31 January 2022.

Main text – 13,900. References – 5505.

RETHINKING NORM PSYCHOLOGY

Cecilia Heyes

Department of Experimental Psychology & All Souls College

University of Oxford

Oxford OX1 4AL

cecilia.heyes@all-souls.ox.ac.uk

https://users.ox.ac.uk/~ascch/

orcid.org/0000-0001-9119-9913

Abstract. Norms permeate human life. Most of our activities can be characterised by rules about what is appropriate, allowed, required, or forbidden – rules that are crucial in making us hyper-cooperative animals. This article examines the current cognitive-evolutionary account of 'norm psychology', of the mental processes that enable normative behaviour, and proposes an alternative that is better supported by current evidence

and better placed to promote interdisciplinary dialogue about norms. The incumbent theory focusses on rules and claims that humans genetically inherit cognitive and motivational mechanisms specialised for processing these rules. The cultural evolutionary alternative defines normativity in relation to behaviour – compliance, enforcement, and commentary – and suggests that it depends on implicit and explicit processes. The implicit processes are genetically inherited and domain-general; rather than being specialised for normativity, they do many jobs in many species. The explicit processes are culturally inherited and domain-specific; they are constructed from mentalising and reasoning by social interaction in childhood. The cultural evolutionary, or 'cognitive gadget', account implies that researchers should not merely chart cultural and developmental variation in normativity, but test whether that variation is due to nature (genetic factors), nurture (learning and social learning), and/or culture (cultural learning). More broadly, the cultural evolutionary perspective suggests that people alive today - parents, peers, educators, elders, politicians, lawyers – have more responsibility for sustaining normativity than the nativist view implies. Our actions don't just shape and transmit the rules, they create in each new generation mental processes that can grasp the rules and put them into action.

Keywords: cognitive gadgets; human cooperation; cultural evolution; domain-general learning; economic games; evolutionary psychology; moral psychology; norm psychology; reinforcement learning; social learning.

1. Introduction

Human lives are drenched in social norms. Our clothes, eating habits, sexual and parental behaviours, and day-to-day modes of interaction with one another – from greeting and speaking to helping and harming – can be described by rules about what is appropriate, allowed, required, or forbidden in different contexts for various members of a social group. Some norms are crisply codified in law (e.g., drive on the right, thou shalt not kill), while others would be difficult for any group member to articulate (e.g., how much eye contact is appropriate in conversation with superiors, subordinates, equals). Some have a moral flavour – a prohibition against unnecessary harm to other people may apply to everyone at all times - while other norms, such as who should wear a particular kind of hat, are obviously transitory and group-specific. Norms vary on many dimensions, but norms of some sort appear to be present in all human cultures (Brown 1991), and to be rare, minimal, or absent in other animals (Jensen 2016; Schmidt & Rakoczy forthcoming; see Andrews 2020 and Fitzpatrick 2020 for contrasting views).

'Norms' have been part of the conceptual toolkit of anthropology, economics, sociology and social psychology for as long as those disciplines have existed. 'Norm psychology' emerged more recently and has a different fan base. About 15 years ago, scholars interested in human evolution - anthropologists, biologists, economists, philosophers, and (a few) psychologists - began to use 'norm psychology' (or 'normative cognition') to refer to a set of cognitive and motivational mechanisms that, they believe, have been specialised by genetic evolution for processing social 'rules' or 'behavioural standards' (e.g. Boyd & Richerson 2005; Chudek & Henrich 2011; Fehr & Schurtenberger 2018; Fitzpatrick 2020; Henrich 2020;

Henrich & Muthukrishna 2021; House 2018; House et al. 2020; Kelly & Davis 2018; Mikhail 2011; O'Neill & Machery 2018; Richersen et al. 2016; Sripada & Stich 2006; Zefferman 2014). 'Norm psychologists', whatever their disciplinary affiliation, believe that the mental processes guiding normative behaviour are important, domain-specific, and genetically inherited. They are important because norms enable cooperation, and the human capacity for cooperation is a large part of what makes us such peculiar, and peculiarly successful, animals. They are domain-specific in the sense of being different from the cognitive and motivational mechanisms that do other jobs in our mental economy, such as predicting inanimate events or detecting predators. And the domain-specific features of norm psychology, rather than basic ingredients found in other animals, are programmed in our genes. According to the gene-culture (or culture-gene) coevolutionary view supported by norm psychologists, norm content – for example, the prescription or proscription of cousin marriage or hat wearing – is learned through social interaction. However, during human evolution the effects on behaviour of norm content provided powerful selection pressure for the genetic evolution of psychological mechanisms specialised for norm processing. On this account, norm content is acquired and implemented by mechanisms that have been tailored for norm processing by natural selection acting on genetic variants.

