*Revista de Humanidades de Valparaíso*, 2022, No 19, 163-179 *Humanities Journal of Valparaíso*
DOI: https://doi.org/10.22370/rhv2022iss19pp163-179 eISSN 0719-4242
Sección Monográfica / Monographic Section

# *Under-* and Overspecification in Moral Foundation Theory. The Problematic Search for a Moderate Version of Innatism

## Sub- *y sobreespecificación en la Teoría de los Fundamentos Morales. La problemática búsqueda de una versión moderada de innatismo*

Rodrigo Braicovich
Universidad Nacional de Rosario, Argentina
rbraicovich@gmail.com

**Abstract**

Jonathan Haidt's *Moral Foundation Theory* has been criticized on many fronts, mainly on account of its lack of evidence concerning the genetic and neurological bases of the evolved moral intuitions that the theory posits. Despite the fact that Haidt's theory is probably the most promising framework from which to integrate the different lines of interdisciplinary research that deal with the evolutionary foundations of moral psychology, *i)* it also shows a critical underspecification concerning the precise mental processes that instantiate the triggering of our evolved moral intuitions, and that *ii)* that underspecification coexists with and overspecification of the structure of human nature when it comes to exploring alternatives to capitalist societies.

**Keywords**: human nature, evolution, social intuitionism, philosophy of mind, intuition, political philosophy.

**Resumen**

La Teoría de los Fundamentos Morales de Jonathan Haidt ha sido blanco de numerosas críticas, fundamentalmente en función de la falta de evidencia vinculada con las bases neurológicas y genéticas de las intuiciones morales evolutivas que dicha teoría propone. A pesar de que la teoría de Haidt aparece hoy en día como el marco teórico probablemente más prometedor al momento de integrar los resultados provenientes de las distintas líneas de investigación interdisciplinaria abocadas al estudio de los fundamentos evolutivos de la psicología moral, argumentaré que *i)* también evidencia

Received: 20/10/2021. Final version: 29/03/2022
eISSN 0719-4242 – © 2022 Instituto de Filosofía, Universidad de Valparaíso

una subespecificación crítica en cuanto a los procesos mentales específicos que operan en la activación de nuestras intuiciones moral evolutivas, y que *ii)* dicha subespecificación coexiste con una sobreespecificación de la estructura de la naturaleza humana al momento de explorar alternativas a las sociedades capitalistas.

**Palabras clave**: naturaleza humana, evolución, intuicionismo social, filosofía de la mente, intuición, filosofía política.

## 1. Introduction

Moral Foundations Theory (henceforth MFT) is probably one of the most ambitious theories developed in the last two decades in the complex and multidisciplinary domain of Moral Psychology, and it has become the framework from which several lines of inquiry concerning the innate basis of many of our moral intuitions have been developed. One of the reasons that explains its growing popularity among researchers is the fact that, rather than pretending to be a rigid investigation program, that came into existence as a spontaneously coherent and perfected theory, MFT has always been explicitly conscious of its complex past and its open future[1]. Although many of the insights that would later become an integral part of MFT were already present in Jonathan Haidt's celebrated 2001 paper (such as the criticism of cognitivism in the explanation of our mental life, the role of *post hoc* rationalizations in moral reasoning, or the idea that the domain of human morality exceeds that of harm), it took at least a decade for MFT to find the nearly definitive form it reached in, for example, Haidt (2012) and Graham *et al*. (2013). Time, however, was not the only input that MFT needed to become the far reaching theory that it is nowadays: it also benefited from the joint labor between Haidt and a great number of researchers, such as G.L. Clore, J. Graham, R. Iyer, A.H. Jordan, C. Joseph, D. Keltner, S. Kesebir, S. Koleva, J.P. Morris, M. Motyl, B.A. Nosek, S.E. Rimm-Kaufman, P. Rozin, S. Schnall, and S.P. Wojcik, which is why I will henceforth refer to MFT as being the brainchild of H+, rather than of Jonathan Haidt alone.

Although MFT needs not much of a presentation, since it has been one of the most cited theories in the last decade and both H+ and their detractors have produced excellent and constantly updated summaries of the theory[2], a minimal reconstruction of the particular aspects I am interested in discussing should include the following premises:

- Moral evaluations are generally the result of intuitions, rather than deliberation or conscious reasoning.

[1] The structure of intellectual autobiography that Haidt chose for what can be considered the most elaborate version of MFT, *The righteous mind* (2012), is a clear expression of the dynamic and flexible nature of the theory.

[2] Graham et al. (2013) provides a particularly useful account of the theory, the criticisms it has received and the corresponding replies by H+.

- A subset of those intuitions are what we can call 'evolved moral intuitions', which are innate in human beings and which have an evolutionary rationale.
- The set of evolved moral intuitions that natural selection endowed us with as a species is linked to adaptive problems our ancestors encountered on a regular basis and which determined our evolutionary success.
- Each evolved moral intuition can be activated by its original trigger or by other (current) triggers that have come to be associated with the original one.
- Although evolved moral intuitions constitute what we could call human nature, they actually produce the 'first draft' of the moral mind, which can be rewritten by culture and lead to historical variations.

