barbeque, she might steel her resolve, calling upon her inner resources by silently exhorting herself: "don't do it; you're a vegetarian now!" She might publicly tell others "I stopped eating meat" in response to questions about potential menu items for an upcoming dinner party. She can assert her newly embraced guideline in conversations with friends to realign their expectations of her, and to enlist their help in keeping her on the straight and narrow. Alice may also begin to prescribe the rule to others, explicitly adding or otherwise indicating that while this is a precept she has personally adopted, she also believes that no one else should eat meat either.

[1] Many thanks to those who read and commented on earlier versions of this paper. They include Lacey Davidson, Taylor Davis, John Doris, Ying-yi Hong, Jenann Ismael, Elijah Millgram, Elizabeth O'Neill, Stephen Setman, Stephen Stich, and Manuel Vargas.

If, in the midst of all this affirmation, Alice was also covertly enjoying a tasty cheeseburger on a regular basis, or was engaged in some other pattern of behavior inconsistent with the rule, it would provoke the reasonable worry that she was merely paying it lip service. This in turn could awaken suspicions of some deeper flaw: inauthentic commitment to vegetarianism, lack of integrity or weakness of character in general, tortured self-deception, or even calculated hypocrisy—burnishing a breastplate of dietary righteousness, virtue signaling without the actual virtue. But an occasional fall off the wagon would be forgivable, since it would not by itself indicate anything more momentous than, say, an isolated misstep or temporary lapse of will. Alice strives to live up to her personal ideals, but intermittently falls short of meeting them, especially at first. It certainly wouldn't be taken to reveal that she is a Machiavellian schemer spouting cheap talk about vegetarian ideals in bad faith, or that the contents of her own mind are systematically opaque to her. When she surprises her family with the strident announcement that "I believe no one should eat factory farmed meat", her claim to know what she believes remains authoritative.[2]

Avowed norms like Alice's *don't eat meat* are, of course, not the only kinds of rules that guide her behavior. Other, non-avowed rules can exert influence over her even though they haven't been personally vetted, and so never enjoyed the same careful attention and explicit endorsement. Like their avowed counterparts, such rules apply to a variety of behaviors, governing things like how much of her income she turns over to the government for taxes, what side of the sidewalk she walks on, which utensils she uses to eat soup and salad, how close she stands to someone she is talking to, what clothes she picks out to wear to a professional meeting, how seriously she takes advice or testimony from different people, and how and when she allows herself to express emotions like anger and grief. This is a motley mix of rules to be sure, and one dimension on which the rules differ is what drives Alice to act in accord with them. She might comply with federal laws out of a conscious self-interested desire to avoid formal reprimands like fines or jail time, even if she thinks those laws are unfair. She might habitually use a pragmatically effective rule of thumb that she picked up from a friend. Her sensitivity to peer pressure might be mainly what motivates her follow with her community's customs about proper dining etiquette. She may also unconsciously ensure her own behavior satisfies its unwritten standards governing appropriate ways to allocate credibility and display emotion, without even noticing she is doing so. Alice may not always realize that she is sensitive to norms of this last sort, but even when she becomes aware she sometimes just continues to comply with them, going along to get along. Sometimes not, though; I'll return to this below.

Alice's behavior reliably conforms to all of these kinds of rules, albeit for different reasons across the different cases. Like those she has avowed, rules in this contrasting set are not united by a shared subject matter. Nor, however, are they united by occupying a similar functional role in Alice's public and mental life; while they all affect her behavior, they do not all do so, socially or psychologically, in the same way. All that they have in common is that each rule influences Alice's behavior even in the *absence* of her explicit endorsement of it, *despite* the fact that she did not consciously consent to be bound by it. Rules in this category exert normative force on Alice even though she never personally avowed them.

[2] Though even this can be controversial; see e.g. Doris 2015 and Haybron this volume.

From this motley bunch I will separate out a subcategory for special attention. In what follows I will call them, for reasons that will become clearer as we go, *internalized norms*. These *do* occupy a specific functional role in Alice's public and mental life. They are socially acquired behavioral rules stabilized by communal practices of intrinsically motivated compliance and enforcement. I will unpack this as we go, and make the case that internalized norms constitute a class of rules that is distinctive and important from the point of view not just of moral psychology, but of the behavioral sciences more generally. Internalized norms are acquired from social interactions in characteristic ways by the dedicated psychological machinery that handles them. Once internalized, they shape cognition and attention, motivate behavior, and may be susceptible and resistant to intervention in distinctive ways as well. The main contrast class to internalized norms for this paper will be what I've been calling *avowed norms*. These, too, constitute a class of rules that is important not just for philosophy and moral theory, but from the point of view of the behavioral sciences more generally. The distinction between the two has been less appreciated than is ideal, however, and our understanding of the psychological underpinnings of avowal remains even more in its infancy than our understanding of internalized norms.

Thus, two main aims of this paper are to clarify the distinction and to characterize key features of each type of norm in a way that might usefully guide future research. In Section II I will identify and describe a number of different lines of research that address human norm-governed behavior. I will compare and contrast how they conceive of their subject matter, and show how the distinction between avowed and internalized norms that I am proposing cross cuts the categories that have organized much of this research. In Section III I turn my focus to cognitive architecture. I describe in broad outline an account of the human capacity for self-regulation provided by McGeer and Pettit (2002), and show how this picture fits with the kinds of dual system architectures now common in the cognitive sciences. In Section IV I use this picture to develop my accounts of avowed and internalized norms, arguing that avowed norms draw on the slower, more deliberate cognitive machinery of self-regulation, while internalized norms are underpinned by a specialized psychological system that handles information and generates motivation in a way that bears many of the characteristics associated with system 1 'fast thinking'. In this section, as in the one that precedes it, I highlight the different motivational features associated with each kind of norm, and attempt to clarify what we know and to formulate some questions that focus attention on what remains unknown. Finally, in Section V I conclude by drawing the strands of the previous sections together and pointing to several issues in the philosophical literature that stand to be illuminated by a better developed and empirically grounded account of the distinctive psychological profiles of internalized and avowed norms.

Section II: An Embarrassment of Riches: A Partial Geography of Categories of Norms

It will first help to situate this distinction with respect to recent work on norms, in no small part because there seem to be so many nearby distinctions on offer (see O'Neill 2017 for a recent survey and endorsement of a reasonable pluralism). Common sense and the vernacular contain an array of intuitive ways to categorize norms, often marking differences between rules based mainly on the kind of activity they regulate. These include sartorial norms concerning how to dress; dining norms concerning how to prepare and consume food; conversational norms regulating the dynamics of dialogue; privacy norms

that manage a whole host of issues, including personal space; and organizational norms that confer powers and duties on actors in different institutional positions. Empirical researchers, on the other hand, have developed theories that group norms together into categories cast at higher levels of generality, often sorting them by reference not only to specific types of behavior to which they apply but also to a more abstract, core value that informs them, such as the values of autonomy, community, and divinity described by Shweder et al (1997).

A prominent landmark in this conceptual geography is the general question of what demarcates the boundaries of the even more abstract category of morality, and of which norms are distinctively *moral* norms. An early attempt to answer this question was made by Kohlberg, whose theory depicted a developmental trajectory that individuals are alleged to take as they learn to distinguish the genuinely moral principles of justice from merely conventional rules or norms of social consensus. Kohlberg's way of carving off the moral from the larger domain of normativity in general was famously criticized by Gilligan (1982) as excluding, or at least taking insufficient account of, women's perspectives. Gilligan also objected that the account was overly restrictive, failing to countenance a variety of behaviors and norms that were putatively moral but did not involve justice, especially those behaviors and norms associated with what she called the ethics of care.

Challenging Kohlberg from another direction, Turiel and his collaborators (Turiel 1983, Smetana 1993, Nucci 2001) disputed the claim that children initially conceive of all rules in the same way, and only gradually come to appreciate important distinctions between them (conventional, instrumental, genuinely moral, etc.) These researchers gathered a wealth of evidence suggesting that even young children conceive of putatively moral and conventional rules in quite different ways. On the view Turiel developed to account for these studies, distinctively moral rules are those that people conceive of as sharing a number of properties: they are judged to be generally rather than only locally applicable in scope, they are judged to be independent of and unchangeable by any authority figure, and they are judged to involve either justice, harm, welfare, or rights. Conventional rules, on this account, are those judged to have the opposite cluster of features, and in experiments violations of these conventional rules were often judged to be less serious than violations of their counterpart moral norms.

Setting aside for a moment the issue of its truth or falsity, the Turiel-inspired account is noteworthy for the crisp picture it suggests, and the relatively clear answer it implies for the question about the domain of morality: moral norms are marked by the fact that they have a number of key features in common, some of which have to do with their content (involving justice, harm, welfare, or rights), and some of which transcend their content (general scope, authority independence). Moreover, the theory holds that these content and content-transcending features all cluster together in a non-accidental, potentially culturally universal way. Given this picture, there would certainly be a good prima facie case that Turiel and his collaborators had succeeded in identifying a plausible candidate for the extension of the term 'moral norm', and thereby provided good reason to think that 'moral' picks out a scientifically interesting and important category, perhaps a psychological natural kind (see Kumar 2015 for thoughts along these lines).

