Chapter 10
The Origin of Moral Norms and the Role of Innate Dispositions

Jessy Giroux

10.1 Introduction[1]

When I hear the question 'What makes us moral?' I am reminded of a more ambitious question of which 'what makes us moral?' is probably a narrowed-down version. That more ambitious or perhaps simply broader question is 'Where does morality come from?' I am mainly interested in the latter question in this article, although my goal is to link that question to the former one. I intend to do this by investigating how 'the things that make us moral' – by which I mean our morally-relevant dispositions and capacities, such as empathy, vicarious distress, our natural propensity to cooperate and reciprocate, our ability to formulate and follow rules, etc. – bring societies to develop similar moral norms.

Let me begin by presenting two general ways of understanding where moral norms come from, how they originate or come about in human societies. First, one can argue that moral norms are no different from conventional norms in their origin. Like etiquette or sartorial norms, for instance, moral norms could be seen as the product of a variety of historical (geographical, religious, political, etc.) contingencies. On this minimalist account, asking why a society prohibits murder is no different from asking why it discourages eating spaghetti with one's hands or wearing white socks with black shoes. Although there certainly is a difference in *degree* between moral and conventional infractions, there is no *qualitative* difference.

Second, one can describe moral norms as naturally emerging from human constitution. According to this model, moral norms, unlike spaghetti or sock-colour

[1] An ancestor of this article was published in 2011 (see Giroux 2011).

J. Giroux (✉)
PhD Student, Department of Philosophy, University of Toronto, Toronto, ON, Canada
e-mail: jessy.giroux@mailutoronto.ca

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_10,

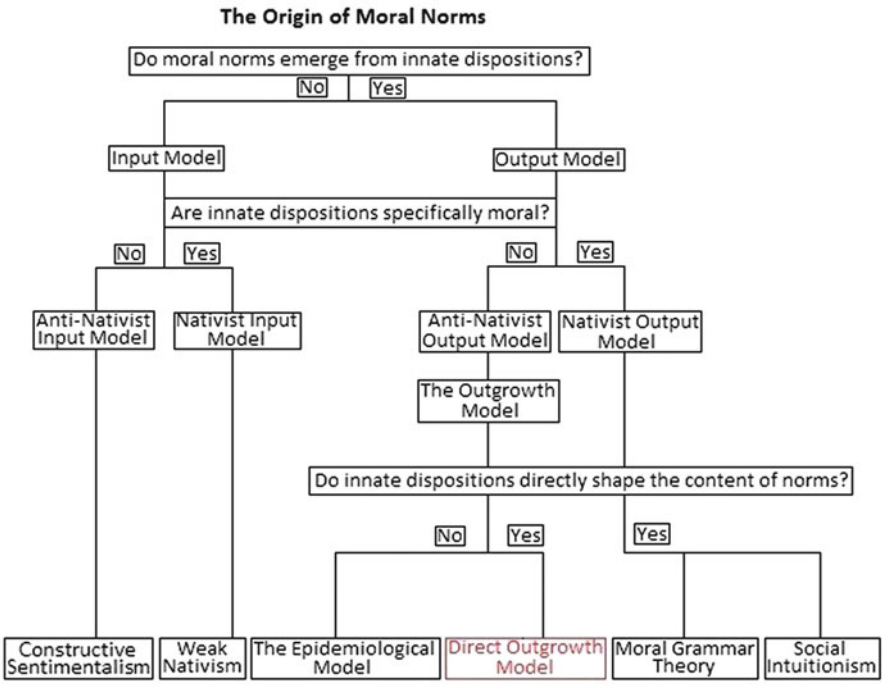

Fig. 10.1 Main questions and theoretical options in the debate on the origin of moral norms

norms, are not arbitrary social constructions but are rather the social extension of psychological traits shared by all human individuals. In this perspective, moral norms more closely resemble *prudential* norms, such as the (implicit) norm discouraging individuals from taking a walk in their underwear at zero degrees Fahrenheit. Such a norm is neither arbitrary nor contingent in that it derives from facts about human individuals, such as human beings' limited resistance to cold weather. One can also argue that human individuals would arrive naturally at such a norm even if it was not taught to them. For this reason, the norm is best described as extending from or emerging from 'human nature' rather than as being arbitrarily created by society and transmitted to individuals through socialization.

These two different kinds of explanations give us two general models for the origin of moral norms, which I will refer to as the Input and the Output models (Fig. 10.1).

In the first model, moral norms are seen as 'inputs' (from the individual's perspective) in that they are external entities transmitted to and assimilated by individuals. In this perspective, culture is the main purveyor of morality: were a given culture to provide radically different moral inputs, an individual from this culture would develop radically different moral judgments, emotions, and behaviours. Therefore, according to the Input model, society (S) creates moral norms (N) which are then assimilated by individuals (I):

Input Model: S → N → I

In the second model, the role of the independent and dependent variables is reversed. Moral norms are seen as 'outputs' (from the individual's perspective) in that they are seen as emerging from individuals – as instantiations of 'human nature' – and consequently imposing themselves on society. The dynamic is therefore a 'bottom-up' rather than a 'top-down' one:

Output Model: I → N → S

Of course, one can legitimately respond that these two models are in no way mutually exclusive. There is a trivial sense in which moral norms are always the product of innate dispositions and culture, of nature and nurture. For instance, a complex moral norm such as 'it is wrong to take money from the poor to give it to the rich' involves concepts that individuals can only acquire through experience. The emergence and transmission of such complex moral norms require not only a cultural context that provides concepts to individuals, but also innate capacities to formulate and grasp such concepts.

Although it is true that moral norms always require 'internal' and 'external' elements to some extent, the input/output dichotomy I am proposing is especially useful when it is applied to *basic* or *elementary* moral principles, which I conceive as the building blocks of more complex moral norms. There are reasons to believe that basic moral principles, such as the principle 'it is wrong to inflict pain on others',[2] are not merely *learned* by individuals but are truly constitutive of their nature – which would fit the Output model. This distinction between basic moral principles and complex moral norms will quickly become essential in the following discussion.