As can be hinted at from this set of premises, MFT aims to provide a framework that serves two simultaneous purposes: to explain the immense *diversity* that human history shows us concerning moral matters while accounting for the *regularities* which that same history provides us. It aims, in other words, to provide a moderate version of the innateness thesis that can rid itself of the problems that plagued the strong versions of the thesis. In the following sections, however, I will try to show that *i)* MFT has a particularly rough time navigating the middle waters that separate the shores of strong innatism and those of radical culturalism, and that *ii)* this difficulty can be attributed to the (deliberate) ambiguity and vagueness that characterizes MFT when providing specific definitions concerning how evolved moral intuitions operate and how original and current triggers are related. As a way of showing the kind of problems that come with the underspecification of the innateness basis of MFT, I will turn to H+'s resort to the old thesis of the incompatibility between human nature and the construction of socialist or communist societies, in order to show that neither that thesis nor its contrary are entailed by MFT and that the incompatibility thesis is therefore unwarranted.

In Section 2, I briefly review the main objections that MFT has been subjected to (which can be skipped by readers who are familiar with the history of MFT) and I focus mainly on a question that, as far as I am concerned, has not been raised by critics, which is the underspecification of the processes that are supposed to activate evolved moral intuitions. In Section 3, I suggest that we find a contrasting -and surprising- overspecification of the contents of those intuitions when H+ defend what I call the 'learning constraints thesis', which seems mainly to serve the function of rehashing the old incompatibility thesis between human nature and Utopian moralities. In Section 4 I argue that the deflation of the innateness thesis leads MFT to problems not only concerning the mentioned overspecification, but also concerning its aspiration to serve as the foundation of the New Synthesis in Moral Psychology. I conclude by suggesting that if MFT is to live up to its promise it must deal with the underspecification and overspecification problems I point out.

## 2. Criticisms and objections to Moral Foundations Theory

MFT is one of those theoretical models that one wishes were true: if it could actually stand the test of time as a plausible model to explain how (many of) our moral intuitions are grounded in our evolutionary story as a species, we would find ourselves in possession of a general framework from within which we would be able to explain an important part of our emotional lives, and we would also be able to grasp the natural limits of the anthropogenic projects we set in motion in order to build better, fairer and more equal and inclusive societies. Although H+ probably thinks that MFT already offers such advantages, I believe that a mere glance at the objections it has been subjected to proves that it is not really the case — at least not yet.

The main objections that have been raised against the theory concern two very different aspects: empirical evidence and explanatory power. Concerning the first aspect, Suhler & Churchland (2011), for instance, have claimed that MFT has failed to provide solid empirical accounts concerning the problem of how the operation of evolved moral intuitions [from now on merely 'evolved intuitions'] is implemented in the brain, concerning the developmental processes by which those intuitions begin to operate as the basis of our actions and decisions, and, finally, concerning the specific set of genes that encodes the development of the neurophysiology of those evolved intuitions. Are Suhler and Churchland right in demanding that H+ provide concrete evidence concerning the developmental, genetic and neurological bases of moral foundations in order for MFT to qualify as a solid research paradigm, or are they setting the bar too high, as H+ claim? As far as I see it, Haidt and Joseph (2012) are right at least in claiming that the genetic demand would set too high a standard, so high that no theory that postulates the existence of innate contents could fulfill such a requirement (at least given what we know today about the human genome). And perhaps it is also true that MFT fares rather well on the domain of developmental psychology, since there does not seem to be a settled body of research that MFT has to explain away concerning the probable ontogenesis of the different moral foundations that it postulates[3]. However, even if MFT can be relieved of the demand to produce a genetic account of moral foundations, and even if we grant that a plausible developmental account of such foundations can be produced in accordance with the present evidence in the field, I personally do not agree with H+ that the same can be said about the neurological requirement demanded by Suhler and Churchland: if it is (arguably) true that we cannot demand that MFT produce a systematic explanation of the neurological foundations of moral intuitions, it must at least prove to be consistent with the relevant research. And that also holds for the ethnographic requirement: although we cannot demand that MFT produce extensive bodies of evidence concerning the existence

[3] It is true, however, that the contrary does not hold either: although there do not seem to be technical obstacles to designing a developmental account of the emergence of moral foundations, positive evidence is certainly lacking, and the evidence from the field of developmental psychology that H+ usually cite only concerns the alleged existence of general innate mental contents, but not the development of specific moral intuitions.

of moral foundations throughout history and across the plurality of cultures that still exist, it has to be in a position to show that the postulation of such foundations is compatible with the existing research, something which has not been done so far.

Apart from the objections concerning the lack of empirical evidence, an important line of criticism has concerned MFT's explanatory power (which was partially summarized and dealt with in Graham et al. 2013, 102-106), and the most frequent objections have aimed at the specific moral foundations (or 'moral domains', in its previous formulation) that the theory posits. On one variant of this objection, what is claimed is that MFT proposes too many moral foundations: according to such objection, many of the domains that MFT proposes in order to 'cut nature at its joints' could be done away and replaced with, for example, the criterion of suffering (Gray et al. 2012) or that of harm (Harris 2011; Schein & Gray 2018). On this particular front, H+'s response has been understandable: since its inception, it has been precisely the goal of MFT to demonstrate that morality *cannot* be reduced to issues of harm or suffering alone, and that a plurality of moral domains must be taken into consideration in order to properly understand where our moral intuitions stem from. To merge the proposed categories into a single dichotomy would therefore imply giving up the specificity of the theory, a specificity that, according to H+, fares better at explaining moral life without relapsing into reductionist approaches that fail to account for the regularities that MFT is actually able to explain (Graham et al. 2013, 103-105; Koleva & Haidt 2012).