This clean picture breaks down, alas, but in instructive ways (Kelly et al 2007, Kelly & Stich 2007). Early critics uncovered deviations from the expected experimental results that

had several noteworthy features. First, people from different countries and socioeconomic groups were liable to ascribe 'moral' content-transcending properties to some norms and activities that they acknowledged had little to do with justice, harm, welfare or rights (Haidt et al 1993). Second, participants in the experiments often tended to 'moralize' (in something like the sense associated with Turiel-inspired accounts) norms and activities that activated a strong emotional response (Rozin et al 1999). Three trends coalesced in the wake of this. One was that theorists were beginning to take more seriously the fact of non-trivial cross-cultural cognitive variation in general, and of normative diversity in particular (Nisbett 2003, Doris and Plakias 2007, Henrich et al 2010, Sommers 2012, Flanagan 2016, Stich et al 2018). Another was that increased effort was directed at formulating models of the cognitive machinery underpinning normative judgments. Many of these explored ways in which the research on moral judgment and data on cross-cultural diversity might be compatible with psychological mechanisms that were at least in part innate, domain specific, and affect- and emotion-driven (Nichols 2004, Prinz 2007, Mikhail 2011, Greene 2014).

Finally, the failure of the Turiel approach to deliver a defensible account of moral norms and the boundaries of the moral domain fueled a free-for-all of empirical theorizing attemping to provide a workable alternative. Theorists took different approaches to finding the distinctive mark of the moral. Some embarked on investigations of the relationship between morality and meat eating (Mameli 2013) or morality and judgments of objectivity (Goodwin and Darley (2008, 2010, 2012). Others attempted to discover relevant subdivisions of what they took to be the moral domain (Graham et al , 2011 and 2013), while still others argued that morality is reducible to something else, like cooperation (Curry 2016, c.f. Kitcher 2011) or to some single fundamental subdomain such as harm (Schein and Gray 2017) or fairness (Baumard et al 2013). Some theorists made the case that the concept of morality is used to pick out different sets of norms and activities from one culture or community to the next (Haidt 2012). This flurry of theorizing also provoked speculation that the concept of morality is itself merely a WEIRD invention: a historically recent, culturally parochial, psychologically uninteresting honorific used by different communities to commend whatever their favored subset of normativity happened to be, and by different researchers for whatever purposes were rhetorically convenient. No position on any of these issues currently enjoys consensus support, and indeed many have voiced skepticism about different parts of the project itself (Sinnott-Armstrong and Wheatley 2012, Sterelny 2012, Stich 2018, c.f. Machery 2012, Davis under review).

Developing alongside—but for the most part independently of—this work in self-styled 'moral' psychology have been lines of research concerned with norms and other putatively similar subject matter. Here, however, the initial point of departure is typically not intracranial psychological machinery but rather patterns in the collective activity of groups of people. This approach includes practitioners who are anthropologists, sociologists, social psychologists, game theorists, computer modelers, evolutionary theorists, economists, and philosophers. It has also yielded its own assortment of taxonomies, categorizing different group-level regularities by appeal to a range of features. Psychology figures in the mix here as well, since distinctions are drawn between different kinds of social patterns by appeal to the cognitive and motivational states of the individual people whose behaviors collectively form each kind of regularity. However, the taxonomies of the subject matter that this research has developed are strikingly different than those on offer in the moral psychology literature described above.

Here, for example, distinctions have been drawn between conventions and moral rules, but also taboos, customs, traditions, descriptive norms, dynamic norms, injunctive norms, and social norms. Even when the pieces of terminology are similar across literatures (i.e. 'convention' and 'moral rule'), the categories those terms are used to express are different, in both their intensions and extensions. Of particular note is that here theoretical divisions between different kinds of group-level regularities are often made not by appeal to content or domain of activity (dining, sartorial, personal space), nor to the prominence of a particular emotion in driving the relevant behaviors (guilt, anger, disgust), nor to the core value associated with the practice (autonomy, fairness, justice). Rather, theoretical divisions are drawn by reference to how each kind of collective social pattern is *stabilized*.

Key contributors to this stability are the clusters of psychological states of the individual members of the group. The mental states posited in these stability-producing clusters are typically similar to those of folk psychology, and also typically social or interpersonally directed—they are *about* the psychological states of the other members of the group. So on this picture, different kinds of endogenously stable social patterns (conventions, descriptive norms, social norms) appear in a community when its members have different combinations of i) communally shared expectations about how most others *will* act in some set of relevant circumstances, ii) communally shared beliefs about how people *should* act in those circumstances, iii) shared beliefs about *the communally shared beliefs* about how people will and should act in those circumstances, together with iv) common preferences individuals hold about if and when they themselves would like to act in accordance with those communally shared expectations and beliefs. (See Lewis 1969, Cialdini et al 1991, Ostrom 2000, Bendor and Swistak 2001, Centola et al. 2005, Schultz et al 2007, Southwood 2011, Southwood, N. and Eriksson 2011, Smith et al 2012, Brennan et al 2013, Bicchieri and Muldoon 2014, Morris et al 2015, Young 2015, Bicchieri 2006, 2016, Sparkman and Walton 2017).

This literature is impressively complicated. Like the literature in moral psychology described above, it too offers an array of cross-cutting distinctions made by different researchers. Likewise, many of these are subtle and contested, but are also apt to be important for purposes both theoretic and practical. For an interdisciplinary reader, though, the sum effect of reading within one of these two literatures, let alone in both of them, can be a frustrating sense of confusion. But this does not indicate anything has gone awry, or even by itself that some theorists are right and others wrong. Purposes are many and varied, and so too will be the categories and distinctions that respectively serve them best. Perhaps there are even important insights to be won from exploring the relationships between these two bodies of work (see Davis et al 2018 and Kelly and Davis 2018 for some initial steps in this direction). For now, the "pluralism about classification schemes for norms" endorsed by O'Neill (2017) is a reasonable position to adopt in light of the embarrassment of classificatory riches already at hand. It is also an attractive one, since in the next sections I will argue that interdisciplinary researchers in the human sciences would benefit from adding another distinction to that embarrassment.

Section III Self-Regulating Minds and Their Routinized Components Parts

Take Alice's pronouncement to her friends concerning her recent conversion to a vegetarian lifestyle: "I believe factory farming is wrong, and so I no longer eat meat." This can be construed as an avowal of a norm (e.g., *don't eat meat*) that Alice has chosen to adopt for herself. There are good reasons to think that the logical, semantic, and epistemological properties of such avowals differ from those associated with other instances of self-ascription, cases in which a person merely reports on one of their own mental states: "I'm hungry", "I find myself becoming convinced that capitalism is an inherently inhumane economic system", "I think I might I love you, Beatrice", "I eventually realized that as a child I had absorbed the idea that women belonged in the home."[3]

In this section I continue making the case that the *psychological* profiles of avowed and internalized norms are distinct. I begin developing the kind of hybrid account of psychological architecture needed to help explain each. Luckily, there are a number of general accounts of two-tiered cognitive architecture on the market in psychology, many of which are compatible with the picture of *self-regulating minds* that that I will unpack presently. More importantly, that picture looks amenable to extension, so that it can help to illuminate important functional joints not just of the psychology of belief formation for which it was initially developed, but of normative cognition more generally.

McGeer and Pettit (2002) offer an account of the capacity to *self-regulate*, an ability they take to be distinctive of humans. They characterize this capacity in terms of the human mind's ability to impose constraints on itself, thus shaping its own activity. They develop their picture in stages, starting with the general characteristics of less sophisticated minds that lack this self-regulating capacity, and adding a series of features that together underpin the ability to exert more reflective self-control over what she believes and which actions she might take or avoid. As will become clear, I do not take the McGeer and Pettit account of self-regulation to succeed as a complete explanation of the range of sophisticated human behavior they discuss; nor is it likely that they offer it as one. Rather, I interpret them as giving a plausible sketch of a kind of cognitive platform likely to be a central component of the more detailed explanations of many of those sophisticated behaviors. I also take their account to provide important insights about the framework within which mechanisms responsible for those more specific capacities might be located.