The first of these claims, about the importance of norm psychology, is rock solid. Questions remain about exactly what types of norms allowed our ancestors to begin cooperating with unrelated others, and ultimately in some societies to cooperate over long timescales and at high risk - for example, in financial markets (Boyd & Richerson 2001, Boyd, Gintis, Bowles & Richerson 2003, Boyd 2016, Henrich 2015, Sterelny 2021). However, no one would deny that norms-of-some-kind were and are crucial for human cooperation. Given their importance, norms need to be explained not only functionally, in terms of their effects on behaviour and fitness, but also at the internal, psychological level. We need to know what it is about our minds that enables us to learn, implement, and enforce norms, and where those features come from. Without this information, our understanding of norms is radically incomplete; there is a missing link between the social science and the natural science of norms; and researchers have limited capacity to inform business, education, and government in designing laws and policy interventions to promote cooperation in contemporary societies (Kelly & Morar, 2020; Raymond, Kelly & Hennes, 2021).

It is precisely because norm psychology is important that the other claims of norm psychologists, about domain-specificity and genetic inheritance, deserve closer scrutiny. The domain-specificity claim is contrary to theories of norm processing rooted in social psychology (Gross & Vostroknutov 2021), philosophy (Bicchieri & McNally 2018; Colombo 2014; Nichols 2021), and cognitive neuroscience (Theriault et al. 2021; Veissiere et al. 2019), but the two camps – norm psychologists and domain-generalists - rarely talk to one another. Norm psychologists cite evidence they regard as consistent with domain-specificity and genetic inheritance, but do not address the evidence interpreted by domain-generalists to show that normative behaviour depends on cognitive and motivational processes that do many other jobs. Similarly, if they refer to the work of norm psychologists at all, domain-generalists usually just state their disagreement. No one is testing the theories against one another, assessing in a systematic, scientifically healthy way to what extent the evidence favours domain-specificity over domain-generality or vice versa (Press, Yon & Heyes 2022).

The standoff may be due to a sense on both sides that they have different purposes. Norm psychologists' want to understand human evolution, whereas domain-generalists have a tighter focus on the character and development of normative thinking and behaviour in contemporary Western societies. This contrast does not justify a lack of engagement or make the standoff less wasteful of intellectual and financial resources. There are clearly areas of common purpose where the two camps are making conflicting claims that need to be resolved. But it is not unusual in academia and everyday life for people with different projects and backgrounds to live in bubbles.

The overriding purpose of this article is to burst the normativity bubbles, to draw norm psychologists and domain-generalists together for productive debate, in Open Peer Commentary and in future research on the psychology of normativity. I aim to do this by outlining a cultural evolutionary account of norm psychology. This 'gadget account' (Heyes 2018a; 2019a) is synthetic in being both 'cognitive-evolutionary' (Kelly & Setman 2021) and compatible with the evidence that domain-general processes are important in the development of normative behaviour. I argue that it fits current evidence better than previous evolutionary accounts, but my purpose is not to show that the gadget account is *right*. Not surprisingly, I think it has many merits, but the function of this article is to provide a framework for future research in which conflicting claims can be tested against one another. The article proposes both an alternative cognitive-evolutionary theory of normativity and a new 'poverty-wealth scheme' for testing it against the original.

I begin by summarising the tenets of 'nativist norm psychology', the evolutionary framework that the gadget account seeks to revise (section 2). I then discuss problems with the nativist view that motivate revision (section 3) and present the cultural evolutionary alternative (section 4). The final section discusses some potential objections to the cultural evolutionary framework, and its implications for research and in the wider world.

2. Nativist norm psychology

Norm psychology descends from and overlaps with 'moral psychology', a project also pursued collaboratively by philosophers and scientists. Norm psychology is broader than moral psychology in its concern with 'conventional' norms (e.g., relating to duration of eye contact) as well as 'injunctive' or 'prescriptive' norms (e.g., relating to harm), and narrower in focussing on norm acquisition and implementation. Unlike moral psychology, norm psychology is concerned with the psychological processes mediating altruism, well-being, character, virtue, and the moral emotions, such as shame and guilt, only insofar as they influence an individual's capacity to acquire and implement social rules (Kelly & Setman 2021). Norm psychology was borne out of discontent with moral psychology (Westra & Andrews, forthcoming). However, the overlap remains so substantial that this article will often refer to the work of moral psychologists (e.g., Bear & Knobe 2017; Cushman 2013; Greene 2017; Haidt 2001; 2012). It will also refer to important, independent empirical work by Tomasello and colleagues (e.g., Göckeritz, Schmidt & Tomasello 2014; Schmidt, Rakoczy & Tomasello 2019). Unlike norm psychology, the theory they have developed depends on a constructivist rather than a computational view of the mind.