The inverted variant of the objection has claimed, on the contrary, that MFT ends up narrowing the domains of moral intuitions to an arbitrarily limited set of moral foundations, thus leaving out (or forcing in) important aspects of our moral lives that merit to be classified as distinct and discrete domains that MFT has failed to take into account. H+'s attitude to this objections constitutes, in my opinion, one the main reasons of MFT's success: instead of standing their ground and defending the particular set of moral foundations proposed in their previous research, H+'s position has been to stress the dynamic and open character of MFT, either by explicitly stating that it does not aim to provide "an exhaustive taxonomy" of moral foundations (Graham et al. 2009, 1041), or by admitting that the selected ones are probably not the only foundations, but "are the most important ones for explaining human morality and moral diversity" (Haidt & Joseph 2007, 386). What is more: through a Popperian and particularly atypical decision within academia, in 2009 H+ openly invited researchers to challenge the set of moral foundations initially proposed in 2007, in order to reassess its accuracy, an invitation that prompted further research concerning the possibility of expanding the initial set and eventually led to the inclusion of the Liberty/oppression foundation (Haidt 2012, 187).

However, independently of all the criticism and the variable success MFT has shown in addressing it, there is a more pressing problem that the theory faces, and it has to do with the question of the specific processes that are supposed to activate evolved intuitions, a problem that can be approached from two different angles. Firstly: how do original triggers operate? What specific mental processes take place when a certain situation triggers an

evolved intuition? In some cases, the process seems to be fairly straightforward: in the domain of Sanctity/degradation, for example, all that is needed for an intuition of disgust to take place is that the individual either tastes or smells something that produces a foul smell or taste. In such simple, 'original' cases, the intuition in question can be reduced to an event that takes place at a physiological level, without involving any higher cognitive processing. But what about the original triggers that have to do with the Loyalty/betrayal foundation? If "intuitions are pattern-recognition systems" (Haidt & Joseph 2004, 64), what are those patterns? How complex or flexible are they? What processes instantiate the recognition, for example, that a certain situation threatens to compromise the formation or maintenance of cohesive coalitions[4]?

The second angle from which the activation of moral intuitions can be approached concerns the question of how original and current triggers are related: what is the relation between the scenarios that originally triggered an evolved intuition, and an current scenario? Do H+ suppose that a certain structural *analogy* holds between the two? If it so, i.e., if we are to believe that for a certain current scenario to trigger an evolved intuition the individual must (either consciously or unconsciously) detect a similarity between that scenario and the type of scenarios that originally triggered certain intuitions, then the cognitive requirements seem to be rather high, so high that, contrary to what H+ repeatedly stress, the theory can only explain human mental phenomena, but not mental events that take place in other kinds of nonhuman animals (as far as we know). If what is involved is not the recognition of a structural analogy that links original and current triggers (and I fail to see that it could, given the cognitive demands that such a process implies), what is?

Unfortunately, H+ have been particularly vague on this topic, merely pointing out, for example, that an original trigger "maps to" a certain set of current ones (Rozin & Haidt 2013, 367), or that the original outputs of certain evolved intuitions become expanded through "an opportunistic accretion of new domains of elicitors" thanks to the process of preadaptation (Rozin et al. 2008, 764) – a process that is perfectly plausible when considered from a general, abstract perspective, but that says nothing concerning the problem of why certain objects or situations became elicitors of a certain intuition and not others. And this silence is particularly troubling given that MFT claims that all of the moral foundations they have posited (except Sanctity: Graham et al. 2013, 112; Haidt & Graham 2007, 9; Haidt & Joseph 2007, 385) have evolutionary roots that go beyond the lineage of *homo sapiens*, which means that the triggering of evolved intuitions cannot rely on cognitive processes that were out of the reach of our ancestors, both within and without the *homo* order. If the linkage

[4] "The original trigger for the Loyalty foundation is *anything that tells you* who is a team player and who is a traitor, particularly when your team is fighting with other teams" (Haidt 2012, 156).

between certain situations and certain evolved intuitions took place before our lineage parted ways with other primates (such as chimpanzees or bonobos), then those intuitions cannot be built on mental capabilities that are (as far as we know) exclusively human[5].