According to that account, simpler minds than ours are those that are *merely* routinized. McGeer and Pettit adopt what they call a "constraint-conforming approach" to

---

[3] This section was inspired by Ismael's discussion (2014, 2016) of the different forms of information processing likely to underpin what she calls the descriptive and performative forms of self-ascription. My jumping off point is one she makes while considering the kind of self-knowledge possible in cases of avowal, where she also notes that

> not all first-personal intentional ascriptions are avowals. To get the right account of self-knowledge, we need a two-tier account along the lines of [McGeer and Pettit 2002], which allows for both descriptive and performative aspects of self-ascription. … There is good motivation for a hybrid account." (Ismael 2014, pg 293)

The distinction I am developing lies within the category of norms, rather than the kinds of cases Ismael is primarily concerned with, but the reasoning that militates for some form of pluralism applies to both.

understanding these merely routinized minds, an approach mostly closely associated with Dennett's (1981) well-known "intentional stance":

> "[to] qualify as 'minded' in some minimal sense, is … to be a system that is well-behaved in representational and related respects … whether an organism or artifice is intentionally minded is fixed by whether it conforms to evidence-related and action-related constraints in a satisfactory measure and manner. … We shall be taking the constraint-conforming approach to mindedness as our starting-point in this paper." (McGeer and Pettit 2002, 282)

In a merely routinized mind, the constraints that govern the flow of information between perceptual input and behavioral output connect the former to the latter in ways that "attain a certain threshold of rational performance" (282). These constraints allow the minded entity to avoid threats and satisfy aims, at least in typical environmental conditions. Such constraints are themselves fairly rigid, but can collectively implement routine behavioral patterns that allow the entity to respond selectively and intelligently to the relevant features of its surroundings—or at least intelligently enough to support the ascription of mindedness and representational content.

The constraints that organize merely routinized minds will typically have an exogenous provenance. They will have been pre-designed and installed by an engineer, in the case of a computer or robot, or will have been shaped over the course of generations of evolution by natural selection, in the case of most non-human organisms. Also characteristic of merely routinized minds is that their constituent constraints are what I'll call *architectural*. Architectural constraints are causally efficacious in channeling the flow information, and may themselves be *vehicles* of intentional content, but are not themselves *represented*, and so are not the subject matter of the mind's own representations. Merely routinized minds are in this sense blind to their own contents and constraints, including to the very constraints that give them their characteristic organization and that constitute them as minds.[4]

On McGeer and Pettit's account, part of what makes human minds special is that they are not *merely* routinized. While they contain routinized subsystems, they include the capacity for self-regulation as well. Moreover, the ability to self-regulate that is distinctive of adult persons operates (when it does) alongside and in concert with these merely routinized subcomponents. Thus, McGeer and Pettit's account appears broadly compatible with views common in cognitive science that subdivide the mind into different strata of psychological mechanisms. These views include modular theories that distinguish between central and more peripheral subsystems, and dual process and dual system theories that distinguish between the broad families of system 1 processes that are fast, intuitive, relatively automated, implicit, and effortless, on the one hand, and system 2 processes that are slow, deliberate, explicit, and guided by effort and attention, on the other. Putting McGeer and Pettit's account together with a view of this sort yields a two-tiered picture of hierarchical

[4] To foreshadow a distinction between representation and motivation that will loom large later in the paper, organisms with merely routinized minds may lack the *capacity* to represent their own architectural constraints, or alternatively they may possess the representational wherewithal but simply lack the *inclination* to use it in this way.

psychological organization that is recognizable in broad outline (see other chapters in handbook?)

An important feature of this picture is that it depicts the lower tier of psychological organization found in human minds as more of a patchwork than a unity. That lower tier is made up of a package of relatively autonomous heuristics and subsystems, a sometimes kludgy collection of adaptive instincts and problem-solving gadgets each with its own primary and auxiliary functions to perform. The operation of each of these is more or less compartmentalized, sectioned off from the others. Each is dedicated to a fairly specific domain and task, so that it has a proprietary set of constraints that regulate the flow of information between the cues and environmental regularities that it searches for signs of in the stream in perceptual input, on the one hand, and the routine set of motivational and behavioral outputs it produces when one of those cues or regularities is detected, on the other. This lower tier of mental organization does not take the form of a single, well integrated, domain general routine, but is rather a patchwork of hubs of locally cohesive structure, a loosely affiliated bundle of subpersonal mechanisms, many of which are given rather than the result of any prior self-regulated activity.[5]

Which leaves self-regulation, and the second of the two tiers of human mental organization. On McGeer and Pettit's constraint-conforming approach, the capacity to self-regulate is underpinned by a suite of abilities that allow humans to do new and different things with constraints. The ability to use natural language looms large, and from it flows sub-capacities for what they call *content-attention*, *constraint-identification*, and *constraint-implementation*.[6] First, it is with language that a person is able to publicly express propositional contents, using words and sentences to broadcast thoughts into the world beyond her own head. Though internal mental states and public sentences are different vehicles, both can be used to express the same kinds of contents; Alice's belief that the rabbit is white has the same content as the English sentence "the rabbit is white". One benefit of the linguistic representational medium, however, is that in speaking or writing, a person is using the medium to publicize certain contents into her surroundings, where her perceptual awareness is naturally trained. In thus externalizing a thought with language, she makes it much easier to draw and focus her *attention* on it; the content itself can become the *object* of her perceptual awareness.

---

[5] For discussion of such dual systems approaches in general see Kahneman 2011, and for an overview of early applications in moral psychology see Cushman et al 2010. Also see Heyes 2018 for a recent defense of the idea that many systems that bear the characteristics of system 1 are nevertheless acquired from culture rather than innately endowed, learned cognitive gadgets rather than inborn cognitive instincts. Dennett (1969) introduced the influential personal/subpersonal distinction to cognitive science. He argued that in explanations of behavior, appeals to the operation of specific subcomponents of a person's mind, rather to the entire person him or her self, still count as legitimately *psychological* explanations; see Drayson 2014 for more recent discussion.

[6] McGeer and Pettit remain silent and presumably neutral, as will I, on the relationship between natural language and other phenomena clearly relevant to their account of self-regulation. These include imagination and reflexivity, mental-time travel and counterfactual reasoning, meta-representation and meta-cognition, self-awareness and self-consciousness, etc. They do not themselves adopt the jargon of dual process theories, or speak explicitly in terms of lower or higher tiers of psychological structure, though some such distinction is implicit in their discussion of routinized and self-regulating minds. Nor do they address the issue of whether the processes that underlie the ability to self-regulate better fit the general characterization associated with system 1 or system 2.

As noted, words and sentences can be used to express the same content that is carried by a person's mental states, including those ensconced in the merely routinized subcomponents of her own mind. She can use language to entertain sentences whose subject matter is something she already found herself believing or desiring, but also contents shared with the architectural constraints that organize the merely routinized parts of her mind. Thus, the contents and constraints of a self-regulating mind can become visible to the mind itself; a person can come to understand herself as a minded entity and a subject of mental states, and can come to know those mental states in a new, reflective way. Once the contents of her mind are brought into view in this way, she might also consider them anew, questioning, assessing, and deliberating upon them, and evaluating them by reference to various standards. Is it true? Do I have enough evidence to believe it? Can I coherently doubt it? Is it something I *want* to be true? If so, is that a desire I should act on right now? If I act on it, what are the best steps to take to fulfill it? Will taking those steps be consistent with other things I think and want? Is consistency something I want to be constrained by?

Second, natural language also provides a medium with which self-regulating minds can discriminate between contents they are attending to, and also to formulate and entertain novel ones. The range of different contents distinguishable with this ability appears theoretically unrestricted, but will be practically limited by the representational richness of the language and the imaginative resources of the person employing it. She can reflect on different contents on her own, allowing her wandering mind to reshuffle bits of memories and daydreams into fresh combinations of contents, or actively directing her creative energies to coming up with new ideas that can serve a particular purpose. She can also publicly discuss ideas, arguing with other people to collaboratively tease out and express new possibilities. She can thus *identify* and distinguish new specific contents of many sorts. Moreover, some of these she will be able to identify as potential *constraints*. They might take the form of imperatives, or any other kinds of rules and standards that might be used to guide and restrict activity in various ways. Does my uncle even care that his religious and political beliefs are wildly inconsistent? What would a reasonable gun control law look like? Should I stop eating meat even though I love barbequed ribs? Have the costs come to outweigh the benefits so much that I finally need to deactivate my Facebook account?

Possession of natural language also allows a self-regulating mind not just to attend to and identify contents but to *ascribe* them. On the constraint-conforming approach, ascribing contents to an entity allows the ascriber to make sense of the entity in intentional terms, to understand what it has done and predict what it will do.[7] A self-regulating mind, moreover, can ascribe contents not just to other entities and organisms but also to it*self*. A person might judge that one content is false, hope another might eventually becomes true, aspire to actively help make another come to be.

Here too, the set of possible self-ascribable contents includes a subset of possible standards and rules—in principle any *constraints* a person can formulate and identify as such. Once a person selects a constraint-content and decides to adopt it as a rule for themselves, they can use a sentence to self-ascribe it: "I'm not going to take Benedick's word for it; I don't trust him anymore," "I'm going to try to not be so persuaded by ad hominem attacks",

[7] On some versions of the approach, the most important function performed by ascription is not predictive or explanatory but *regulative*; see especially McGeer 2007, 2015.

or "I believe factory farming is wrong, and so I no longer eat meat." For example, when Alice self-ascribes a constraint publicly, it signals to others that she embraces it as a standard against which she is willing to be evaluated, and will dedicate herself to trying to keep her various epistemic and practical pursuits in line with it. In ascribing the rule to herself she accepts it, voluntarily consents to the restrictions it will impose on her, and pledges to make an effort to act in ways that will satisfy it. In doing so, she exercises her capacity to self-regulate.