Sripada and Stich (2006) coined the term 'norm psychology' and provided a compelling manifesto in a chapter titled *A Framework for the Psychology of Norms*. They gave 1) a ‘characterisation’ of norms, 2) a survey of evidence that norm psychology is ‘innate’, and 3) a preliminary model of the psychological mechanisms enabling the acquisition and implementation of norms. The following summary of nativist norm psychology uses their framework as a springboard because, even now, it captures the key tenets more clearly than any other statement I have found. Also, crucially, it remains representative of the field. With minor exceptions, to be noted, most nativist norm psychologists work within Sripada and Stich’s framework. They use similar definitions, the same categories of evidence, and, in many cases, the same key examples.

2.1 Characterisation - rules

A norm is ‘a rule or principle that specifies actions which are required, permissible or forbidden’, that has ‘independent normativity’ and ‘intrinsic motivation’ (Sripada & Stich 2006). Independent normativity means that to qualify as a norm a rule may be, but need not be, recognised and enforced by social institutions and laws. Intrinsic motivation means that ‘people are motivated to comply with norms as *ultimate ends*, rather than as a means to other ends’ (Sripada & Stich 2006, see also Gavrilets & Richerson 2017; Henrich & Ensminger 2014). ‘Violations of norms, when they become known, typically engender *punitive attitudes*, like anger, condemnation, and blame, directed at the norm violator, and these attitudes sometimes lead to punitive behaviour’.

2.2 Evidence of innateness

The empirical case for the innateness of norm psychology can be summarised in four propositions relating to universality, importance, development, and motivation.

Norms are universal. There is a broad consensus that norms are present in all human cultures, and, although there are common themes, that norm content is highly diverse (Brown 1991; Henrich 2020; Wilson & Sober 1998). Due to this combination of commonality and diversity, norm psychologists tend to be agnostic about whether norm content is innate. They leave open the question of whether humans are born with a genetically inherited propensity to believe that specific behaviours are prescribed or forbidden. However, they assert that norm content is acquired and implemented by innate psychological mechanisms, cognitive and motivational processes that have been tailored for norm processing by natural selection acting on genetic variants.

The combination of commonality and diversity is obvious in some domains, such as clothing and body adornment. Most societies have sartorial norms, but the prescribed and proscribed items vary from penis sheaths to powdered wigs. Less obviously, the commonality-plus-diversity pattern is present in domains more closely associated with cooperation and morality. Most societies have norms prohibiting killing and physical assault, and promoting sharing, reciprocation and helping, but there is wide variation in tolerance of harmful behaviour, especially against women and out-group members, and in the extent of helping expected in various contexts and by people in different social roles.

Norms are important. Sripada and Stich emphasised, uncontroversially, that norms were important in human evolution and that they continue to play crucial roles in contemporary life. Norms 'govern a vast array of activities, ranging from worship to appropriate dress to disposing of the dead. And while some norms deal with matters that seem to be of little importance, others regulate matters like status, mate choice, food and sex that have a direct impact on people's welfare and their reproductive success' (p. 282).

Normative behaviour develops early and without teaching. There is plenty of evidence that normative behaviour appears early in childhood. Sripada and Stich (2006) cited evidence that children can distinguish prescriptive from descriptive social rules at 3 to 5 years of age (Nucci 2001), and that cross-cultural variation in fairness norms is in place by 9 years of age (Henrich et al. 2004). Since 2006, developmental studies have revealed more precocious normative achievements. For example, 6-year-olds punish unequal distribution of resources between third parties when delivering punishment incurs a cost (McAuliffe, Jordan & Warneken 2015), and 3-year-olds engage in 'normative protest', saying things like 'No, it does not go like this!' when a puppet fails to do the same thing as an adult (Kenward 2012; Rakoczy, Warneken & Tomasello 2008; Schmidt & Rakoczy forthcoming).

Normative behaviour is intrinsically motivated in the sense that norms have a 'unique kind of subjective authority which differs from standard instrumental motivation', that makes people 'disposed to comply with norms even when there is little prospect for instrumental gain, future reciprocation or enhanced reputation, and when the chance of being detected for failing to comply with the norm is very small' (Sripada & Stich 2006, p.285). Sripada and Stich pointed out that this has been a standard view in moral philosophy, sociology (Durkheim 1912), and parts of social psychology (Batson 2014) for many years, and mention plausible, everyday examples such as tipping in a restaurant to which you know you will never return, and jumping in a river to save a drowning person (Frank 1988). For hard data, they turned to behavioural economics, highlighting evidence that in experimental games people cooperate – for example, give more money than necessary to another player – when players are anonymous and aware that they will have just one exchange (Marwell & Ames 1981). People also punish norm violation – for example, failure to contribute to common goods – when all players are anonymous and delivering punishment incurs a cost (Fehr & Gachter 2002). This kind of 'costly punishment' of 'free-riders' occurs both when punishers have lost out because of norm violation, and, in the case of 'third party punishment', when they have merely observed norm violation in a game where the punisher had no stake (Fehr & Fischbacher 2004).