MFT's lack of specifications concerning not only the operations that link a trigger to an intuition but also the relation between original and current triggers may well be deliberate: remaining elusive concerning both issues allows H+ to remain open to new explanatory models in the domain of psychology and neuroscience, while committing themselves to a precise alternative on either issue would welcome objections and criticisms from several fronts. If we consider H+'s attitude to the thesis of the modularity of the mind, that certainly seems to be the case: "you do not have to embrace modularity, or any particular view of the brain, to embrace MFT. You only need to accept that there is a first draft of the moral mind, organized in advance of experience by the adaptive pressures of our unique evolutionary history" (Graham et al. 2013, 63)[6]. Remaining (or, rather, becoming) noncommittal about the modularity thesis has allowed MFT to walk away unscathed after the avalanche of criticism that such thesis has been subjected to in the last two decades (Fox & Friston 2012; Glascher et al. 2010; Kaskan et al. 2005; Lindquist & Barrett 2012; Oosterwijk et al. 2012; Palecek 2017; Prinz 2006; Thagard & Schröder 2015). Whether it has been a deliberate attempt to steer clear of possible attacks on any proposed alternative or not, MFT's extremely broad strokes on the precise way original triggers operate an on the articulation of original and current triggers has been particularly useful for its aim of stressing that, through culture, almost *anything* can become a trigger of an evolved intuition, provided that we know how to 'hook it up' to certain stimuli. That advantage, however, seems to compensate rather poorly for certain problems that, as we will see, it presents the theory with when we delve into what constraints our evolved intuitions place on economic, political and cultural reforms.

---

[5] It could be argued that there need not be any objective relation between original triggers and current ones, and that the only link between them is the fact that they activate the same mental processes, as Kelly's "co-opt thesis" claims concerning the original and the current functions of the mechanism of disgust (2011, 129-135). That would seem to go against the occasional claims we find in MFT that posit -though always in a vague fashion, as I have pointed out- that there is an actual (but unspecified) relation between the original triggers and the current ones. As we shall see, however, the real problem is that even if MFT resorted to the some version of the co-opt thesis, it wouldn't change the fact that it would remain an utterly underspecified model concerning how evolved intuitions are triggered.

[6] "While the five foundations theory is a nativist theory, it does not need any version of modularity to be true. We suspect that the human mind does contain a number of social-cognitive and social-emotional abilities that are modular «to some interesting degree» […]. For our version of nativism to be true, all we need is the sort of «preparedness» that is widely accepted throughout psychology" (Haidt & Graham 2007, 28). "Each of these five is a good candidate for a Sperber-style learning module. However, readers who do not like modularity theories can think of each one as an evolutionary preparedness […] to link certain patterns of social appraisal to specific emotional and motivational reactions. All we insist upon is that the moral mind is partially structured in advance of experience so that five (or more) classes of social concerns are likely to become moralized during development" (Haidt & Joseph 2007, 381). Cf. also (Graham et al. 2011, 382).

## 3. Human nature, Utopian moralities and learning constraints

The question about the (in)compatibility of human nature and the Utopian moralities on which, for example, socialist and communist modes of production are built on can be approached from at least two angles. One of them is to resort to empirical evidence and claim that the long history of failed struggles, revolts and revolutions that began with the Paris Commune in 1871 shows that every attempt to build a socialist or communist society is doomed to fail. As far as I can tell, H+ never resort to this first strategy, which is not only a clear *non sequitur*[7] but also frequently premised on the assumption that Stalin's, Pol Pot's or Castro's dictatorships, for example, were actually communist societies, something that the most vital fraction of the Marxist Left nowadays vehemently denies (Cinatti 2005; Trotsky 2006).

Another (philosophical rather than historical) strategy, which is the one adopted by H+, is to claim that socialist and communist modes of production are premised on certain principles that are fundamentally at odds with the moral intuitions that constitute our evolved human nature. (Thus, rather than resorting to History in search of evidence, H+ resort to MFT to explain why History turned out the way it did). But what are exactly those intuitions that have proven to be insurmountable hindrances to every socialist or communist project? At first glance, they are intuitions that have to do with the allocation of resources: H+'s evolutionary narrative for the moral foundation of Fairness is that natural selection has provided us with a tendency to feel flashes of approval or disapproval when we perceive certain patterns in the way resources are allocated. If those patterns imply, for example, "sharing material goods with [our] close kin" (Haidt & Joseph 2007, 375-376), we will feel a flash of approval or even pleasure (Graham et al. 2013, 62-63); if, on the contrary, people, for example, "try to cheat us or take advantage of us" (Haidt 2012, 152-153), we will experience a gut reaction that will lead to a negative assessment of the situation. What the historical Left has failed to understand is that, because of this, if a certain society is organized on a mode of production that forces its individuals to face violations to the fairness principle in a recurrent fashion, that society will face resistance and will only be able to persist in time through authoritarian measures and through "the increasing application of threats and force to compel cooperation" (Haidt 2012, 416).

But, one should be allowed to ask: what is it exactly about the evolved intuitions concerning the Fairness foundation that turns capitalism into humankind's best friend? What are the contents of those intuitions that render communist or socialist principles unbearable

[7] It is not only that "WEIRD societies (Western, Educated, Industrial- ized, Rich, and Democratic) are bad places to look for human universals, because WEIRD people are outliers in so many ways" (Koleva & Haidt 2012, 177). That had already been claimed by Rousseau (against Hobbes's grim view of human affairs) and had been restated over and over by every revolutionary, from Marx to Kropotkin. The point is rather that, strictly speaking, *no* historical evidence whatsoever can be used as evidence on itself of the existence of a particular human nature.

in the long run to individuals? What precise definition of the Fairness principle do those intuitions imply? If H+ are confident that those evolved intuitions are sufficient to decide what kind of societies are compatible with us as a species and which are not, one would expect a precise description of what is involved in them. Unfortunately, as I said before, the expected explanations tend to be rather scanty and vague, and leave us wanting the details of which patterns of resource allocation trigger negative gut reactions and which do not, or how that process of pattern recognition is implemented in the minds of the species that, according to the authors, share with us at least the building blocks of the moral foundation of fairness.