My goal in this article is to examine the Input and the Output models and determine which is the more plausible account of the origin of moral norms.[3] In doing so, I will discuss contemporary versions of each model, taking Jesse Prinz's 'Constructive Sentimentalism' as an example of the Input model (3), and Jonathan Haidt's 'Social Intuitionism' and the 'Moral Grammar' theory as examples of the Output model (4). I will then introduce my positive thesis, the 'Direct Outgrowth' model, which, although primarily an Output theory, captures important dimensions of Prinz's version of the Input model (5).

Before I can begin my presentation of the Input model, however, I need to set the scene by distinguishing three central and closely related debates.

---

[2]I will always presuppose a *pro tanto* – and perhaps *ceteris paribus* – clause behind any such basic moral principle.

[3]To provide a more complete picture of the different theoretical options, one should include the Realist model, according to which moral norms are or result from objective moral properties, often seen as 'supervening' on non-moral properties. I take an agnostic stance on the existence of moral properties in this article, and I will not discuss the Realist model, mainly because I wish to consider only properties whose existence is not controversial.

10.2 Origination, Nativism, and Universalism

The distinction I am proposing between an Input and an Output model may seem like a mere duplication of the more generally accepted distinction between 'nativist' and 'anti-nativist' theories. Indeed, both dichotomies answer the question of whether or not morality 'comes from within', and the new dichotomy I am proposing may seem to bring nothing new. It is therefore important for me to explain how my proposed dichotomy differs from the more common one.

The main differences can be found in the *specific object* addressed by each dichotomy as well as in the *scope* characteristic of each. While the nativism debate deals with whether or not human beings are 'equipped' with morality at birth, and therefore investigates the moral phenomenon at the individual level, the debate on the origin of moral norms discusses whether and/or to what extent moral norms found in societies historically emerge from innate human dispositions, and therefore investigates the moral phenomenon at the population level.

Of course, the position one endorses in the nativism debate very often determines the position that one endorses in the origination debate – and vice versa. One who believes that humans are morally constituted at birth will generally believe that this innate constitution is the root of the moral norms found in societies, while one who believes that morality is not innate will generally believe that moral norms emerge from a different source. However, even though Input and Output models often combine respectively with anti-nativist and nativist theories, different combinations of the two dichotomies are perfectly conceivable, and the theory I will defend in Sect. 10.5 is just such a different combination.

Another debate that should not be confused with the origination debate is the one on the *universality* of moral norms. This distinction is clearly less controversial than the previous one, but it is nevertheless important to understand the specific role that the universalism debate plays in its relation to the origination debate. Essentially, the position one defends in the universalism debate usually serves to provide evidence for the position defended in the other debate. One who highlights the apparent universality of moral norms will often do so in order to support the claim that moral norms have their root in universal human dispositions. Chandra Sekhar Sripada nicely illustrates the kind of data used by advocates of the Output model to provide evidence of their model:

> Most societies have rules that prohibit killing and physical assault (Brown 1991). Most societies have rules promoting sharing, reciprocating, and helping, at least under some circumstances (Cashdan 1989). Most societies have rules regulating sexual behavior among various members of society, especially among adolescents (Bourguignon and Greenbaum 1973), and most societies have rules promoting egalitarianism and social equality (Bohem 1999). (Sripada 2008, p. 322)

The claim of moral universality is also reinforced by experiments in moral psychology which indicate that individuals the world over share similar judgments and intuitions when faced with various moral dilemmas. For instance, Hauser et al. tested the now famous 'trolley problems' on people from different countries and

found that people offer similar answers to different versions of the dilemmas.[4] Even the Hadza, 'a small and remote group of hunter-gatherers living in Tanzania, show similar patterns of responses'(Hauser et al. 2008, p. 135). The apparent universality in moral judgments can be interpreted as supporting the idea that moral norms derive from innate dispositions (Output model), as well as the idea that humans are born 'equipped with' morality (nativist position).

On the other side of the debate, the anthropological record is used to identify exotic tribes or ancient civilizations that promote(d) what we would consider to be highly immoral or barbaric practices. Such examples are usually cited in order to show that moral norms are much more diverse than what the Output model or the nativist position can account for. In response to such anthropological and historical counterexamples, advocates of Output or nativist theories will often distinguish between *foundational* and *content* moral universalism, arguing that counterexamples only serve to refute the latter type of universalism. Counterexamples disprove *content* universalism by showing that moral norms can differ significantly in content from one society to another. Yet, such counterexamples do not disprove the possibility of a 'foundational similarity' between norms: perhaps even the most antagonistic norms ultimately rest on the same fundamental moral principles. Diversity in moral norms could result from combining differently the same basic moral principles, depending on various historical factors. Jonathan Haidt and Craig Joseph defend a form of foundational universalism when they say:

> humans come equipped with an intuitive ethics. . . . These intuitions undergird the moral systems that cultures develop, including their understandings of virtues and character. By recognizing that cultures build incommensurable moralities on top of a foundation of shared intuitions, we can develop new approaches to moral education and to the moral conflicts that divide our diverse society. (Haidt and Joseph 2004, p. 56)

For the rest of this article, I will take for granted that a plausible account of the origin of moral norms needs to accommodate *both* a degree of moral universality and a degree of moral diversity. If one does not endorse 'foundational universalism', one needs to offer an alternative explanation of why individuals and societies share such moral similarities – as described not only in Sripada (2008) and Hauser et al. (2008), but in other influential works such as Brown (1991). Likewise, a plausible theory will be one that recognizes the existence of moral diversity and will not endorse 'content universalism'. It is my hope that these will not seem like arbitrary or unfounded conditions.

With these distinctions and clarifications established, I can now turn to a contemporary version of the Input model: Jesse Prinz's 'Constructive Sentimentalism'.