2.3. Model

Figure 1 shows Sripada and Stich's (2006) model. They suggest that both main components, the acquisition mechanism and the execution mechanism, are innate and domain-specific; they have been shaped by genetic evolution to operate in a different way from mechanisms

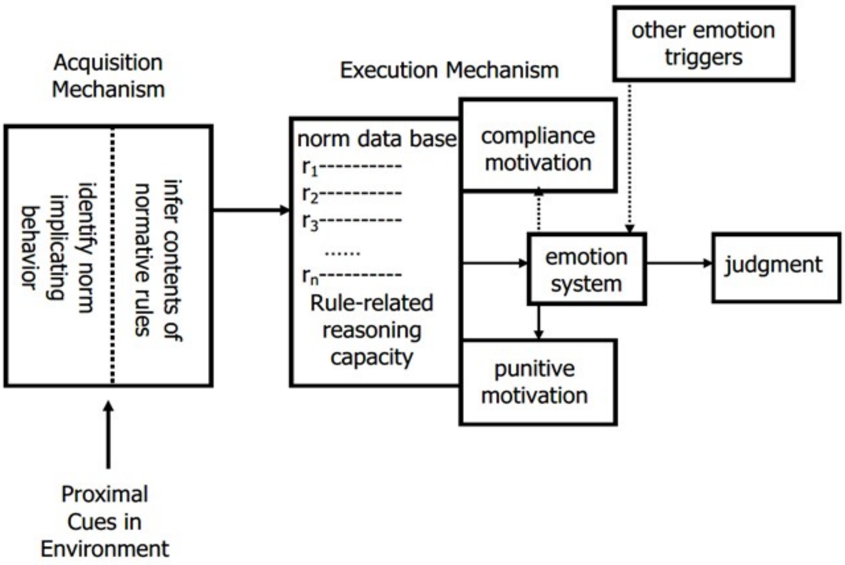

Figure 1. Sripada and Stich's 'sketch of the mechanisms underlying the acquisition and implementation of norms' (2006, p. 298).

that acquire and implement rules that are not norms. The acquisition mechanism operates automatically from early in development. It detects behavioural cues indicating that there is a norm in the local cultural environment, infers the content of that norm, and passes this information about norm content to the implementation system, where it is stored and used. The implementation mechanism maintains 'a data base of normative rules', supplied by the acquisition mechanism, generates intrinsic motivation to comply with those rules, detects violations, and generates intrinsic motivation to punish violators.

Elaborating on this sketch, Sripada and Stich suggested that some norms are easier than others to detect, infer, remember and/or recall because the acquisition mechanism is constrained by innate biases that are specific to social learning – learning from others, rather than direct interaction with the inanimate environment. The candidates here are 'Sperberian biases' – preferences, aversions, and emotions that make some ideas more 'attractive' than others (Sperber 1996) – and 'social learning strategies' that incline people to learn more from older, more prestigious models ('prestige bias'), or to adopt the most common cultural variant ('conformity bias'; Henrich and Gil-White 2001; Henrich and Boyd 1998; see Lewens 2015 for a recent critique of conformity bias).

Sripada and Stich assumed innate motivation to punish norm violation, driven by anger, contempt and disgust, but, like subsequent norm psychologists, stopped short of claiming there are norm-specific emotions (Kelly 2020). Their model suggests that beliefs, judgment, and explicit reasoning are parts of a semi-detached system– what I will call the 'explicit system' – shown on the right side of Figure 1[1].

3. Problems with the nativist view

Sripada and Stich's framework gave norm psychology a strong start, but its use in the last 15 years, and independent developments, have highlighted conceptual and empirical problems.

3.1 Explanatory target

Sripada and Stich defined norms as 'rules' and norm psychology as a system for processing these rules. This is troublesome because it is not clear what distinguishes a mental representation of a rule (or a rule-like mental representation) from any other kind of mental representation. Paradigmatic examples of rules are written or spoken statements such as 'Thou shalt not kill' or 'Drive on the right'; they are carved in stone, a computer, or the airwaves, not in the head. If we take the paradigmatic cases to mean that mental rules are sentences in the head, then a rule-based definition of norms presupposes what research on norm psychology is intended to discover – the nature of the psychological processes underlying normative behaviour. Instead of encouraging open enquiry about the kinds of mental representations involved in norm processing, it presupposes that normative behaviour is mediated by complex, language-like mental representations. If, like Sripada and Stich, we avoid this assumption, a rule-based definition of norms leaves us deeply uncertain about the explanatory target of norm psychology (Westra & Andrews, forthcoming). What is the nature of a 'rule' when it is *not* written, spoken or inscribed in the mind in sentential form?