The closest H+ come to providing us with an (indirect) answer to those questions is the idea that there are 'learning constraints' that set limits to what kind of beliefs, traits of character and attitudes can effectively be instilled in a human being by way of education and instruction:

> Does anyone seriously believe that it would be as easy to teach children to love their enemies as to hate them? Or that betrayal of friends and family is as intuitively pleasing as is loyalty to them? (Such "unnatural" beliefs may have been taught in Mao's China, but only imperfectly and with great effort. Loyalty to kin is far more easily learned than its opposite). (Haidt & Graham 2007, 28)

Those learning constraints are obviously the result of the fact that, contrary to what many anthropogenic projects have assumed, the mind is not a blank slate, which means that the manipulation of the political and economical variables that frame the subject's daily actions will not be enough to modify his subjectivity, and some of the outcomes that social reformers envision concerning the malleability of human behavior and attitudes will simply not be possible:

> A fundamental problem with many virtue theories is they assume that virtues are learned exclusively from environmental inputs. They implicitly endorse the old behaviorist notion that if we could just set up our environment properly, we could inculcate any virtue imaginable, even virtues such as 'love all people equally' and 'be deferential to those who are smaller, younger, or weaker than you.' Yet one of the deathblows to behaviorism was the demonstration that animals have constraints on learning: some pairings of stimuli and responses are so heavily prepared that the animal can learn them on a single training trial, while other associations go against the animal's nature and cannot be learned in thousands of trials. Virtue theories would thus be improved if they took account of the kinds of virtues that 'fit' with the human mind and of the kinds that do not. Virtues are indeed cultural achievements, but they are cultural achievements built on and partly constrained by deeply rooted preparednesses to construe and respond to the social world in particular ways. (Haidt & Joseph 2004, 62-63)

> If there were no first draft of the psyche, then groups would be free to invent Utopian moralities (e.g., "from each according to his ability, to each according to his need"), and they would be able to pass them on to their children because all moral ideas would be equally learnable. This clearly is not the case. (Graham et al. 2013, 63)

All of this provides us with two elements: a theoretical thesis and a rule of thumb. On the side of the theory, we can rule out the existence of "equipotentiality in moral learning" (Haidt & Joseph 2007, 375), because we are endowed with evolved intuitions that do not seem to be overridable (if they were, they wouldn't be a constraint to acquiring behavioral patterns that were contrary to what those intuitions tell us). On the practical side, we can now ask ourselves, before promoting a certain social, cultural, economical or political program, whether it collides with *any* of our evolved intuitions (either from the domain of Care, Fairness, Loyalty, Authority, Sanctity or Liberty); if it does, it will be received with a gut reaction that will -immediately or eventually- lead to its rejection. This being so, research into the precise contents of the set of evolved intuitions that cover the proposed moral foundations should put us in a position to finally know, at least in broad strokes, which paths can be safely traveled by political legislators, social reformers and moral philosophers, thus placing MFT as a first step into what E.O. Wilson envisaged as the most urgent task of a biologically informed philosophy:

> Above all, for our own physical well-being if nothing else, ethical philosophy must not be left in the hands of the merely wise. Although human progress can be achieved by intuition and force of will, only hard-won empirical knowledge of our biological nature will allow us to make optimum choices among the competing criteria of progress. (Wilson 1978, 7)

> Self-knowledge will reveal the elements of biological human nature from which modern social life proliferated in all its strange forms. It will help to distinguish safe from dangerous future courses of action with greater precision. We can hope to decide more judiciously which of the elements of human nature to cultivate and which to subvert, which to take open pleasure with and which to handle with care. (Wilson 1978, 96-97)

However, with the exception of Utopian moralities[8], MFT does not provide us with such specific guidelines concerning safe and dangerous courses of action, and neither does it seem to want to do so. As we will see, this is not incidental, but rather an intended consequence of MFT initial objectives.

[8] There are other scattered exceptions, of course, as when we are warned of the futility of trying to teach children to 'love all people equally' or to force people to 'be deferential to those who are smaller, younger, or weaker than" them (Haidt & Joseph 2004, 62-63). But apart from those rather general exceptions, MFT maintains a cautious attitude concerning what paths to avoid.