Well, almost. Imagine Alice self-ascribing a rule, and giving herself a morning pep talk in the mirror: "I will stand up for myself and not be interrupted in the staff meeting today!" When the moment comes, however, she still might not be able to bring herself to live up to the standard she has set for herself. She has seen what happens to outspoken women at her office, and during the meeting her resolve might crumble, overridden by fear, or her pressing desire to avoid the kinds of grimly effective social sanctions that have stifled female assertiveness in the past.[8] In self-ascribing the constraint, she will have put herself in a position to self-regulate; she will have done some preparatory work, decided on a self-regulatory agenda, and perhaps set the wheels in motion to achieve it.

But the third of the three subcapacities that underpin self-regulation on McGeer and Pettit's account is not self-ascription but *implementation*. Successful implementation—solving the problem of getting her activity to actually conform to the constraint she has identified and verbally ascribed to herself—is by no means an entirely linguistic or representational undertaking. If Alice is going to do more than just give lip service to any self-ascribed constraint, she has to somehow *enforce* it—give it functional oomph. Rather than just entertain the content, she must *impose* it upon herself in a way that allows it to effectively shape what she believes and does. This is a challenge exactly because doing so will in some cases require her to redirect herself, often overriding other desires that are at odds with the constraint, or stifling impulses and urges pulling her in the opposite direction. A plausible psychological story about implementation needs to say something not just about content, representational media, and language, but also about *motivation*.

I will return to motivation below, since implementation is quite a bit messier than exercise of the first two subcapacities. McGeer and Pettit have much more to say about self-regulation, but not much is directly about the motivational side of the picture on which I will focus (though see 287-290). Moreover, I will broaden their two-tiered picture to include avowed norms that can govern overt behavior. Most of McGeer and Pettit's discussion focuses on epistemic matters, the construction and maintenance of one's own regime of representational hygiene, and so deals with constraints that guide the formation and managing of beliefs and other belief-like states.[9] In so broadening it, I may be putting their picture to purposes they did not intend, and might not endorse. Nevertheless, McGeer and

[8] Thanks to Lacey Davidson for the example, and the suggestive comment that the phenomenology of cases like this tend to be very different than the phenomenology of, say, trying and failing to comply with other self-ascribed rules like *don't eat meat*. In the conclusion I briefly discuss cases like the former, in which a norm that an individual has personally avowed is at odds with another norm that she has internalized because it has been ascribed to her by her community.

[9] See Millgram 2014 for an illuminating discussion that is similar in spirit, and which introduces the terminology of representational hygiene. See Stich 1978 and Frankish 1998 for discussions more informed by cognitive science that concern different kinds of epistemic states.

Pettit's elegant account of merely routine minds and their architectural constraints, on the one hand, and self-regulating minds and their represented and self-imposed constraints, on the other, provides a useful, fairly high-level framework within which to situate a psychological distinction between internalized and avowed norms.

Section IV: Internalized Norms and Avowed Norms

In this section I continue to articulate the differences between internalized and avowed norms. In addition to differences in how each type of norm is typically initially adopted, there is reason to think there are concomitant differences between how internalized and avowed norms are psychologically realized. These differences in the functional role they occupy in an individual's mind, in turn, influence how instances of each type of norm relates to internal motivation, to introspection, to choice and willpower, to social pressure, and to how they might be incorporated into an individual's identity and self.

I will add detail and raise questions in a moment, but for a rough initial approximation this will suffice: a person has *internalized* a norm once it is represented in what I'll call her norm system. Internalized norms are typically automatically acquired, identified by dedicated psychological processes associated with imitation and social learning, soaked up from observing and participating in the interpersonal interactions of her community. Once a person has internalized a norm, she thereby becomes intrinsically motivated to act on it. There is growing enthusiasm in the cognitive and behavioral sciences for the idea that the lower tier of human minds comes equipped with a such a subsystem, a set of subpersonal routines dedicated specifically to norms and norm internalization (Sripada & Stich 2007, Chudek & Henrich 2011, Gelfand 2018, Kelly & Davis 2018). Evolutionary theorists have posited that this subsystem and our unique adeptness with socially learned and socially enforced rules is key to explaining our species' virtually unprecedented successes in spreading across the globe and dominating the planet (for better or worse). Our natural, intuitive sensitivity to such rules, social sanctions, and punishment-stabilized behavioral patterns in our social world would thus be largely responsible for our ability to thrive in a variety of habitats and to sustain the kind of social coordination needed to support large scale cooperation and collective action (Boyd and Richerson 2005, Henrich 2015, Boyd 2017, c.f. Sterelny 2014).

This specialized piece of human minds allows groups of people to generate and sustain collective patterns of behavior, but it can be analyzed at the level of the individual psychology as well. The principle functions of a person's norm system are to detect and acquire norms from her social environment, and to generate motivations to keep her own and other's behavior in line with those norms she has internalized. In the case of her own behavior, this will take the form of motivation to comply with the norm, while in the case of other people's behavior it will take the form of motivation to enforce it by sanctioning transgressors.

Making the full case for this idea will include providing more detailed functional specifications and accounts of the mechanisms that perform them, as well as a presentation of the current state of the evidence that supports it (see Setman and Kelly in preparation). For present purposes a few points can suffice. A useful first point is that on this view different kinds of norms can be internalized and executed by this system: dining norms,

sartorial norms, purity norms, epistemic norms, aesthetic norms, norms concerning care or justice. There is no *psychological* feature that imposes restrictions based on the specific domain of activity or value associated with internalized norms. Nor does this view entail commitment to a specific conception of morality; the subsystem is not reserved exclusively for *moral* norms, nor do rules *become* moral norms when they are acquired and become represented in a person's norm system. The basic view posits a specific functional role that internalized norms will come to occupy in an individual's mind, but does not advance any content-oriented limitations. Norms concerning virtually any subject matter might be internalized and thus come to occupy that role.

Second, the idea of a norm system fits within the picture of multi-tiered psychological organization discussed in the previous section. To a first approximation, the norm system operates like other subpersonal machinery of the mind, and the account portrays it as having many of the properties associated with subpersonal mechanisms in general; it performs its functions automatically, implicitly, non-deliberatively, without voluntary choice and sometimes in spite of conscious effort to the contrary. In McGeer and Pettit's terminology, the lower tier of human minds contains a merely routinized subsystem dedicated to norm internalization and execution. Like other merely routinized subsystems in the lower tier of the psychological hierarchy, this one searches the stream of perceptual input for cues and signs of environmental regularities relevant to its proprietary functions. These will include cues about the position and status of other people, and as well as regularities in their behavior—especially those regularities which, when deviated from, are sanctioned by others. When performing its acquisition function, the norm system will make inferences, likely guided by various constraints, about the rule being exemplified by the behavior and sanctioning pattern, and will deliver a representation of that rule to the database of internalized norms. In occupying this functional role in her mind, the norm becomes coupled to the person's motivational apparatus in a distinctive way. Once internalized, detection of the circumstances and types of people to which the norm applies will typically produce the system's routine set of motivational and behavioral outputs, pushing the person to conform to the norm and punish violations of it.

Third, there is a plausible, though still contested (see Andrews and Monsó, this volume), case that *only* human minds have this kind of routinized, norm-dedicated subcomponent. If this is right, then there is indeed a sense in which normativity is uniquely human, but perhaps not *only* in the reflective, individual-centric ways on which philosophers tend to focus.[10] Moreover, if this is right, then 'doing norms', like recognizing faces or being disgusted, and like detecting agency or parsing the meaning of a sentence in your native language, is not something you personally do. Rather it is, at least in some cases, something your mind does for you.

This leads to a key fourth point. A number of interesting similarities look to hold between the disgust system and the norm system: both appear to have universal, perhaps innately shaped structural features, but, via their associated domain specific mechanisms for acquisition and social learning, both are able to support considerable cross-cultural variability

[10] This focus is understandable, given many philosophers' interest in and lionization of individual autonomy and the associated processes of self-fashioning and self-constitution; see e.g. Korsgaard 1996, 2009; Anderson and Lanier 2001).

as well. The analogy with disgust is also particularly apt in light of the emotion's motivational properties. While a person's disgust system bears many features of merely routinized, system 1 subcomponents of human minds, some of the most striking are the downstream effects it has on a person when she detects something disgusting. Her disgust system will produce a nausea-like phenomenology; it will make her face into the instantly recognizable expression of the gape; it will unleash its characteristic influence on how she tends to think about the object of disgust, pushing her to conceive of it as offensive, dirty, and polluting; and it will generate strong motivation for her to get away from and continue to avoid the disgusting entity (see Kelly 2011 for details). Moreover, a person's disgust system initiates the routines that produce all these effects, including the motivational ones, automatically, without volition, and sometimes despite what the person reflectively knows or thinks about the thing that activates the wave of revulsion—turd shaped fudge and rubber vomit are two common examples.