This problem has not been ironed out since Sripada and Stich's seminal work was published. Following their lead, philosophers pursuing a cognitive-evolutionary approach have offered detailed, rule-based characterisations of norms (e.g., Kelly & Setman 2021). Norm psychologists from other disciplines either do not define norms or norm psychology at all (e.g., Henrich & Muthukrishna 2021), or offer something very brief such as 'learned behavioral standards shared and enforced by a community' (Chudek & Henrich, 2011, p.218), 'customary rules that govern behavior in groups or societies' (House 2018), or 'mutually accepted behavioral standards of a group' (Gockeritz, Schmidt & Tomasello 2014). Thus, 'standards' or 'regularities' (Westra & Andrews, forthcoming) sometimes take the place of 'rules' but, with no specification of what is meant by 'standards' and 'regularities', the explanatory target of norm psychology remains obscure.

3.2 Evidential relations

Norm psychologists typically do not explain for each category of evidence – universality, importance, early development, intrinsic motivation - why they take it to support the existence of innate, domain-specific mechanisms for norm processing. Sripada and Stich simply said 'it is hard to see how the facts we've assembled *could* be explained without positing innate psychological mechanisms that perform the functions we've sketched' (2006, pp 290-291). This is problematic because the evidential relations are far from obvious.

Universality. The mere presence of norms in all human societies does not imply that norms are acquired and implemented by norm-specific, genetically inherited mechanisms. In principle, normativity could be due to convergence. All human societies, confronting similar problems of cooperation, and with a common kit of domain-general psychological processes, may have found roughly the same solutions (Buckholtz & Marois 2012). This would be implausible if all societies had found exactly, rather than roughly, the same solutions. It would be implausible if norm content were universal – for example, with all societies forbidding cousin marriage and violence against women and having the same fairness norms – but norm psychologists freely acknowledge that this is not the case. It would also be implausible if norms were equally important in all societies, but research has revealed marked cross-cultural variation in the 'tightness' of social norms. Tight cultures (e.g. India, Japan) have many, strict norms and low tolerance of deviant behaviour, whereas loose cultures (e.g. the Netherlands, Ukraine) have fewer, weaker norms and greater tolerance of deviant behaviour (Gelfand et al. 2011; Gelfand 2018). Variation in tightness does not preclude innate norm psychology, but it does highlight the difficulty of making inferences from cross-cultural data to psychological mechanisms. It is not clear how much norm variation, and of what kind, would favour an empiricist over a nativist view of norm psychology, or vice versa (Sterelny 2010).

Importance. Gelfand's work on tightness suggests that norms are less important in some societies than in others. This does not undermine the general claim that normativity is an important feature of human lives, but it raises the question whether the importance of norms should incline us to believe that norm processing mechanisms are innate. Reading and writing are important in most societies but due to their recent historical origin, we know that reading and writing are not mediated by dedicated, innate mechanisms. Human characteristics can be pervasive and important without being genetically inherited. So perhaps the inference from importance to innateness depends on the observation that some norms ‘regulate matters like status, mate choice, food and sex that have a direct impact on people’s .. reproductive success’ (Sripada & Stich 2006, p.282). If so, the inference remains cryptic. If norms have an impact on reproductive fitness, there is the potential for *something* related to their processing to be favoured by genetic evolution, but that something need not be Big Special psychological processes – complex, uniquely human mechanisms dedicated to processing a specific kind of input. Instead, any norm-related genetic adaptations could be Small Ordinary – simple, quantitative modifications of mechanisms present in other animals (Heyes, 2018a, 2019a).

Early development. Chomsky (1965) made a compelling ‘poverty of the stimulus’ argument linking development with innateness. Focussing on language, he argued that we have reason to believe a characteristic is innate – that a specific propensity to develop the characteristic is genetically inherited – when it appears before children have been exposed to enough information in their environment to support development of the characteristic. For example, we would have reason to believe in an innate ‘language acquisition device’, if children’s linguistic development runs ahead of the information about language provided by their social interactions with adults.

Whatever its merits for language in particular, the poverty argument provides a solid, general basis for inferring genetic inheritance from developmental data (Heyes 2018a, 2019b). However, like subsequent norm psychologists (and unlike some moral psychologists, e.g., Hauser 2006), Sripada and Stich did not advance poverty arguments. For example, they did not consider whether there is enough information in the developmental environment for children to learn, via domain-general mechanisms, to distinguish prescriptive from descriptive rules before they are 3-5-years old (Nucci 2001). They cited ethnographic evidence that it is not necessary for children to be taught to punish norm violation – a point amplified by more recent experimental work with Western children (e.g., Schmidt & Rakoczy, forthcoming) – but, viewed as a poverty argument, this is not compelling for two reasons. First, teaching is just one way in which information can be provided through social interaction, one potential contribution to 'wealth' rather than 'poverty' of the stimulus (Ray & Heyes 2011). Adults, and other experts, can impart information about norms and normativity without intending to do so, 'leaking' their attitudes in emotionally charged behaviour and casual remarks about the actions of others (Heyes 2019b). For example, in conversation with their children over a picture book, Western parents refer more often to fairness when explaining why in-group, rather than out-group, members should be helped and not harmed (Chalik & Rhodes 2015). Second, whatever may or may not be necessary to support normative behaviour in the laboratory, there is ample anecdotal and ethnographic evidence that, in the wild, normative behaviour, such as sharing and helping, are taught in infancy and learned through social play (Lew-Levy et al. 2018). This evidence suggests wealth, rather than poverty, of the stimulus – that there may be sufficient information in children's environments to support the development of normativity. More generally, it reminds us that early development of a characteristic is not, by itself, evidence of innateness.