## 4. Moderate innateness

Even if we don't find explicit and systematic elaborations concerning the concept in the publications that develop the core tenets and projections of MFT, the notion of 'learning constraints' is particularly important in H+'s overall argument because of the fact that it contributes to an important goal that MFT set itself from its very beginning, which is to provide a moderate version of the innateness thesis. The 'New Synthesis in moral psychology' that H+ felt that loomed on the horizon could not be attained from a blank slate, constructionist approach that pinned every mental content to culture; it needed to be rooted in a scientific explanation of the inner workings of the human mind that took into consideration its evolutionary past and the traces that evolution had left in our cognitive architecture. But it also had to be able to explain the apparent plurality of habits, customs, beliefs and traits of character that we find throughout history, something that the strong versions of the innateness thesis were not able to do. And the notion of learning constraints seems to do precisely that job: instead of claiming that natural selection has endowed our species with certain innate, hardwired contents that determine us to act in certain ways, it states that the innate contents of our mind (be they modular or not) merely determine that certain habits, attitudes traits of character, etc., will be actually acquirable while others won't be – or that, at least, some of them will be more easily acquirable than others[9]. Thus, rather than stating that human societies will show certain inescapable regularities throughout space and time, the learning constraint approach stresses that human societies can be amply diverse and that such diversity will be the result of the precise interaction between culture and nature, between the particularities of the society we are born in and the learning constraints that we are universally born with[10].

This approach provides us with a model that is highly sensitive to variations, both at the communal and the individual level. At the communal level, H+ emphasize that despite there being (at least) six moral foundations, communities (or cultures) tend to rely disproportionately on only some of them. This disproportionate reliance on the different moral foundations can be either spontaneous[11] or deliberate (Graham et al. 2013, 110), and

[9] In both Haidt & Graham (2007, 28) and Graham et al. (2013, 62) we find, intermingled with the learning constraints thesis, this 'ease of learning' thesis. The fact that the authors see no conflict between both theses and that they do not care to elaborate on the consequences of each is further evidence, I believe, of the fact I am trying to emphasize here that the main drive behind these type of reflections on H+'s part is to build and alternative to strong versions of innateness.

[10] The complementary preference for the 'teeming' (Sperberian) variant of the modularity thesis over the Fodorian one is an expression of this same search for a moderate position on the problem of innateness (Haidt & Graham 2007, 28; Haidt & Joseph 2007, 379-386).

[11] Before MFT, Haidt even talked of the "selective loss of intuitions" as a somewhat necessary process in the development of culture: "A culture that emphasized all of the moral intuitions that the human mind is prepared to experience would risk paralysis as every action triggered multiple conflicting intuitions" (Haidt 2001, 827).

can even lead to the virtual disappearance of one or more of those foundations, which means that in some cultures or communities certain foundations will be entirely absent from the realm of moral evaluations[12]:

> The foundations are not the finished buildings, but the foundations constrain the kinds of buildings that can be built most easily. Some societies might build a tall temple on just one foundation, and let the other foundations decay. Other societies might build a palace spanning multiple foundations, perhaps even all five. […] Similarly, the moral foundations are not the finished moralities, although they constrain the kinds of moral orders that can be built. Some societies build their moral order primarily on top of one or two foundations. Others use all five. (Graham et al. 2013, 65)

To see how the communal and the individual levels complement each other, we can take as an example the foundation of Sanctity, which includes the evolved intuitions that natural selection provided us with as a means of avoiding sources of contamination, and the whole set of original and current triggers that activate those intuitions. On the communal level, does MFT predict that, given the importance of avoiding contaminants in the Environment of Evolutionary Adaptedness, every historical culture will show to rely heavily on that foundation, making Sanctity a central issue on moral matters? Not at all. What it predicts is that, when considered from a global perspective, the frequency with which cultures tend to rely on Sanctity concerns cannot be explained merely by chance or cultural influence, but that, nevertheless, its importance will depend on each community, to the point that in some cultures Sanctity concerns will be completely absent.

To this source of intercultural variation, we must add the personal (or intracultural) variable: even if we live or have been raised in communities where Sanctity concerns generally play an important part in moral life, we can still shift from one foundation to another when considering a certain moral issue, which will lead us to view it from a different perspective and, if other evolved intuitions are triggered, perhaps change our evaluation of the situation completely. Such "code switching" -as Haidt & Hersh called it (2001, 217)- can take place as a result of several processes: we can reevaluate our initial assessment through a process of conscious reasoning (deliberation); we can be made to see things under a different light in the course of an exchange of arguments (dialogue), or as a result of manipulation aimed at making us decide in a certain way (persuasion). As Haidt has particularly emphasized, this is something that politicians, lawmakers and publicists have known for a long time, which explains why mastering the skill of linking their ideas and products to our evolved intuitions has been a central part of their strategy: "Political parties and interest groups strive to make their concerns become current triggers of your moral modules. To get your vote, your money, or your time, they must activate at least one of your moral foundations" (Haidt 2012, 149).

[12] "Secular westerners have gradually *lost touch with* the ethics of divinity, shrinking the moral domain mostly to what Shweder called 'the ethics of autonomy'…" (Rozin & Haidt 2013, 368; my italics); "If your moral matrix *rests entirely on* the Care and Fairness foundations, then it's hard to …" (Haidt 2012, 183).