---

[4]First introduced by Philippa Foot (1978), the 'trolley problems' are thought experiments used to test individuals' intuitions when faced with specific moral dilemmas. Although there have been different versions of the initial problem, all testing different intuitions, Foot's original dilemma was formulated as follows: 'A trolley is running out of control down a track. In its path are five people who have been tied to the track by a mad philosopher. Fortunately, you could flip a switch, which will lead the trolley down a different track to safety. Unfortunately, there is a single person tied to that track. Should you flip the switch or do nothing?' (Foot 1978, p. 20).

10.3 The Input Model

Jesse Prinz is one of the most prominent contemporary advocates of anti-nativism. In *The emotional construction of morals* (2007) and in subsequent papers (Prinz 2008a, b, c, 2009), Prinz cogently argues that moral values are social constructions with an essential foundation in human emotions.[5] His main contention is that whatever innate dispositions humans may have, none of them is *specifically moral*. Even the human disposition for vicarious distress, which is arguably the most likely candidate for a specifically moral disposition, is interpreted by Prinz as serving a primarily non-moral function:

> Doesn't vicarious distress show that we have an innate predisposition to oppose harm? Perhaps, but it's not a moral predisposition. Consider the communicative value of a conspecific's scream. The distress of others alerts us to danger . . . . It's an indication that trouble is near. It's totally unsurprising, then, that we find it stressful. (Prinz 2008b, p. 374)

Prinz has a very specific understanding of the concepts 'specifically moral' and 'innate'. According to Prinz, for a psychological phenotype P to be specifically moral is for it to be innate in the following sense: P is innate if and only if it is 'acquired by means of psychological mechanisms that are dedicated to P, as opposed to psychological mechanisms that evolved for some other purpose or for no purpose at all' (Prinz 2008a, p. 370). There are therefore two elements in Prinz's notion of innateness: a faculty or phenotype is innate only if (1) it is subserved by dedicated machinery – which will often involve specialized modules – and only if (2) it is an evolved adaptation that was directly selected for its fitness-enhancing qualities. For the sake of simplicity and to avoid any confusion, I will always have this definition in mind when using the concept 'innate' in this article.

There is no doubt that Prinz rejects the nativist picture of morality. According to Prinz, our moral sentiments, as well as our capacity to formulate moral judgments, are mere 'by-products' or 'spandrels'[6] of other capacities: 'Morality . . . is a by-product of capacities that were not themselves evolved for the acquisition of moral rules' (Prinz 2007, p. 270). Since no disposition ever evolved for the purpose of morality, morality does not meet Prinz's second criterion for innateness. Nor does it meet Prinz's first criterion; on multiple occasions, Prinz rejects the claim that there are specialized moral modules or any mechanisms specifically dedicated to morality (see for instance section 3.3 in Prinz 2008a).

---

[5] Prinz mainly refers to moral values, but most of his argumentation can apply to moral norms as well.

[6] A spandrel is a characteristic or trait that is not a direct product of adaptive selection, but which is instead a by-product of some other characteristic or trait that was specifically selected. A related concept is that of 'exaptation', which refers more specifically to a 'shift of function' in the process of evolution – such as bird feathers which evolved for the purpose of weather regulation, but which were eventually 'co-opted' for the act of flying. For a better description of these concepts and how they apply to morality, see Fraser (2010).

Although there is no doubt that Prinz endorses moral *anti*-nativism, it should be noted that one cannot *infer* from Prinz's anti-nativism an endorsement of what I call the Input model. This is simply because a disposition like vicarious distress, even if it primarily serves a non-moral function, could still be the source of moral norms found in human societies. If one construes moral norms as the natural extension of innate dispositions, one truly endorses the Output model – even if it turns out that those dispositions do not serve a primarily moral purpose. Parts of Prinz's argumentation seem to support such a view of the origin of moral norms: 'Natural selection has probably furnished us with a variety of behavioral and affective dispositions that contribute to the emergence of moral values . . . . There is a trivial sense in which every norm we have owes something to our biological makeup' (Prinz 2007, p. 255).

There is ample evidence however that Prinz really endorses the Input model. That is because, according to him, none of the innate dispositions favouring morality can ultimately outweigh the 'process of enculturation' (Prinz 2007, p. 257) or socialization. Culture is thus the real force to be reckoned with, and there is arguably no limit to what a society can come to endorse as a rule of conduct for its members. What is condemned in society A can very well be revered in society B:

> I tend to think, somewhat cynically, that the range of moral rules is relatively unconstrained . . . . I adamantly believe that we could teach people to value recreational torture of babies . . . . I'm sure a search of the anthropological record would uncover groups that tortured babies for fun – especially if the babies belonged to enemy groups that were defeated in battle. (Prinz 2008c, p. 429)

Despite his failure at finding groups that torture babies for fun, Prinz does provide examples of remote tribes that perpetuate shocking traditions, such as the Llongot tribe in the Philippines whose coming-of-age ritual for boys consists in bringing back the head of an innocent member of a neighbouring tribe (Rosaldo 1980, from Prinz 2008b, p. 373). Such examples clearly serve the purpose of showing that whatever innate dispositions human beings may have, none of them is strong enough to outweigh the pressures of socialization. And precisely because innate dispositions are *weak* or *non-pervasive* in that way, they cannot accurately be described as the source of moral norms. Therefore, the Output model is not the right model of the origin of moral norms.

However, as was discussed in Sect. 10.2, anthropological counterexamples cannot provide sufficient evidence against the Output model, because such counterexamples do not disprove the theory of 'foundational universalism', i.e. the claim that even radically different moral norms rest on a foundation of universally shared basic moral principles. For example, one could argue that the Llongots recognize the *pro tanto* wrongness of murder, but that this *pro tanto* wrongness is outweighed by metaphysical beliefs held by the tribe, such as the belief that neighbouring tribes are evil or somehow inhuman, or that such murders are necessary for the tribe's survival. Such beliefs could be seen by them as legitimizing an otherwise condemnable act. An advocate of the Output model could go on to argue that the recognition of the *pro tanto* wrongness of murder is a 'natural extension' of human dispositions rather than a social construction, and that the case of the Llongots only serves to illustrate that

the same basic moral principles can find different expressions in different cultural contexts. Therefore, in order to provide a satisfactory defence of the Input model, Prinz would need to explain why basic moral principles, such as the *pro tanto* wrongness of murder, enjoy such universality.