Intrinsic motivation. Sripada and Stich's claim about intrinsic motivation was clear. In their view, intrinsic motivation - evidenced by selfless acts in everyday life, and delivery of costly punishment in the lab – could be produced by norm-specific, genetically inherited mechanisms but not by domain-general processes. The claim was clear, but there was little argument. For example, like subsequent norm psychologists, they did not explain why, in their view, costly punishment could not be due to mistaken beliefs about the likelihood of future interactions with the same agents, generalisation from the rewarding effects of selfish punishment, or domain-general, rather than norm-specific, social learning (Seymour, Singer & Dolan 2007). Similarly, norm psychologists have not explained why we should expect motivation from these sources to be less effective than motivation arising from domain-specific, genetically inherited processes in converting normative thought into action (Hume 1748/1975).

In the last 15 years, additional evidence has been added to each of Sripada and Stich's categories but the evidential relations remain unclear. No one has spelled out why evidence of universality, importance, early development, and intrinsic motivation – independently or in combination – should encourage us to believe that humans have an innate norm psychology rather than normative competence based on domain-general cognitive processes and cultural learning.

3.3 Counter-evidence

Sripada and Stich's model remains a high point in the history of norm psychology. As far as I am aware, it is the only attempt to date to specify the mechanisms involved in norm processing, and – aside from questions about 'rules' (section 3.1) and evidential relations (section 3.3) – the specification was clear and plausible in the light of evidence available at the time. However, subsequent empirical and theoretical developments have challenged key features of the model.

3.3.1 Common is right

A substantial body of evidence now shows that children and adults conflate what is descriptively normal (common or frequent) with what is prescriptively normal (allowed or required) (e.g., Bear & Knobe 2017; Eriksson, Vartanova, Ornstein & Strimling 2021; Foster-Hanson, Roberts, Gelman & Rhodes 2021; McAuliffe, Raihani & Dunham 2017; Roberts, Ho & Gelman 2019; Roberts, Guo, Ho & Gelman 2018). For example, after being told that listening to a certain kind of music is common within a group, children between 4- and 13-years express disapproval of a group member who listens to a different kind of music, and explain their disapproval using prescriptive language (e.g., 'that's not allowed') (Roberts, Gelman & Ho 2017). When adults from 30 European countries were asked about questionable behaviours (e.g., casual sex, paying cash to avoid taxes) there was a positive correlation between their ratings of the frequency and justifiability of the behaviour (Eriksson et al. 2021). And the evidence of descriptive-prescriptive conflation is not only correlational. Adults learning a prescriptive norm from scratch, the ideal length of a fictional hunting tool called a 'stagnar', blend statistical and evaluative information in the training set. After learning, their representation of a 'normal' stagnar depends on the distribution of stagnar lengths to which they were exposed during training (descriptive), as well as information given with each exemplar about 'how good that stagnar is for hunting' (prescriptive; Bear & Knobe 2017, Study 4).

This evidence of conflation does not sit well with the assumption that prescriptive norms – the focus of norm psychology – are acquired and implemented by dedicated mechanisms. Sripada and Stich's model suggests that the acquisition mechanism is switched on by prescriptively normative behaviour, infers norms from this behaviour exclusively, and that the products of these inferences are the only entries in the execution mechanism's 'norm data base'. If this were the case, one would not expect statistical information – information about what is average, or common, rather than what is approved or ideal – to influence learning of prescriptive norms or judgements about what is allowed or justifiable. The influence of descriptive information suggests either that prescriptive norm learning is not mediated by innate, domain-specific mechanisms or that such mechanisms exist but do not function well because their operation is regularly contaminated by information from outside their proper domain (Andreoni et al. 2021)[2].

3.3.2 Domain-general learning is important

There is evidence that domain-general processes, especially reinforcement learning, play a major role in the development of normativity. For example, recent work on descriptive-prescriptive conflation shows that children and adults disapprove of atypical behaviour, not only in other people (relative to social categories), but in nonhuman animals (relative to biological categories; Foster-Hanson et al. 2021). Similarly, studies of typically developing adults in multiplayer games indicate that their learning of new norms from robot players is subject to the same biases as learning character traits from observable behaviour (a social but not normative task; Hertz 2021).