But if almost anything can be made to trigger our evolved intuitions (provided that we know how to frame it) this is because what natural selection has provided us with is a set of pattern *recognition* mechanisms: they are mechanisms that lead us to avoid not situations that *objectively* affect (or affected in the EEA) our chances of survival and reproduction, but rather situations that we unconsciously *identify* as ultimately threatening our survival and reproduction. (Danger, in other words, is in the eye of the beholder). What we will tend to avoid, therefore, are those situations where we recognize a certain pattern. But whether that pattern actually reveals an objective state of the world or not is altogether irrelevant. And whether we perceive a pattern in a certain situation will depend on a plurality of factors, some of them related to our community and the moral foundations on which it is built, some of them to our particular biography, and some of them to chance and our interaction with others. If, for example, a certain course of action I am about to take *objectively* threatens the cohesion of my community (Loyalty/betrayal foundation) but I am not aware that it does so, I will not be deterred from doing it by my innate setup, because my evolved intuitions will not be triggered. And even if I become aware that it may lead to that undesired outcome, I may decide to, or happen to, or be made to look at it from a different perspective that presents that course of action as actually desirable (for some reason or another), which will make the incompatibility between my action and my evolved intuitions concerning the Loyalty foundation meaningless in practice and, therefore, effectless.

This approach certainly allows MFT to explain the variations we see in history concerning how strongly individuals and communities rely on certain moral foundations, but it lands H+ in a particularly problematic position. In the first place, the cost of being able to explain historical variations is a *substantial deflation of the innateness thesis*: if we are born with an innate set of intuitions concerning moral matters, but it is actually a modifiable, malleable first draft that can virtually be rewritten, what effect can it actually be said to have in our decisions? What is more: what explanatory and predictive power does the moderate (deflated) version of innateness provide MFT with? This is particularly important given the fact that, while one of the main goals of MFT was to offer an alternative to the strong versions of innatism (which were not able not account for historical variation), the other goal was, as I stressed earlier, to avoid becoming reduced to a merely sociological or culturalist account of moral life no different in kind from, for example, that of R. Fiske, who Haidt frequently acknowledges as a source of insight but whose account lacks, from the latter's standpoint, a complementary evolutionary explanation of the universality of the psychological patterns studied by Fiske. Abandoning the innateness thesis would have meant forfeiting the possibility of achieving the New Synthesis in Moral Psychology (i.e, the synthesis between the social and the natural sciences or, more specifically, between ethics and Evolutionary Psychology) that H+ believed that MFT was finally capable of delivering.

Another problem that the deflation of the innateness thesis presents is that it seems to be completely at odds with the picture I presented in section 3 concerning H+'s assessment of the viability of Utopian projects and the 'learning constraints' thesis: while the deflated

innateness position leans towards a picture of humans as hermeneutic machines in whose highly variable, fluctuating, plastic and subjective *recognition* pattern mechanisms lies the eventual activation of an evolved intuition, H+'s approach to Utopian projects rather seems like a problematic return to the stronger versions of innatism. When we revisit the claim about the incompatibility between human nature and socialist or communist societies after having delved into the deflated version of innateness to which the learning constraints thesis leads MFT, something seems to have been glued together not too carefully. Am I saying that when it comes to the question of exploring alternatives to Capitalist societies there is an ideological overspecification in MFT that is absent when dealing, for example, with American liberal and conservative politics? I see no other explanation[13]. Is that ideologically motivated overspecification central or structural to MFT or SIM? Not at all. Quite on the contrary, I believe that (as Antonio Gramsci did with Machiavelli) both SIM and MFT can be reinterpreted as cookbooks for the Radical Left, providing tactical advice to those who aim to build societies that do not lead to the destruction of the planet and of the human race in their search for profit. That is, without a doubt, beside the point of this article. But it shows, I believe, how unwarranted the thesis of the fundamental incompatibility between human nature and socialist or communist societies is when considered from the very premises and conclusions of MFT.

## 5. Conclusions

As I stated in the beginning, one of H+'s main goal in proposing MFT was to provide an alternative to strong innatism (which cannot account for variation), on the one hand, and to radical culturalism (which cannot account for regularities), a middle way that can be synthetized in the idea that "morality is innate *and* highly dependent on environmental influences" (Graham et al. 2013, 61). As I have tried to show, however, although some of the objections that have been raised against the theory can be considered at least partially unjustified, there are other problems that MFT cannot deal with very easily (at least not yet). One of them is the problem of giving a precise account of how evolved moral intuitions operate (i.e., what cognitive mechanisms instantiate them); the other is the problem of the linkage between original and current triggers.

As I suggested, the ambiguous and non-committal attitude that H+ have shown towards the modularity thesis has allowed them to stay afloat during the storm of criticism that started to fall upon that thesis, but it did so at the price of an underspecification of the precise processes that are at the basis of the activation of evolved intuitions, an underspecification that contrasts with the surprising overspecification that we find when dealing with the possibility

[13] Every bit as unwarranted by MFT's own principles would be to state that, given the environmental destruction that lies at the core of the capitalist mode of production, it is incompatible with the evolved intuition that concerns the Care foundation and that, as a consequence, it will not be able to persist in time without resorting to dictatorial measures.

of exploring alternatives to capitalist societies, which feels as an uncalled for remnant of the strong versions of innatism that cannot be accounted for on the grounds that MFT was built upon.

Although it it needless to say that the New Synthesis in Moral Psychology cannot be built on ideological prejudices, neither will a mere deflation of the innateness thesis suffice, since that can only lead to the murky waters that MFT is now sailing, leaning one minute to the shores of strong innatism (when spooked by the ghost of communism) and another to the shores of culturalism (when faced with the uncanny diversity that a glance at our moral history presents us with). If MFT is to actually become the solid research framework it aims to be, it must face both its under and its overspecifications, ridding itself of ideological assumptions and producing a precise account of how evolved intuitions operate and of what makes specific scenarios capable of becoming current triggers of those intuitions.