Such an explanation is provided by Prinz. When considering the apparent universality of basic moral principles, he offers an explanation inspired in large part by game theory:

> There are some social pressures that all human beings face. In living together, we need to devise rules of conduct . . . . Cultures need to make sure that people feel badly about harming members of the in-group and taking possessions from their neighbors . . . . This is a universal problem, and given our psychological capacities (for emotion, reciprocation, mental state attribution, etc.), there is also a universal solution. (Prinz 2008c, p. 405)

According to Prinz, this 'universal solution' to this 'universal problem' is the main force that constrains the otherwise limitless range of moral norms. It is because there is a universal solution to a universal problem of coordination that there is such universality in basic moral principles. The fact that similar principles are held cross-culturally is therefore not so much the result of our sharing similar dispositions as it is the result of our facing similar problems and coming up with similar solutions.

Prinz's argument about the importance of coordination pressures in understanding the origin of moral norms is in itself rather uncontroversial. Hardly anyone will deny that groups have their own dynamics, from which certain imperatives naturally arise, and that these dynamics can play a role in the emergence and prevalence of moral norms in human societies. The problem with Prinz's thesis however is that he implies that coordination pressures are *sufficient* for explaining the universality of basic moral principles. It is not at all clear that coordination pressures *alone* can account for the extent of moral similarities across cultures. There are also good reasons to doubt that innate dispositions are as 'weak' as Prinz thinks they are, or that socialization and conditioning mechanisms are such irresistible forces. As will be discussed in the next section, many dispositions and 'prepared emotional reactions' actually seem quite robust and can become serious obstacles to unusual socialization projects.

If it turns out that innate dispositions are indeed robust, they may very well be an essential factor, alongside other factors such as coordination pressures, in explaining the origin of moral norms. What remains to be determined however is the *specific nature* of the relevant dispositions as well as the *specific role* that they play in bringing about moral norms. Different conceptions of the nature and role of dispositions will give us different versions of the Output model.

10.4 The Output Model

The Output model can take many different forms, and the innate dispositions construed as the sources of moral norms can range from a general 'learning preparedness' to a full-fledged moral sense. I will focus here on two contemporary theories

which distinguish themselves in the amount of attention they receive from researchers. It should be noted that, just as with Jesse Prinz's Constructive Sentimentalism, the two theories I will discuss here are primarily associated with the *nativism* debate, and their current application to the *origination* debate involves an element of interpretation on my part. It is perfectly possible therefore that some advocates of each theory would be more sympathetic to the Input model, and I will briefly discuss in Sect. 10.5 how such a nativist version of the Input model is conceivable.

The two theories I wish to discuss are the 'Moral Grammar' theory, whose main advocates are Marc Hauser (2006, 2008), Susan Dwyer (2008), and John Mikhail (2008); and the theory of 'Social Intuitionism', developed and mainly defended by Jonathan Haidt (Haidt and Joseph 2004; Haidt and Bjorklund 2008). The two theories differ significantly in their understanding of the *nature* of the relevant innate dispositions, and I will focus in this section on their diverse conceptions of the nature of innate dispositions. In the next section, I will address the question of the *role* that innate dispositions play, according to the Output model, in bringing about moral norms in human societies.

I begin with the Moral Grammar theory (MG). Often referred to as the 'linguistic analogy', because of its roots in Noam Chomsky's linguistic model, MG describes the relevant innate disposition as a specialized moral faculty or competence. This faculty or competence is construed as an innate 'grammar', i.e. a set of abstract, general principles which unconsciously guide the individual's interpretation of social phenomena and facilitate the acquisition of a moral 'language' or system. This is certainly an unusual way of conceiving moral dispositions, and it is therefore important to understand how advocates of MG arrive at such a conception.

First, they note that a rule or principle such as 'it is wrong to inflict pain on others' appears to be 'endorsed' remarkably early by young children, long before socialization is able to leave its full imprint on them.[7] This is just one instance of the general phenomenon of the *precociousness* of human morality, of which John Mikhail offers an interesting overview which is worth presenting at length:

> Three- and four-year-old children use intent or purpose to distinguish two acts with same result (Baird 2001). They also distinguish 'genuine' moral violations (e.g. theft, battery) from violations of social conventions (e.g. wearing pajamas to school; Smetana 1983; Turiel 1983). Four- and five-year-olds use a proportionality principle to determine the appropriate level of punishment for principals and accessories (Finkel et al. 1997). Five-year-olds display a nuanced understanding of negligence and restitution (Shultz et al. 1986). One man shoots and kills his victim on the mistaken belief that he is aiming at a tree stump. A second man shoots and kills his victim on the mistaken belief that killing is not wrong. Five- and six-year-olds distinguish cases like these in conformity with the distinction between mistake of law and mistake of fact, recognizing that false actual beliefs may exculpate, but false moral beliefs do not (Chandler et al. 2000). Five- and six-year-olds also calibrate the level of punishment they assign to harmful acts on the basis of mitigating factors, such as provocation, necessity, and public duty (Darley et al. 1978). Six- and seven-years-olds exhibit a keen sense of procedural fairness, reacting negatively when punishment is inflicted without affording the parties notice and the right to be heard (Gold et al. 1984). (Mikhail 2008, p. 354)

[7] If not intellectually, at least *practically*, i.e. as reflected by their behaviours and actions.

And the list goes on. Using this list of claimed premature moral 'knowledge' as evidence, Mikhail, like other advocates of MG, argues that children must be born with a set of moral 'principles and parameters'. Without such innate principles and parameters, children would develop a moral system only through the stimuli received from their environment. The problem, argue MG advocates, is that 'moral stimuli' are so *poor* that they could hardly account for the level of moral knowledge possessed by children. *Ergo*, there has to be something resembling an innate moral 'grammar' guiding their experience – in the same way that an innate grammar is said to guide the acquisition of a natural language in the Chomskian linguistic paradigm.