Returning to the norm system, an idea worthy of more investigation is that claim that the motivations associated with it are similar to this in many respects. Call these—conative states produced by the norm system as it performs its function of inducing an individual to comply and enforce internalized norms—*normative motivations*. There is a core set of open questions concerning the nature of these normative motivations, centered on the details of their neural and psychological implementation and evolutionary history, and their susceptibility to the influence of self-control and other forms of personal and collective level intervention. A particularly pressing empirical puzzle about normative motivations is their relation to other conative states and processes. Are they best understood as being composed out of other, more familiar mental states like desires and emotions, or are they better conceived as constituting a *sui generis* category, perhaps psychologically constructed in a unique way?[11] Normative motivations may be *intrinsic* in some sense, and certainly appear to be distinct from, and can in some cases be more powerful than, a person's self-interested desires and the kinds of personal preferences that initiate more instrumentally motivated behaviors. Their associated phenomenology often has a distinctive potency as well, leading one recent commentator to remark that they appear to be made up of a "puzzling combination of objective and subjective elements" (Stanford 2018, 2).

There is much research to be done here, and the distinction between internalized and avowed norms will be useful in structuring it. For whatever the character of the normative motivations generated by the norm system and associated with internalized norms, it appears to be *markedly different* from whatever sources of motivation a person needs to draw on in order to keep her activity in line with norms she has chosen for herself. Deliberately reflecting on a norm, and then selecting, avowing, and consciously imposing it on oneself—implementing a constraint, in McGeer and Pettit's terminology—is part of the activity of self-regulation rather than mere routine, and will likely be underpinned by very different psychological machinery.

Such differences are likely to be found along a number of dimensions, representational as well as motivational. For instance, internalized norms will often be architectural, but avowed norms by definition will be reflectively represented (c.f. Clark

[11] See Feldman Barrett (2017) for more on the idea of psychologically constructed emotions, drives, and other motivating mental states.

2000). There might be other differences in the representational *format* of the internalized and avowed norms as well; after all, cognitive science has discovered variety in the format of mental representations that drive categorization and classification, e.g. exemplars, prototypes, stereotypes, concepts, etc. (Machery 2011). There is no prima facie reason the human mind might not contain a similar variety of representational formats for norms as well.[12] These could be teased apart and investigated using the same kind of careful experimentation used in the literature on categorization.

There may also be limits on the abstractness of internalized norms that do not apply to avowed norms. For instance, Alice might make a genuine New Year's resolution: "I will lead a healthier lifestyle in 2019". On its own, however, this does not straightforwardly operationalize into any specific action or rule. It is clearly more abstract that "avoid the bar tonight", "don't eat meat", or "run three miles every morning". "Be healthier" or "make healthier decisions" are both less actions or rules than they are expressions of a more general goal, or of a broad value that Alice might embrace. Her commitment to the value of health can in turn help guide her formulation of more specific norms she can avow and impose on herself, behavior guiding rules with more articulated cues and conditions in which they apply, and more determinate behaviors that she will attempt to produce in response to them.[13] Given the way that internalized norms are automatically acquired from social interactions, and the nature of the routinized links between cue and response supported by the subpersonal mechanisms that underlie them, it may be that only rather concrete rules can get into a person's norm box, where "concrete" means having fairly detailed specifications of their application conditions and appropriate responses.[14]

As noted at the outset of Section III, an individual's claims to self-knowledge about those norms she has internalized versus those norms she has personally avowed are likely to be importantly and interestingly different as well. Self-ascriptions about internalized norms are likely to be descriptive, mere reports that are susceptible to the same kinds of inaccuracies and failures as ascriptions of mental states to other people, and perhaps

---

[12] See Stich (1993) for an early defense of the idea that norms might be represented as prototypes and exemplars.

[13] Ismael (2016) considers the example of health in the context of her theory of self-governing cognitive systems, which posits the same kind of broadly two-tiered picture of merely routinized and self-regulated psychological organization that I have been developing in this paper. The goal to "be healthy" is an example of a self-imposed mental state so abstract that it

> "can't itself be embodied in a drive or appetite because it doesn't have a built-in connection to a particular set of behaviors. Achieving good health demands different behaviors in different circumstances. Sometimes it means eating less, sometimes it means eating more, sometimes it means exercising more, and sometimes it means rest. It is the paradigm of a goal whose connection to behavior is mediated by explicit representation of the agent's circumstances, the desired end, and choice of action that depends on the relationship between them ("that is where I want to be, this is where I am, how do I get there?"). Appetites don't have this structure. They have a built-in drive to perform a particular kind of behavior: eat, drink, have sex" (68).

[14] For a similar example, compare abstract goals one might adopt like "be more racially egalitarian" or "I will strive to be less sexist", on the one hand, to implementation intentions, on the other. Implementation intentions are very specific if-then rules one can deliberately self-impose, for instance rehearsing to oneself "if I see a Black face, I will think 'safe'". These work by rerouting a particular cue (Black face) from a response it was previously paired with (fear, aversion) to a new response (safe). Perhaps surprisingly, these have been shown to be effective in helping to mitigate the effects of implicit biases (Gollwitzer and Sheeran 2006; also see Brownstein et al forthcoming for discussion of the recent controversies about the Implicit Association Test).

underpinned by the same kinds of mentalizing psychological mechanisms, directed at oneself rather than at others.[15] On the other hand, the act of avowing a norm (or any other mental state) is less purely descriptive than it is performative, less of an observation and more of a pledge. In these cases of self-ascription that are avowals, the individual is making a conscious, personal level decision and undertaking voluntary mental action. It is plausible that when it comes to avowed norms, a person can indeed claim a different kind of epistemic privilege and special sort of first-person authority.[16]

However, the core and perhaps most fundamental differences between internalized and avowed norms are likely to be linked to motivation. An initial recommendation, inspired by the discussion in Section II, is that given the vexed issue of what counts as 'moral', psychological research into motivation associated with norms may make better progress if it is structured by questions concerning the differences between internalized and avowed norms, and not by questions about the character of moral motivation or moral norms. There is still no agreed upon account of which norms are 'moral', and continuing to frame questions and results in terms of 'moral motivation' or 'moral cognition' without one is likely to add to the Tower of Babel-esque confusion (c.f. Haidt 2001). As the analogy with disgust suggested, the normative motivations imbued to internalized norms appear to share many properties with the kind of motivation associated with other subsystems in the routinized part of human minds. The motivation associated with avowed norms is a thing apart, and appears to have more in common with the subject matter of other areas of research.

Exercising self-control is often notoriously difficult, and using conscious will power to shape one's behavior is a recognizably distinct kind of struggle, whether it be to briefly refrain from eating a marshmallow, or to forgo cigarettes and ribeye steak forever, or to get up and run every morning, stop procrastinating, speak out at a staff meeting, or to put more trust in women's testimony. While there is little empirical psychological literature on avowal and norms—at least not under this description—several areas of extant research look promising as starting points and building blocks.[17] Fulfilling a personally avowed norm—satisfying a constraint one imposes on oneself in an act of self-regulation—is likely to initially be a continuous struggle no matter how epistemically convincing one finds the case in favor of doing so. As a result, effectively keeping oneself bound by an avowed norm will require a package of elements: occurrent (maybe self-activated) conative states, short term tactics, and a long-term strategy, the deployment of which constitutes a process that is extended in space and time. It is likely to involve the same psychological resources that drive

[15] See especially Carruthers 2011 for defense of the idea that a person's ability to read her own mind is not different in kind or with respect to underlying mechanism from her ability to read others' minds; she just has more evidence about her own behavior than she does about anyone else's. Also see Wilson 2002 for the idea that most people are 'strangers to themselves' with respect to large swaths of their own psychological makeup.
[16] There are, of course, complications. Some of the more interesting cases are those that go beyond the difficulties associated with merely paying lip service to a norm, and into the territory of alienation and estrangement. See especially Moran 2001 on estrangement and self-knowledge. Also see Doris 2015 for discussion of the role of verbalization and rationalization in supporting agency. Doris proceeds from a perspective that stays closer to the contemporary empirical picture and takes seriously the effect of automatic and implicit psychological machinery in producing behavior.
[17] The lack of psychological attention contrasts with philosophical work, where Gibbard's (1990) development of a norm-expressivist metaethical theory sparked a substantial literature in response. Most of that, however, focuses on logic and semantics, and to a lesser extent the metaphysics, rather than working out theories of the cognitive and motivational machinery that underpins avowed norms.

willpower and *self-control* (Sripada 2014), and to leverage an ability to form and keep to *habits* (Brownstein 2018, especially section 3). A full account of how people marshal their own motivation to comply with avowed norms will also take note of human's hypertrophied ability to take and exert *ecological control* over themselves, to adopt technologies and actively construct their own environments in ways that support their agency, channeling their behaviors towards their reflective goals and towards ends that they evaluatively endorse[18]. With internalized norms, motivation is intrinsic, and so 'comes for free'. Not so for avowed norms. In the latter case, an individual has to figure out how to get that norm into her motivational driver's seat so that she will satisfy it, to find ways to allow the norm to guide and restrict her own behavior, even in the face of competing motivations, urges, and impulses when they arise. As others have noted, this picture of struggling to satisfy an avowed norm mimics the general structure of commitment problems, and the formal understanding of *commitment devices* could be useful in shedding light on the social and psychological resources humans have developed to navigate them (Frank 1988 and Kelly 2011 chapter 3 for discussion, Elster 2000, and Nesse 2001a and 2001b). The field is ripe for exploration.