## References

Cinatti, C. (2005). La actualidad del análisis de Trotsky frente a las nuevas (y viejas) controversias sobre la transición al socialismo. *Estrategia Internacional*, *22*. https://www.ft-ci.org/La-actualidad-del-analisis-de-Trotsky-frente-a-las-nuevas-y-viejas-controversias-sobre-la?lang=es

Fox, P. T., Friston, K. J. (2012). Distributed processing; distributed functions? *NeuroImage*, *61*(2), 407-426. https://doi.org/10.1016/j.neuroimage.2011.12.051

Glascher, J., Rudrauf, D., Colom, R., Paul, L. K., Tranel, D., Damasio, H., Adolphs, R. (2010). Distributed neural system for general intelligence revealed by lesion mapping. *Proceedings of the National Academy of Sciences*, *107*(10), 4705-4709. https://doi.org/10.1073/pnas.0910397107

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. *Advances in Experimental Social Psychology*, *47*, 55-130. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Graham, J., Haidt, J., Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029-1046. https://doi.org/10/fhfs36

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366-385. https://doi.org/10/cq64hc

Gray, K., Young, L., Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychological Inquiry*, *23*(2), 101-124. https://doi.org/10/bqc4

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814-834. https://doi.org/10.1037//0033-295X.108.4.814

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. London: Vintage Books.

Haidt, J., Graham, J. (2007). When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. *Social Justice Research*, *20*(1), 98-116. https://doi.org/10.1007/s11211-007-0034-z

Haidt, J., Hersh, M. A. (2001). Sexual morality: The cultures and emotions of conservatives and liberals 1. *Journal of Applied Social Psychology*, *31*(1), 191-221.

Haidt, J., Joseph, C. (2004). Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus*, *133*(4), 55-66. https://doi.org/10.1162/0011526042365555

Haidt, J., Joseph, C. (2007). The moral mind. How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. En P. Carruthers (Ed.), *The Innate Mind. Vol. 3: Foundations and the future*, pp. 367-392. New York: Oxford University Press.

Harris, S. (2011). *Moral landscape: How science can determine human values*. New York: Free Press.

Kaskan, P. M., Franco, E. C. S., Yamada, E. S., de Lima Silveira, L. C., Darlington, R. B., & Finlay, B. L. (2005). Peripheral variability and central constancy in mammalian visual system evolution. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1558), 91-100. https://doi.org/10.1098/rspb.2004.2925

Kelly, D.R. (2011). *Yuck! The Nature and Moral Significance of Disgust*. Cambridge: MIT Press.

Koleva, S., Haidt, J. (2012). Let's Use Einstein's Safety Razor, Not Occam's Swiss Army Knife or Occam's Chainsaw. *Psychological Inquiry*, *23*(2), 175-178. https://doi.org/10/ggdmm7

Lindquist, K. A., Barrett, L. F. (2012). A functional architecture of the human brain: Emerging insights from the science of emotion. *Trends in Cognitive Sciences*, *16*(11), 533-540. https://doi.org/10.1016/j.tics.2012.09.005

Oosterwijk, S., Lindquist, K., Anderson, E., Dautoff, R., Moriguchi, Y., Barrett, L. (2012). States of mind: Emotions, body feelings, and thoughts share distributed neural networks. *NeuroImage*, *62*, 2110-2128. https://doi.org/10.1016/j.neuroimage.2012.05.079

Palecek, M. (2017). Modularity of Mind: Is It Time to Abandon This Ship? *Philosophy of the Social Sciences*, *47*(2), 132-144. https://doi.org/10.1177/0048393116672833

Prinz, J. J. (2006). Is the mind really modular? En R. Stainton (Ed.), *Contemporary debates in cognitive science*, pp. 22-36. Oxford: Blackwell.

Rozin, P., Haidt, J. (2013). The domains of disgust and their origins: Contrasting biological and cultural evolutionary accounts. *Trends in Cognitive Sciences*, *17*(8), 367-368. https://doi.org/10.1016/j.tics.2013.06.001

Rozin, P., Haidt, J., McCauley, C. R. (2008). Disgust. En M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of Emotions*, pp. 757-776. New York: Guilford Press.

Schein, C., Gray, K. (2018). The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm. *Personality and Social Psychology Review*, *22*(1), 32-70. https://doi.org/10/gcvxrf

Suhler, C. L., Churchland, P. (2011). Can Innate, modular «foundations» explain morality? Challenges for Haidt's Moral Foundations Theory. *Journal of Cognitive Neuroscience*, *23*(9), 2103-2116; discussion 2117-2122. https://doi.org/10/fsr92d

Thagard, P., Schröder, T. (2015). Emotions as Semantic Pointers: Constructive Neural Mechanisms. En L. F. Barrett & J. A. Russell (Eds.), *The psychological construction of emotion*, pp. 144-167. London: The Guilford Press.

Trotsky, L. (2006). *La revolución permanente*. Buenos Aires: Utopía Libertaria.

Wilson, E. O. (1978). *On Human Nature*. New York: Penguin.