The problem with this line of argument, however, as expressed by multiple authors, is that the moral stimulus appears 'poor' only if it is conceived entirely in terms of rational rules or principles that are expressed by parents or 'imbibed' by children through experience. But, as Kim Sterelny notes, 'children get more than verbal feedback. Audiences respond [to moral infractions] with emotional changes, and humans respond emotionally to their very own actions and to the effects of those actions on others' (Sterelny 2010, p. 293). Once emotional responses are considered to be moral stimuli, the claim regarding their poverty suddenly appears quite . . . poor.

The point here is not so much to reject the claim of moral precociousness in children as to reject MG's interpretation of the phenomenon. The fact that children exhibit moral behaviour and have a certain understanding of morality at a young age is generally recognized. For instance, Shaun Nichols, who is not himself an advocate of MG, comes to similar conclusions after reviewing a series of studies, noting that 'children have a strikingly early grasp of core moral judgment' (Nichols 2008, p. 261). He observes that some moral 'facts' grasped by children, such as the moral/conventional distinction, do seem quite precocious: 'These findings on the moral/conventional distinction are neither fragile nor superficial. They have been replicated numerous times using a wide variety of stimuli.' The point therefore is not to deny the phenomenon of 'early morality' but to understand it in light of the central role played by *emotions.*

One way of understanding the role of emotions is in terms of emotional *reinforcement*. This is the strategy adopted by Jesse Prinz to account for the phenomenon of early morality. Reinforcement strategies such as 'love-withdrawal', 'power assertion', and 'induction of empathic distress' (Prinz 2008b, p. 431) are all used by parents and other moral educators to bring children to assimilate their society's norms. The advantage of such an explanation is of course its great parsimony. The theory requires nothing more than a general responsiveness of children to conditioning to account for their early moral behaviours and judgments. The disadvantage of the explanation, however, is that it presupposes something that greatly resembles 'equipotentiality', a theory that is largely discredited nowadays in psychology: 'Garcia and Koelling (1966) demonstrated that equipotentiality – the equal ability of any response to get hooked up at any stimulus – was simply not true. It is now universally accepted in psychology that some things are easy to learn

(e.g. fearing snakes), while others (fearing flowers or hating fairness) are difficult or impossible' (Haidt and Bjorklund 2008, p. 183).

The implication of a rejection of equipotentiality is that children are not as malleable as is sometimes claimed, and they will often resist attempts at emotional reinforcement:

> Children routinely resist parental efforts to get them to care about, value, or desire things. It is just not very easy to shape children, unless one is going with the flow of what they already like. It takes little or no work to get 8-year-old children to prefer candy to broccoli, to prefer being liked by their peers to being approved of by adults, or to prefer hitting back to loving their enemies. Socializing the reverse preferences would be difficult or impossible. (Haidt and Bjorklund 2008, p. 201)

This leads to a second way of understanding the role of emotions to account for the phenomenon of early morality, which is developed by Jonathan Haidt in the theory of Social Intuitionism. In addition to emotional reinforcement, one needs to take into consideration the emotional *preparedness* of children. According to Haidt, children are born equipped with 'an innate preparedness to feel flashes of approval or disapproval toward certain patterns of events involving other human beings' (Haidt and Joseph 2004, p. 56). More specifically, Haidt argues that these prepared emotional reactions correspond to five basic moral domains: harm/care, fairness/reciprocity, in-group/loyalty, authority/respect, and purity/sanctity. He essentially arrives at such a conclusion by reviewing important works detailing, among other things, commonalities in moral judgments across cultures as well as animal precursors of morality. According to Haidt, these five moral domains are the best way of capturing what is found in the literature, and he postulates the existence of a specialized module for each of these domains.

It is not clear however how Haidt arrives at the conclusion that one should see these five fundamental moral domains as being encoded in the human mind. The simple fact that we can successfully subsume moral phenomena into five categories is no proof that these categories are present in the brain at birth. As Ron Mallon puts it, referring to 'principles' rather than 'domains': 'The mere fact that we can describe principles that seem to capture intuitions about a set of moral cases gives us exactly no reason at all to think that those principles are themselves implemented directly in a computationally discrete way or by a computationally discrete faculty' (Mallon 2008, p. 151). The domains described by Haidt may very well turn out to be universal and the source of moral norms found cross-culturally, but these facts alone can hardly be seen as evidence, let alone proof, of the existence of five specialized modules.

It should be noted that the same criticism applies to many advocates of the Moral Grammar theory as well. The main weakness of theories that postulate the existence of specialized moral modules, especially in the Fodorian tradition, is that they are cognitively costly, which renders them dubious in the eyes of their anti-nativist critics. Why should one adopt a modular picture of morality when a less costly alternative is available? The answer offered by many nativists is that only this kind

of nativist framework can efficiently account for the universality of basic moral norms and judgments.[8]

My personal contention, shared by anti-nativists, is that one does not need such a nativist framework to account for moral universality. Unlike most anti-nativists, however, I share the nativist's essential intuition, which is that moral universality speaks in favour of a conception of morality as emerging *from within* human individuals. I do not believe however that specialized moral modules or faculties are necessary to explain how morality comes from within, and for this reason I have introduced a distinction between nativist theories and the Output model. The Output model can provide an explanation of moral norms as being rooted in human constitution even in the absence of specialized modules or faculties. It is such an anti-nativist version of the Output model that I now wish to defend.

10.5 The Outgrowth Model, or Output Non-nativism

I have not yet provided a full picture of the different theoretical options in the debate on the origin of moral norms. Thus far, I have only presented one version of the Input and the Output model, namely, the anti-nativist version of the Input model and the nativist version of the Output model. Two theoretical options remain to be addressed.