Section V: Concluding Philosophic Postscript

Recall Alice. Her efforts to figure out who she is and become who she wants to be should look familiar, but hopefully the familiarity doesn't obscure how wonderful and mystifying and important and terrifying and fulfilling and psychologically intricate the whole thing can be. It is a process which mixes the private and the public, description and performance, and in which "the distinction between discovery and creation breaks down in a fascinating and distinctive way" (Ismael 2016, 13).

I've devoted the bulk of this paper to norms, making the case that there is an important psychological distinction between norms that an individual like Alice adopts by personally avowing them and norms that she has internalized from her social environment because a specialized part of her mind detected and acquired them for her. I located that distinction with respect to the larger literatures on norms, moral psychology, and collective behavior, and went on to develop some theoretical resources that might be used to account for it. I began integrating McGeer and Pettit's constraint-conforming approach to self-regulation with a two-tiered and patchwork account of human psychological organization, and pointed to a body of literature that is making the case that one of the subpersonal, routinized mechanisms in the lower tier of human minds is dedicated to acquiring norms and generating a special kind of motivation to comply with and enforce them. The picture is attractive, but questions remain, especially concerning motivation, and there is much exciting research yet to be done.

Parts of Alice's project can be made sense of using these resources. Another common milestone on a journey like hers may begin with a personal revelation, of the sort than can be either slowly dawning or can come in an eruptive, flashbulb burst of self-awareness. However it unfolds, Alice realizes that not only has she been subject to a sexist norm, but that she herself has internalized that same norm from her patriarchal community.

[18] See Clark 2007 on ecological control, and Holroyd and Kelly 2016 for the distinction between taking and exerting it.

It is a norm that she never consented to, and upon reflection does not endorse (e.g. *the testimony of men is more credible than the testimony of women*, or *women should not be assertive or express anger in the workplace*). She can respond to her newfound knowledge by publicly denouncing the norm, taking steps to uproot it in herself, and avowing a new feminist norm that is at odds with the old sexist one. Her discovery and rebellion, however, may not by themselves completely loosen the hold the old norm has over her, or fully cancel the effects it has on her behavior and judgment. Merely disavowing the sexist norm she has internalized is unlikely to immediately dislodge it from her norm system, or fully defuse the internal pull it exerts to keep her behavior in line with it.[19]

Though I have been focused on their internal psychological differences in this paper, it is worth noting that the public lives of internalized norms and avowed norms are likely to be interestingly different as well. For example, in and of itself, Alice's revelation and disavowal of a sexist norm that prevails in her patriarchal community probably fails to remove it straightaway from her own mind, and it will obviously not delete it from everyone else's minds, either. Even her public rejection of the norm will not completely block the influence of the external social pressure those others apply to her in order to keep her in compliance with it. Alice unfortunately does not get to decide whether or not she is subject to this norm in this way, and despite her denouncement of it and her avowal of a new feminist norm, she will continue to be penalized by her community when she violates the old sexist one. One can easily imagine her getting angry about the situation, and how doubly infuriating it must be when expressing that very anger is seen as another transgression, drawing more communal reprimand.

This illustrates the ascriptive character of many norms, especially role-specific ones. Many such norms will be *ascribed* to Alice by others simply because she occupies a particular social role within her community (in this case the social role of being a woman). In virtue of this, she will be evaluated by, and her behavior will become sensitive to, those ascribed norms regardless of whether she has agreed to them or not, of whether she has avowed or disavowed them, and of whether she is even consciously aware of them (Witt 2011). Of course, not all norms that influence Alice are ascribed by her community in this way, but over the course of her lifetime some of the social roles and norms with which she will have to wrangle certainly will be. But other social roles she will be able to more *voluntarily* opt into and out of; likewise, other norms she will be able to select and self-impose, or to reject. Still other social roles—like competent surfer, Civil War buff, marathon runner, or US Senator—she can *aspire* to, and then intentionally pursue and perhaps successfully achieve (also see Callard 2018). The role of private, individual choice looms much larger in these latter voluntary and achieved cases, while the role of public factors like cultural practices, social structures, and other members of Alice's community are more prominent in the former, ascribed ones.[20]

---

[19] The situation described here is meant to parallel the kind of dissociation and conflict between explicit and implicit attitudes that has been much remarked up on the literature on implicit bias (Brownstein and Saul 2016). Also see Stich 1983 for an earlier discussion of the idea, similar in spirit if not detail (he is concerned with belief-like states rather than norms), that the human mind "keeps two sets of books" (231).

[20] See Davidson and Kelly (2018) for an examination of Witt's position, a discussion of norms, social roles, and soft social structures, and an initial expression of the kind of pluralism developed in this paper. The internalized / avowed distinction is not quite the same as the ascribed / chosen distinction, but for the most part, ascribed

This merely scratches the surface of the differences in the public lives of avowed norms and internalized norms, differences that lay beyond the psychological roles they occupy in the minds of individuals who have adopted them. But appreciating differences in the psychological underpinnings of avowed and internalized norms can shed light on how each type behaves in more public contexts, and thus on a number of issues of philosophical interest. For example, norms of each type may be interestingly different in how they interact not just with individual reflection and non-verbal kinds of social pressure, but with *norm talk*: language and verbal persuasion, public opinion in the form of linguistically articulated justification and interpersonal criticism (Lamm 2014, Bicchieri 2016, Mercier and Sperber 2017, Shank et al 2018; c.f. Summers 2017). Indeed, the distinctive types of normativity and agency associated with avowal and internalization, respectively, may typically be more or less collaborative, and in different ways (Doris and Nichols 2012, Doris 2015).

Moreover, many of our social practices are vaguely sensitive to these kinds of differences in norms and norm-governed behavior, and more generally to differences between behaviors that originate in processes found in higher versus lower levels of the hierarchy of human psychological organization. Those of us in WEIRD individualistic cultures like Alice are especially keen on *choice* and individual *selfhood*. Thus, we are attuned to whatever features of a behavior might signify that it was voluntarily chosen, and so may accurately reflect some genuine inner self. Our practices appear to treat intentional identification and avowal as evaluative, and as being intertwined with responsibility: those actively chosen behaviors that are seen as expressions of a true and authentic identity are also taken to be worthier of praise and blame. Conversely, we seem more willing to dismiss as incidental those things that merely happen to a person, or those behaviors that are produced in a more passive way—things her mind made her do but that she wouldn't reflectively endorse. Indeed, we seem content to allow a person to disavow the latter kinds of behaviors because we act as if that whatever caused them, it wasn't really *her* (Strohminger and Nichols 2014, Strohminger et al 2017). Much effort has been spent trying to reconstruct philosophically defensible versions of this distinction, between things a person does and things that happen to her, and to characterize what is distinctive and special about actions that are taken to be genuinely one's own. These efforts are informed by concerns about how social practices surrounding moral responsibility, praise, and blame should deal with the distinction (Wolf 1993, Smith 2012, Vargas 2013, Sripada 2016), what exactly it has to do with the structure of agency, (Bratman 2007) and how it is related to the metaphysics of personal identity (Millgram 2013).[21]

Of course, the public and the private blend into each other. Indeed, the fluid boundaries between the two are constantly being renegotiated (Igo 2018), and cultural conceptions of selves and individuals vary and evolve along with those negotiations (Ross

---

social roles are likely to involve mostly internalized norms, while voluntary and achieved social roles are likely to involve a mixture of ascribed and avowed norms.

[21] Millgram also raises the worry that too many philosophical demands have been made of accounts of this kind of distinction, and that philosophers attempting to capture it have been led astray by trying to serve too many masters. This point is especially pertinent for the purposes of this paper, given his claim that it is also "part of philosophical commonsense to have qualms about how *psychologically realistic* such elaborate constructions can be" (Millgram 2013, page 240, italics added).

2012). These complications are just part of what make the construction of good psychological, social, and moral theories of all of these fascinating and all-too-human phenomena so maddeningly difficult. They are also part of what make the personal project itself, of deciding on a set of norms and values, of weaving together an identity from what one has been given and what can be been chosen, of aspiring to and establishing a self of one's own, so disorienting and crucial and fraught and thrilling. But don't take my word for it. Go ask Alice.