A range of formal and informal models show that reinforcement learning would be an efficient way to acquire norms (e.g., Buckholtz & Marois 2012; Ho, MacGlashan, Littman & Cushman 2017; Morris & Cushman 2018). Some of the empirical evidence implicating reinforcement learning comes from studies in which adults are asked to evaluate each of a series of stimuli (e.g., the attractiveness of faces or pieces of music), and told after each judgement whether it agreed with other people's evaluations. Information indicating agreement activates the brain's reward system, and information indicating disagreement activates areas processing punishment. In other words, a signal indicating to a person that their behaviour was or was not normative, activates the same neural mechanisms as delivery of food or money (reward), or electric shock (punishment) for button pressing in response to inanimate stimuli (Germar & Mojzisch 2019; Klucharev et al. 2009; Schnuerch & Gibbons 2014; Wu, Luo & Feng 2016). Increasing the resolution of these neurophysiological results, a recent study showed that learning about normative responses (which of two options is preferred by a social group) and about the inanimate world (points associated with selecting one of two boxes) is mediated by the same dopamine-dependent neurochemical mechanisms (Rybicki, Sowden, Schuster & Cook 2021).

A nativist norm psychologist might object that the foregoing studies relate to descriptive rather than prescriptive norms, for example, to preferences that happen to be typical of a group but that are not explicitly endorsed by group members as the right preferences to have. The findings reviewed in 3.3.1 suggest that prescriptive norms are not psychologically distinct from descriptive norms. However, for those who are not persuaded by descriptive-prescriptive conflation, there is further evidence of domain-general processing where the subject matter is squarely prescriptive or even moral. For example, psychopaths – people with a developmental disorder that interferes with moral judgement and increases the risk of antisocial behaviour – show impaired domain-general learning (Blair 2017). Similarly, a study using electrophysiological and behavioural measures confirmed that infants are more likely to approach an agent categorised by adults as 'helping' rather than 'hindering' (e.g., Hamlin, Wynn & Bloom 2010; Hamlin 2015). However, it also indicated that this preference for pro-social actors is due to the same neural mechanisms that mediate approach to and avoidance of inanimate objects (Cowell & Decety 2015; Decety & Yoder 2017). Consistent with these data, connectionist modelling indicates that the helper-hinderer effect in infants could be due to domain-general associative learning (Benton & Lapan 2021; Heyes 2019b; Scarf, Imuta, Colombo & Hayne 2012).

Nichols and colleagues have focussed most consistently on unequivocally prescriptive norms (e.g., Nichols et al. 2016; Nichols 2021; Partington, Nichols & Kushnir 2021). Answering 'moral nativists' (e.g., Hauser 2006; Levine, Leslie & Mikhail 2018), and using Bayesian modelling alongside behavioural experiments, they have shown that complex features of moral normativity could be due to domain-general processing - for example, tendencies to interpret rules as act-based rather than consequence-based, to assume that it is permissible to do things that are not explicitly prohibited, and to respond differentially to the violation of moral and conventional norms (Nichols 2021). This 'rational learning' model of morality assumes rich, genetically inherited cognitive resources – including the concepts of agent, intention, and cause – while making a powerful case that specifically moral concepts are acquired via domain-general processes of learning.

3.3.3 Social learning biases are innate *or* domain-specific

Recent work does not accord with Sripada and Stich's suggestion that 'Sperberian biases' (Sperber 1996) and 'social learning strategies' provide an opportunity for genetically inherited domain-specific psychological mechanisms to bias norm learning.

Sperberian biases have proved difficult to investigate empirically because they comprise a mixture of weakly specified learning, motivational and ecological factors (Buskell 2017). Social learning biases are more empirically tractable, and research in the last 15 years has indicated the kinds of models that children, adults and nonhuman animals are most likely to copy (Kendall et al. 2018). However, this research suggests that social learning biases are either innate and domain-general or culturally learned and domain-specific; they are not, as Sripada and Stich supposed, innate and domain-specific (Heyes 2016a, 2016b).

Consider prestige bias as an example. There is evidence that chimpanzees (Horner et al. 2010), children (Chudek, Heller, Birch and Henrich 2012; Fusaro & Harris 2008; McGuigan 2013), and adults (Henrich & Henrich 2010) are more inclined to copy high-status models, or models to whom other agents attend (there is some ambiguity about the meaning of 'prestige'; Chellappoo 2020), but careful scrutiny of these studies suggests that the prestige biases of animals and children up to about 5-years-old are due to phylogenetically ancient, domain-general attentional processes (Heyes & Pearce 2015; Heyes 2016c; 2017a; 2017b). In the simplest cases, the prestigious models are more likely to be copied because they are salient – bigger, more vocal, more likely than others to be found by gaze following. In contrast, at least some prestige bias in adulthood is due to domain-specific, deliberative, culturally inherited metacognitive processes (Heyes 2016a; Heyes, Bang, Shea, Frith & Fleming 2020)[3].