First, one could defend a nativist version of the Input model. A theory of this type would essentially argue that humans are morally constituted at birth, *but* that moral dispositions are highly malleable and hold very little if any weight in shaping the content of a society's norms. One could call such a theory 'Weak Nativism', where 'weak' refers to the non-pervasive nature of innate moral dispositions. Because moral dispositions are non-pervasive, they cannot accurately be described as the source of moral norms found in human societies. Whether moral norms are seen as resulting mainly from contingent factors (geographical, political, religious, etc.), or from necessary ones (such as Prinz's coordination pressures), the relevant factors will be *external* to the individual's moral constitution.

Second, one could defend an anti-nativist version of the Output model – which is the kind of theory I endorse. The main idea is that moral norms have an essential source in innate dispositions but that none of these dispositions is specifically moral in nature. Not being specifically moral implies that these dispositions did not evolve for the purpose of morality and are not subserved by dedicated machinery – such as specialized modules. Yet, despite their not being specifically moral in such a

[8]Of course, not all moral nativists are committed to the existence of specialized moral modules. One can distinguish between 'strong' and 'weak' kinds of nativism, noting that only the former kind is committed to modules. Jesse Prinz offers a similar distinction between three kinds of nativism which he labels 'immodest', 'modest', and 'minimal' nativisms (Prinz 2009, p. 168). I am focusing here on 'strong nativism' simply because it is the type endorsed by most advocates of the Moral Grammar theory and Social Intuitionism.

Table 10.1 Four theoretical options for the origin of moral norms

	Nativism	Anti-nativism
Input model	Weak nativism	Constructive sentimentalism
Output model	Moral grammar theory; social intuitionism	The outgrowth model

way, these various dispositions are nevertheless the main cause of the development of moral norms in human societies. Morality should therefore be construed as an 'outgrowth' of those dispositions (Table 10.1).[9]

Moral norms can be seen as outgrowths of innate dispositions in at least two different ways, which I will refer to as the 'direct' and the 'indirect' versions of the Outgrowth model.

According to the *indirect* version of the Outgrowth model, moral norms grow indirectly out of innate dispositions in the sense that innate dispositions indirectly shape the content of norms. To indirectly shape norms is to play a *constraining* role in the determination of norms; when a norm is 'introduced' in a given society, innate dispositions are the main factor that determines whether or not the 'candidate norm' will successfully impose itself, i.e. be adopted by society and persist through time. Something like a Darwinian mechanism is in play which allows for only the *fittest* candidate norms to pass the test of time – and for a candidate norm to be *fit* it must be compatible with humans' natural constitution.

To illustrate this idea, one can take the example of cooperation norms. Imagine a society that would try to promote cheating, betrayal, and free-riding as moral ideals. Such a project would certainly be short-lived, and one obvious reason is the one raised by Jesse Prinz regarding coordination pressures: a society or any collective enterprise will sooner or later collapse if its members are unable to cooperate and trust each other. However, coordination pressures are only one part of the explanation. Indeed, if humans were simply unreceptive to cooperation imperatives and were inclined only to cheat, betray, and free-ride, cooperation norms could not successfully impose themselves, and society would simply collapse.[10] Imagine for instance a society composed entirely of psychopaths. Given psychopaths' natural selfishness and lack of empathy,[11] one can doubt that they would be able to

[9] It should be remembered that this kind of theory is described as 'anti-nativist' only insofar as it rests on Prinz's specific use of the concept 'innate'. If one were to adopt a different definition of the concept – for instance, if one were to say that a faculty can be called 'innate' even if it is a by-product or spandrel of other faculties – one could very well consider this kind of theory to be 'nativist'. This is why, using a different definition of 'innate', I presented this theory in a different article (Giroux 2011) as a form of 'moderate nativism' rather than as a form of anti-nativism.

[10] One could argue that social institutions could still be preserved if individuals were strongly constrained by external forces, such as in a police state. However, in this scenario, the individuals in charge of enforcing cooperation would themselves only be serving their personal interest. Therefore, in the absence of a genuine capacity for cooperation, society can only rest on very shaky grounds.

[11] See Blair et al. (2005) for a thorough description of psychopaths' unusual constitution.

truly assimilate cooperation norms; as a result, their society's institutions would eventually collapse, leaving them with something resembling a Hobbesian state of nature. This is but one example of how human dispositions render certain norms very likely, while others are rendered highly improbable, if not impossible.

Theorists who construe innate dispositions as 'indirectly shaping' the content of moral norms in this way are usually inspired by Dan Sperber's 'epidemiology of representations' (Sperber and Hirschfeld 2004). The main philosophers who adopt the 'Epidemiological model' are Shaun Nichols (2004, 2008), Chandra Sekhar Sripada (2008), and Steven Stich (2006).[12] Nichols defends a version of the Epidemiological model called 'affective resonance', which focuses on the constraining role of *emotional* dispositions: 'The affective resonance hypothesis predicts that, *ceteris paribus*, norms that prohibit actions that are independently likely to excite negative emotions should be more likely to survive than norms that are not connected to emotions' (Nichols 2008, p. 270). Sripada incorporates a wider range of dispositions, which he calls 'Sperberian biases': 'When their effects are summated over populations and over time, they generate a fairly strong population-level force which can have the effect of changing the distribution of norms in the direction favored by the Sperberian bias' (Sripada 2008, p. 333).

The idea that innate dispositions constrain the range of possible moral norms is a truly elegant explanation of why one finds so many similarities in the moral norms adopted by otherwise very different societies. I do not believe, however, that this indirect version of the Outgrowth model provides a complete account of the origin of moral norms. My claim is that innate dispositions play an even stronger role in shaping the content of moral norms: they provide the elementary moral principles, or 'building blocks' used by all societies in the creation of more complex moral norms. Because they actually *provide moral content*, as opposed to merely imposing general constraints, innate dispositions should be seen as *directly* shaping the content of moral norms.