References

Anderson, R. L. and Landy, J. (2001). "Philosophy as Self-Fashioning: Alexander Nehamas's Art of Living," *Diacritics*, 31(1): 25-54.

Andrews, K. and Monsó, S. (forthcoming). Animal moral psychologies. *The Oxford Handbook of Moral Psychologies*, eds M. Vargas and J. Doris. New York: Oxford University Press.

Baumard, N., Andre, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, 36(1), 59–78.

Bendor, J. and Swistak, P. (2001). "The Evolution of Norms," *American Journal of Sociology*, 106(6): 1493–1545.

Bicchieri, C. (2006). *The Grammar of Society: the Nature and Dynamics of Social Norms*, New York: Cambridge University Press.

Bicchieri, C. (2016). *Norms in the Wild* Oxford: Oxford University Press.

Bicchieri, C. and Muldoon, R. (2014). "Social Norms", The Stanford Encyclopedia of Philosophy (Spring 2014 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2014/entries/social-norms/>.

Boyd, R. (2017). *A Different Kind of Animal: How Culture Transformed Our Species* Princeton University Press, Princeton, NJ.

Bratman, M. (2007). *Structures of Agency: Essays*. New York: Oxford University Press.

Brennan, G., Eriksson, L., Goodin, R. and Southwood, N. (2013). *Explaining Norms*. Oxford, Oxford University Press.

Brownstein, M. (2018). *The Implicit Mind: Cognitive Architecture, the Self, and Ethics*. Oxford: Oxford University Press.

Brownstein, M., Madva, A. and Gawronski, B. (Forthcoming). "Understanding Implicit Bias: Putting the Criticism into Perspective," *Pacific Philosophical Quarterly*.

Brownstein, M. and Saul, J. (2016). *Implicit Bias and Philosophy, Volumes 1 & 2*, Oxford: Oxford University Press.

Callard, A. (2018). *Aspiration: The Agency of Becoming*. Oxford: Oxford University Press.

Carruthers, P. (2011). *The Opacity of the Mind*. New York, Oxford University Press.

Centola, D., Willer, R. and M. Macy (2005). "The Emperor's Dilemma: A Computational Model of Self-Enforcing Norms," *American Journal of Sociology* 110:4, 1009-1040.

Chudek, M. and Henrich, J. (2011). 'Culture–gene coevolution, norm-psychology and the emergence of human prosociality,' *Trends in cognitive sciences*, *15*(5), 218-226.

Cialdini, R. B., Kallgren, C. A. & Reno, R. R. (1991). A focus theory of normative conduct: a theoretical refinement and reevaluation of the role of norms in human behavior. *Adv. Exp. Soc. Psychol.* 24, 201–234.

Clark, A. (2000). "Word and Action: Reconciling Rules and Know-How in Moral Cognition," *Canadian Journal of Philosophy*, 30(1): 267-289.

Clark, A. (2007). "Soft Selves and Ecological Control," in D. Spurrett, D. Ross, H. Kincaid and L. Stephens (eds) *Distributed Cognition and the Will*. Cambridge, MA: The MIT Press.

Curry, O. (2016). "Morality as Cooperation: A Problem-Centred Approach," in *The Evolution of Morality*, T.K. Shackelford and R.D. Hansen, Editors., Springer International Publishing. p. 27-51.

Cushman, F., Young, L., and Greene, J. (2010). 'Multi-system Moral Psychology,' *The Oxford Handbook of Moral Psychology*, Eds. J. Doris et al. New York: Oxford University Press, page 47 – 71.
Davidson, L. and Kelly, D. (2018). 'Minding the Gap: Bias, Soft Structures, and the Double Life of Social Norms,' *Journal of Applied Philosophy*, 1-21.
Davis, T. (under review). "The Scope and Structure of the Moral Domain: An Empirical Study".
Davis, T., Hennes, E. & Raymond, L. (2018). "Normative Motivation and Sustainable Behavior: New Insights from an Evolutionary Perspective," *Nature: Sustainability* 1: 218 – 224.
Davis, T. and Kelly, D. (2018). 'Norms, Not Moral Norms: The Boundaries of Morality Don't Matter' commentary on Kyle Stanford "The Difference Between Ice Cream and Nazis: Moral Externalization and the Evolution of Human Cooperation," *Behavioral and Brain Sciences*, 18-19
Dennett, D. (1969). *Content and Consciousness*. New York: Routledge & Kegan Paul Books Ltd.
Dennett, D. (1981). "True Believers: The Intentional Strategy and Why it Works," in A. F. Heath, ed., *Scientific Explanation*, (the Herbert Spencer Lectures at Oxford), Oxford University Press.
Doris, J. (2015) *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford: Oxford University Press.
Doris, J. and Nichols, S. (2012). "Broadminded: Sociality and the Cognitive Science of Morality." In E. Margolis, R. Samuels, and S. Stich (eds.), *The Oxford Handbook of Philosophy and Cognitive Science,* pp. 425-53. Oxford: Oxford University Press
Doris, J. and Plakias, A. (2007). "How to Argue About Disagreement: Evaluative Diversity and Moral Realism." In W. Sinnott-Armstrong (Ed.), *Moral Psychology, vol. 2: The Biology and Psychology of Morality* (pp. 303-332). Oxford: Oxford University Press.
Drayson, Z. (2014). "The Personal/Subpersonal Distinction," *Philosophy Compass*, 9(5): 338-346.
Elster, J. (2000). *Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints.* Cambridge: Cambridge University Press.
Feldman Barrett, L. (2017). *How Emotions Are Made: The Secret Life of the Brain*. New York: Mariner Books.
Flanagan, O. (2016). *The Geography of Morals: The Varieties of Moral Possibility*. New York: Oxford University Press.
Frank, R. (1988). *Passions within reason: The strategic role of the emotions*. New York: W.W. Norton & Company.
Frankish, K. (1998). "A Matter of Opinion," *Philosophical Psychology*, 11(4): 423 – 442.
Gelfand, M. (2018). *Rule Makers, Rule Breakers*. New York: Scribner.
Gibbard, A. (1990). *Wise Choices, Apt Feelings*. Harvard University Press
Gilligan, C. (1982). *In a different voice*. Cambridge: Harvard University Press.
Gollwitzer, P. M., & Sheeran, P. (2006). "Implementation intentions and goal achievement: A meta-analysis of effects and processes," In M. P. Zanna (Ed.), *Advances in experimental social psychology*, 69–119. New York: Academic Press.
Goodwin, G. P., & Darley, J. M. (2008). The psychology of meta-ethics: Exploring objectivism. Cognition, 106(3), 1339–1366.

Goodwin, G. P., & Darley, J. M. (2010). The perceived objectivity of ethical beliefs: Psychological findings and implications for public policy. Review of Philosophy and Psychology, 1(2), 161–188.
Goodwin, G. P., & Darley, J. M. (2012). Why are some moral beliefs perceived to be more objective than others? Journal of Experimental Social Psychology, 48(1), 250–256.
Graham J, Nosek B, Haidt J, Iyer R, Koleva S, Ditto P. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2): 366–385. doi:10.1037/a0021847
Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The Pragmatic Validity of Moral Pluralism", In *Advances in Experimental Social Psychology* (Vol. 47).
Greene, J. (2014). *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. New York: Penguin Books.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*. 108, 814-834.
Haidt, J. (2012). *The Righteous Mind*. New York: Pantheon Books.
Henrich, J. (2015). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, NJ: Princeton University Press.
Heyes, C. (2018). *Cognitive Gadgets: The Cultural Evolution of Thinking*. Cambridge, MA: Harvard University Press.
Holroyd, J. and Kelly, D. (2016). 'Implicit Bias, Character, and Control', *From Personality to Virtue: Essays in the Philosophy of Character*. Eds. A Masala and J. Webber. Oxford University Press, pages 106 - 133.
Igo, S. (2018). *The Known Citizen*. Cambridge, MA: Harvard University Press.
Ismael, J. (2014). 'On Being SomeOne,' In A. Mele (eds.), *Surrounding Free Will: Philosophy, Psychology, Neuroscience*, Oxford University Press, page 274 – 297.
Ismael, J. (2016). *How Physics Makes Us Free*. Oxford, UK: Oxford University Press.
Kahneman, D. (2011). *Thinking Fast and Slow*. New York: Farrar, Straus and Giroux.
Kelly, D. 2011. *Yuck! The Nature and Moral Significance of Disgust*. Cambridge, MA: The MIT Press.
Kelly, D. and Davis, T. (2018). 'Social Norms and Human Normative Psychology,' *Social Philosophy & Policy*. 35(1): 54 – 76.
Kelly, D. and Stich, S. (2007). 'Two Theories of the Cognitive Architecture Underlying Morality,' *The Innate Mind Vol 3.: Foundations and Future Horizons*, Eds. Peter Carruthers, Stephen Laurence and Stephen Stich. New York: Oxford University Press. Pages 348-366.
Kelly, D., S. Stich, K. Haley, S. Eng, and D. Fessler. 2007. Harm, affect, and the moral/ conventional distinction. *Mind and Language* 22 (2): 117–131.
Kitcher, P. (2011). *The Ethical Project*. Cambridge, MA: Harvard University Press.
Kohlberg, L. (1981). *The Philosophy of Moral Development: Moral Stages and the Idea of Justice (Essays on Moral Development, Volume 1)*. Harper & Row.
Korsgaard, C. (1996) *The sources of normativity*. Cambridge University Press.
Korsgaard, C. (2009) *Self-constitution: Agency, identity, and integrity*. Oxford University Press.
Kumar, V. (2015). "Moral judgment as a natural kind," *Philosophical Studies* 172 (11): 2887-2910.