According to the direct version of the Outgrowth model, innate dispositions directly shape the content of moral norms by helping every 'normally constituted' human individual to develop naturally the same basic moral principles.[13] Those basic principles provide the general structure on which human societies develop moral norms. Societies will diverge by giving more importance to certain basic principles rather than others, and by identifying different criteria for their application, but they will still incorporate a similar set of basic moral principles. Only exceptionally strong factors, such as extreme metaphysical beliefs, could potentially lead societies

[12] Jesse Prinz also defends a version of the epidemiological model, arguing that 'cultural transmission is a function of fitness' (Prinz 2007, p. 220). However, as was described in Sect. 10.3, Prinz does not assign a real constraining role to innate dispositions, and for that reason I did not include him as an advocate of the 'indirect Outgrowth' model: 'Biologically based behaviors are not quite a constraint on the genealogy of moral rules, because culture can override them, but they are often a central ingredient' (Prinz 2007, p. 274).

[13] Again, with the exception of individuals such as psychopaths.

to not incorporate some of the basic moral principles.[14] This view amounts to an endorsement of what was dubbed 'foundational universalism' in Sect. 10.2.

At this point, one may legitimately ask what those basic moral principles actually are. Since the present article only aims at clarifying the main theoretical options in the debate on the origin of moral norms, this is certainly not the place to defend an exhaustive list of basic moral principles. However, as a general indicator of what I have in mind, one can refer to W.D. Ross's list of *prima facie* moral duties, which includes principles of fidelity, reparation, gratitude, non-maleficence, justice, beneficence, and self-improvement (Ross 1930). I believe that individuals are naturally led to develop similar basic principles for all or most of Ross's domains. My specific contention is that individuals are neither 'born with' those principles nor do they merely 'internalize' them as a result of socialization. Rather, such principles naturally derive from human predispositions, even in the absence of specialized moral modules.

Putting aside pressures of socialization, at least two factors can be seen as leading individuals to develop similar basic moral principles. The first factor is what Shaun Nichols calls 'natural elicitors': certain events naturally elicit emotional reactions in individuals independently of culture. A good example is vicarious distress: 'even newborn infants respond aversely to some cues of suffering (e.g. Simner 1971) . . . . There is good reason to suppose that the emotional response to suffering in others is universal and innately specified' (Nichols 2008, p. 271). Innately prepared vicarious distress, even if it did not evolve for the purpose of morality, could lead individuals to develop intuitions about the 'negativity' of suffering in others. Other 'natural elicitors' could have a similar impact.

Second, humans naturally project on others the many rules and principles that they come to perceive as applying to themselves: being treated unfairly makes me feel outraged, therefore it is something that will make others (who are like me) feel outraged, and it is therefore something I have a *pro tanto* reason to avoid. Replace 'being treated unfairly' with 'being lied to' or 'having a promise made to me broken', and individuals can be brought to see that if it is true for themselves, it will usually be true for others, for the simple reason that others are *like them.* This is what Erik J. Wielenberg calls the 'likeness principle': 'If I believe that I am a bearer of certain moral barriers and that others are similar to me with respect to their known properties, I am disposed to form the belief that those others possess similar moral barriers' (Wielenberg 2010, p. 446). Steven Pinker defends more or less the same principle: 'No creature equipped with the circuitry to understand that it is immoral for you to hurt me could discover anything but that it is immoral for me to hurt you' (Pinker 2002, p. 193).

[14]It is quite hard to find actual examples of societies that do not incorporate the basic moral principles. An example that may come to mind is the practice of human sacrifice that was condoned by certain ancient religious traditions. Perhaps these killings were viewed by some societies as *wholly positive* (because they pleased the gods), but it is more likely that they were viewed as *necessary evils* that retained an element of *prima facie* wrongness which was simply outweighed by a greater good. For an example to be truly accurate here, it would have to show that a society does not even consider killing (or lying, not keeping a promise, etc.) as a *wrong-making feature* of an act.

The existence of natural elicitors and the likeness principle are only two examples of how individuals can develop similar basic moral principles even in the absence of moral modules. Of course, if it turns out that such modules actually exist, there will be an even stronger case for viewing dispositions as *directly* shaping norms.[15] My goal in this section has been simply to show that moral nativism, understood in a *Prinzian* fashion, is not a prerequisite for construing moral norms as emerging from within individuals.

10.6 Conclusion

With all the pieces in place, we have a preliminary picture of what could be a new theory of the origin of moral norms. According to the Direct Outgrowth model that I have presented, moral norms should be seen as the natural extension of human dispositions rather than as social constructions. In this perspective, moral norms are directly shaped by humans' emotional preparedness and by a variety of other innate dispositions. These innate dispositions are not specifically moral, which means that they did not evolve for the purpose of morality and that they are not subserved by dedicated machinery. Still, they lead every 'normally constituted' human individual to develop naturally certain basic moral principles, such as the seven principles described by W.D. Ross. These basic moral principles constitute the normative backbone of every human society, and moral diversity should be understood as the assignment of different weights to the same basic principles.

Of course, this model can and should incorporate other important factors outlined by other models. For instance, one should see 'coordination pressures' as emphasizing certain imperatives which are essential to any collective enterprise, and one should recognize the role of emotional reinforcement, alongside emotional preparedness, as a key factor accounting for the phenomenon of early morality. With the inclusion of these different elements, one arrives at a potentially more complete picture of the origin of moral norms.

A lot more needs to be said in defence of this model, however, before it can claim to be more plausible than alternative models. The model's main challenges will be to demonstrate that innate dispositions are as robust as it claims, and that basic moral principles that individuals naturally develop are truly pervasive. Those are challenges that I hope to address in the future.

Acknowledgments I am very grateful to Christine Tappolet, Benoît Dubreuil, Bruno Guindon, Jean-François Grégoire, and Joseph Owens for their helpful comments on this article. I also wish to thank participants at the international conference "What Makes Us Moral?" in Amsterdam, Netherlands, on June 23–24, 2011, for their comments on my presentation. I am especially grateful to Carel van Schaik and Carsten Nielsen for their comments.

[15]Output nativists generally seem to endorse the view that innate dispositions *directly* shape the content of a society's norms (Haidt and Joseph 2004, p. 56; Dwyer 2008, p. 414).

References

Baird, J.A. 2001. Motivations and morality: Do children use mental state information to evaluate identical actions differently? Paper presented to the biennial meeting of the Society for research in child development, Minneapolis.

Blair, J., D. Mitchel, and K. Blair. 2005. *The psychopath: Emotion and the brain*. Malden: Wiley-Blackwell.

Bohem, C. 1999. *Hierarchy in the forest*. Cambridge: Harvard University Press.

Bourguignon, E., and L. Greenbaum. 1973. *Diversity and homogeneity in world societies*. New Haven: HRAF Press.

Brown, D. 1991. *Human universals*. New York: McGraw-Hill.

Cashdan, E. 1989. Hunters and gatherers: Economic behavior in bands. In *Economic anthropology*, ed. S. Plattner. Palo Alto: Stanford University Press.

Chandler, M., B. Sokol, and C. Wainryb. 2000. Beliefs about truth and beliefs about rightness. *Child Development* 71(1): 91–97.

Darley, J., E. Klosson, and M. Zanna. 1978. Intentions and their contexts in the moral judgments of children and adults. *Child Development* 49: 66–74.

Dwyer, S. 2008. How not to argue that morality isn't innate: Comments on Prinz. In *Moral psychology*, vol. 2, ed. W. Sinnott-Armstrong, 407–418. Boston: MIT Press.

Finkel, N., M. Liss, and V. Moran. 1997. Equal or proportionate justice for accessories? Children's pearls of proportionate wisdom. *Journal of Applied Developmental Psychology* 18: 229–244.

Foot, P. 1978. *Virtues and vices and other essays in moral philosophy*. Berkeley/Los Angeles: University of California Press.

Fraser, B.J. 2010. Adaptation, exaptation, by-products, and spandrels in evolutionary explanations of morality. *Biological Theory* 5(3): 223–227.

Garcia, J., and R. Koelling. 1966. Relation of cue to consequence in avoidance learning. *Psychonomic Science* 4: 123–124.

Giroux, J. 2011. The origin of moral norms: A moderate nativist account. *Dialogue: Canadian Philosophical Review* 50(2): 381–406.

Gold, L., J. Darley, J. Hilton, and M. Zanna. 1984. Children's perceptions of procedural justice. *Child Development* 55: 1752–1759.

Haidt, J., and F. Bjorklund. 2008. Social intuitionists answer six questions about moral psychology. In *Moral psychology*, vol. 2, ed. W. Sinnott-Armstrong, 181–217. Boston: MIT Press.

Haidt, J., and C. Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus: On Human Nature* 133: 55–66.

Hauser, M.D. 2006. *Moral minds: How nature designed our universal sense of right and wrong*. New York: Ecco Press.

Hauser, M.D., L. Young, and F. Cushman. 2008. Reviving Rawls's linguistic analogy: Operative principles and the causal structure of moral actions. In *Moral psychology*, vol. 2, ed. W. Sinnott-Armstrong, 107–143. Boston: MIT Press.

Mallon, R. 2008. Reviving Rawls's linguistic analogy inside and out. In *Moral psychology*, vol. 2, ed. W. Sinnott-Armstrong, 145–155. Boston: MIT Press.

Mikhail, J. 2008. The poverty of the moral stimulus. In *Moral psychology*, vol. 1, ed. W. Sinnott-Armstrong, 353–359. Boston: MIT Press.

Nichols, S. 2004. *Sentimental rules: On the natural foundations of moral judgment*. Oxford/New York: Oxford University Press.

Nichols, S. 2008. Sentimentalism naturalized. In *Moral psychology*, vol. 2, ed. W. Sinnott-Armstrong, 255–274. Boston: MIT Press.

Pinker, S. 2002. *The blank slate: The modern denial of human nature*. New York: Viking Penguin.

Prinz, J.J. 2007. *The emotional construction of morals*. Oxford: Oxford University Press.

Prinz, J.J. 2008a. Is morality innate? In *Moral psychology*, vol. 1, ed. W. Sinnott-Armstrong, 367–406. Boston: MIT Press.

Prinz, J.J. 2008b. Reply to Dwyer and Tiberius. In *Moral psychology*, vol. 1, ed. W. Sinnott-Armstrong, 427–439. Boston: MIT Press.
Prinz, J.J. 2008c. Resisting the linguistic analogy: A commentary on Hauser, Young, and Cushman. In *Moral psychology*, vol. 2, ed. W. Sinnott-Armstrong, 157–170. Boston: MIT Press.
Prinz, J.J. 2009. Against moral nativism. In *Stich and his critics*, ed. M. Bishop and D. Murphy, 167–189. Malden: Miley-Blackwell.
Ross, W.D. 1930. *The right and the good*. Oxford: Oxford University Press.
Shultz, T., K. Wright, and M. Schleifer. 1986. Assignment of moral responsibility and punishment. *Child Development* 57: 177–184.
Simner, M. 1971. Newborn's response to the cry of another infant. *Developmental Psychology* 5: 136–150.
Smetana, J.G. 1983. Social cognitive development: Domain distinctions and coordinations. *Development Review* 52: 1333–1336.
Sperber, D., and L.A. Hirschfeld. 2004. The cognitive foundation of cultural stability and diversity. *Trends in Cognitive Sciences* 8: 40–46.
Sripada-Sekhar, C. 2008. Nativism and moral psychology: Three models of the innate structure that shapes the contents of moral norms. In *Moral psychology*, vol. 2, ed. W. Sinnott-Armstrong, 319–343. Boston: MIT Press.
Sterelny, K. 2010. Moral nativism: A skeptical response. *Mind and Language* 25(3): 279–297.
Stich, S. 2006. Is morality an elegant machine or a kludge? *Journal of Cognition and Culture* 6: 181–189.
Turiel, E. 1983. *The development of social knowledge: Morality and convention*. New York: Cambridge University Press.
Wielenberg, E.T. 2010. On the evolutionary debunking of morality. *Ethics* 120(3): 441–464.