Lamm, E. (2014). "Forever united: the coevolution of language and normativity," in Daniel Dor, Chris Knight and Jerome Lewis (eds.) *The Social Origins of Language: Early Society, Communication and Polymodality*. Oxford University Press, pages 267 – 283.
Lewis, David, 1969. *Convention*. Cambridge: Harvard University Press.
Machery, E. (2011). *Doing Without Concepts*. New York: Oxford University Press.
Machery, E. (2012). Delineating the moral domain. *Baltic International Yearbook of Cognition, Logic and Communication*, 7(1).
Mameli, M. (2013). Meat made us moral: a hypothesis on the nature and evolution of moral judgment. *Biology & Philosophy* 28: 903–931.
McGeer, V. and Pettit P. (2002). 'The Self-Regulating Mind,'. *Language & Communication*. 22: 281-299.
McGeer, V. (2007). "The Regulative Dimension of Folk Psychology." In *Folk Psychology Re-Assessed*, edited by D. Hutto and M. Ratcliffe, 137–156.
McGeer, V. (2015). "Mind-making practices: the social infrastructure of self-knowing agency and responsibility," *Philosophical Explorations*, 18(2): 259–281.
Mercier, H. and Sperber, D. (2017). *The Enigma of Reason*. Cambridge, MA: Harvard University Press.
Mikhail, J. (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge University Press.
Millgram, E. (2014). "Private Persons and Minimal Persons," *Journal of Social Philosophy*, 45 (3): 323–347.
Millgram, E. (2015). "Segmented Agency," In M. Vargas and G. Yaffe, *Rational and Social Agency: The Philosophy of Michael Bratman* (Oxford: Oxford University Press, 2014): 152-189. Reprinted (with postscript) in *The Great Endarkenment* (New York: Oxford University Press, 2015).
Moran, R. (2001.) *Authority and Estrangement: An Essay on Self-Knowledge*, Princeton, NJ: Princeton University Press.
Morris, M., Hong, Y., Chiu, C. and Liu, Z. (2015). "Normology: Integrating insights about social norms to understand cultural dynamics," *Organizational Behavior and Human Decision Processes*, 129: 1–13.
Nesse, R. ed. (2001a). *Evolution and the Capacity for Commitment*. New York: Russell Sage Foundation Publications.
Nesse, R. (2001b). "Natural Selection and the Capacity for Subjective Commitment," in *Evolution and the Capacity for Commitment*, ed. R. Nesse, New York: Russell Sage Foundation Publications.
Nichols, S. 2004. *Sentimental rules: On the natural foundations of moral judgment*. New York: Oxford University Press.
Nisbett, R. (2003). *The Geography of Thought*. New York: The Free Press.
Nucci, L. (2001). *Education in the moral domain*. Cambridge: Cambridge University Press.
O'Neill, E. (2017). "Kinds of Norms," *Philosophy Compass*, 12(5): 1-15. https://doi.org/10.1111/phc3.12416
Ostrom, E. (2000). "Collective Action and the Evolution of Social Norms," *Journal of Economic Perspectives*, 14(3): 137–158.
Ross, D. (2012). "The evolution of individualistic norms," In Kim Sterelny, Richard Joyce, Brett Calcott & Ben Fraser (eds.), *Baltic International Yearbook of Cognition, Logic and Communication*. MIT Press. pp. 1-33.
Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes

(community, autonomy, divinity). *Journal of Personality and Social Psychology*, 76(4), 574-586.
Schein, C., & Gray, K. (2017). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology,* 27, 1–39.
Schultz, P. W. et al. e constructive, destructive, and reconstructive power of social norms. *Psychol. Sci.* 18, 429–434 (2007).
Schultz, P. W. et al. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science.* 18, 429–434.
Setman, S. and Kelly, D. (in preparation). "The Psychology of Normative Cognition," entry for the *Stanford Encyclopedia of Philosophy*.
Shank, D. B., Kashima, Y., Peters, K., Li, Y., Robins, G., & Kirley, M. (2018). Norm talk and human cooperation: Can we talk ourselves into cooperation? *Journal of Personality and Social Psychology.* Advance online publication.
Shweder, R., Much, N., Mahapatra, M. and Park, L. (1997). The "big three" of morality (autonomy, community, and divinity), and the "big three" explanations of suffering. In A. Brandt & P. Rozin (eds.), *Morality and Health.* Routledge.
Sinnott-Armstrong, W. & Wheatley, T. (2012). 'The Disunity of Morality and Why it Matters to Philosophy,' *The Monist* 95(3): 355-377.
Smetana, JG (1993). Understanding of social rules. In M. Bennett (Ed.), *The development of social cognition: The child as psychologist*. New York, NY, US: Guilford Press, pp. 111-141.
Smith, A. (2012). "Attributability, answerability, and accountability: In defense of a unified account." *Ethics* 122(3): 575–89.
Smith, J. R. et al. (2012). Congruent or conflicted? The impact of injunctive and descriptive norms on environmental intentions. *J. Environ. Psychol.* **32**, 353–361 (2012).
Sommer, T. (2012). *Relative Justice: Cultural Diversity, Free Will, and Moral Responsibility.* Princeton, NJ: Princeton University Press.
Southwood, N. (2011). The moral/conventional distinction. Mind, 120(479), 761–802.
Southwood, N. and Eriksoon, L. (2011). "Norms and Conventions," *Philosophical Explorations,* 14(2):195–217.
Sparkman, G. & Walton, G. (2017). Dynamic Norms Promote Sustainable Behavior, Even if It Is Counternormative. *Psychological Science.* Vol. 28(11) 1663–1674.
Sripada, C. (2014). "How is Willpower Possible? The Puzzle of Synchronic Self-Control and the Divided Mind," *Nous 48*(1), 41-74.
Sripada, C. (2016). "Self-expression: a deep self theory of moral responsibility," *Philosophical Studies* 173 (5): 1203-1232
Sripada, C. and S. Stich. (2007). A framework for the psychology of norms. P. Carruthers, S. Laurence, & S. Stich (eds.), *The innate mind: Culture and cognition.* New York: Oxford University Press, pages 280-301.
Stanford, K. (2018). "The Difference Between Ice Cream and Nazis: Moral Externalization and the Evolution of Human Cooperation," *Behavioral Brain Sciences*, 1 – 13.
Sterelny, K. (2012). "Morality's Dark Past," *Analyse & Kritik* 34 (1):95-115
Sterelny, K. (2014). "Cooperation, Culture, and Conflict," *The British Journal for the Philosophy of Science*, 67(1): 1 – 31
Stich S. (1978.) "Beliefs and sub-doxastic states," *Philosophy of Science*, 45: 499–518.
Stich, S. (1983). *From folk psychology to cognitive science: The case against belief.* Cambridge, MA: The MIT Press.

Stich, S. (1993). "Moral Philosophy and Mental Representation," in R. Michod, L. Nadel & M. Hechter (eds.), *The Origin of Values*. Aldine de Gruyer. pp. 215-228.
Stich, S. (2018) "The Quest for the Boundaries of Morality," in Karen Jones, Mark Timmons & Aaron Zimmerman, eds., *The Routledge Handbook of Moral Epistemology*, (New York: Routledge).
Stich, S., Mizumoto, M. and McCready, E, (eds). (2018) *Epistemology for the Rest of the World*. New York: Oxford University Press.
Strohminger, N., and Nichols, S. (2014). "The essential moral self," *Cognition* 131: 159–171
Strohminger, N., Knobe, J. and Newman, G. (2016). "The True Self: A psychological concept distinct from the self," Perspectives in Psychological Science 12(4):551-560
Summers, J.S. (2017). "Rationalizing our Way into Moral Progress," *Ethical Theory and Moral Practice,* 20(1), 93-104.
Turiel, E. 1983: *The Development of Social Knowledge*. Cambridge: Cambridge University Press.
Vargas, M. (2013). *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.
Wilson, T. (2002.) *Strangers to ourselves*. Cambridge, MA: Harvard University Press.
Witt, C. (2011). *The Metaphysics of Gender*. New York: Oxford University Press.
Wolf, S. (1993). *Freedom within reason*. New York, NY: Oxford University Press.
Young, P. (2015). "The Evolution of Social Norms," *Annu. Rev. Econ*. 7: 359–87.