3.3.4 Costly punishment is rare and old

A central feature of nativist norm psychology is the idea that humans are intrinsically motivated to punish norm violation. Inflicting harm on a norm violator, exacting ‘retribution’, is an end-in-itself. In support of this view, Sripada and Stich cited evidence from economic games showing that people are willing to pay a price to punish norm violators both when the violation reduces the pay-off of the punisher (costly second-party punishment; Fehr & Gachter 2002) and when it reduces the pay-offs of other players (costly third-party punishment; Fehr & Fischbacher 2004); that is, when the punisher merely observes the violation. Since 2006, laboratory research using economic games has provided further evidence of costly second-party and third-party punishment in adults (e.g., Crockett, Ozdemir & Fehr 2014) and children from 6 years of age (Marshall, Yudkin & Crockett 2021; McAuliffe, Jordan & Warneken 2015). It has also indicated that retribution is a sufficient motive. For example, in the laboratory, adults engage in costly punishment even when it is ‘hidden’ – that is, the punisher believes that the violator will not be aware of having been punished for their transgression, and therefore that punishment cannot discourage the violator from repeating the transgression in future (Crockett et al. 2014).

So, the claim that people are intrinsically motivated to punish norm violation is in good shape. However, the nativist view that intrinsic motivation to punish is due to a human-specific, norm-specific mechanism is incompatible with recent evidence. Experiments using naturalistic methods, rather than economic games, indicate that costly third-party punishment occurs infrequently in Western populations when no one is watching (Balliet et al. 2021; Hofmann et al. 2014; Kriss et al. 2016; Pedersen, McAuliffe and McCullough 2018; Molho et al 2020). Alongside ethnographic evidence that people in small-scale societies rarely deliver costly punishment unless they or their kin have suffered serious harm (Boehm, 2008; Ericksen & Horton, 1992), this suggests that behaviour in economic games, where the experimenter is always watching, over-estimates the power of intrinsic motivation to punish norm violation by confounding it with the desire to please or impress the people conducting the study (Pedersen et al. 2018; 2020).

If costly third-party punishment rarely occurs among humans in the wild - outside economic games - the gap between humans and other animals is smaller than nativist norm psychologists assume. Dominant chimpanzees punish conspecifics who steal their food (Reidl, Jensen, Call & Tomasello 2012), and many animals – including wasps, mole rats, and fairy wrens – respond aggressively to conspecifics that fail to show cooperative behaviour (Clutton-Brock & Parker 1995). We do not know how often or how much this retaliatory behaviour costs the perpetrators, and therefore whether it meets a strict definition of costly punishment (Jensen 2010; Raihani et al. 2012), but research on 'appetitive aggression' (Rasa 1976) suggests that it is intrinsically motivated – that the aggressors enjoy it. For example, studies of laboratory rodents – rats, mice, and Syrian hamsters – show that they will work for the opportunity to fight with a conspecific, and develop a preference for places where they have fought in the past (Aleyasin et al. 2018).

3.4 Summary of problems

Nativist norm psychology has encountered conceptual and empirical challenges. On the conceptual side, it has become evident that defining norms with reference to 'rules', 'regularities', or 'standards' breeds uncertainty about the explanatory target of norm psychology. In addition, it is not clear why the nativist view assumes that the universality of norms, along with their importance, early development, and intrinsic motivation, is a sign that norm processing must be done by innate, domain-specific mechanisms. On the empirical front, the nativist view is incompatible with evidence that adults and children conflate descriptive and prescriptive norms; that domain-general learning plays a significant role in norm acquisition; and that the social learning biases modulating norm acquisition are either innate and domain-general, or culturally learned and domain-specific, rather than domain-specific and innate. The evidence remains strong that people are intrinsically motivated to punish norm violation, but recent work suggests that the mechanisms responsible are continuously with those found in nonhuman animals. These conceptual and empirical problems suggest that we need a new framework for norm psychology.

4 A cultural evolutionary framework for norm psychology

This section outlines a new cognitive-evolutionary framework for norm psychology (summarised in Figure 2). It proposes that, in humans, normative competence depends on domain-general psychological processes plus a culturally evolved 'cognitive gadget' (Heyes 2018a; 2019a). A cognitive gadget (cf. 'cognitive instinct', Pinker 1995) is a distinctively human, domain-specific cognitive process (or integrated set of cognitive processes) that is assembled through social interaction during childhood. At the population level, cognitive gadgets are shaped to do their jobs by cultural selection. Cultural selection is a Darwinian process of variation-and-selective-retention operating on socially inherited, rather than genetically inherited, variants (Birch 2017; Birch & Heyes 2021; Campbell 1965; Lewens 2015)[4].

Like nativist norm psychology, the cultural evolutionary or 'gadget' account assigns important roles to nature (genetic inheritance), nurture (learning), and culture (social inheritance) in the development of normativity, but the balance between them is different.

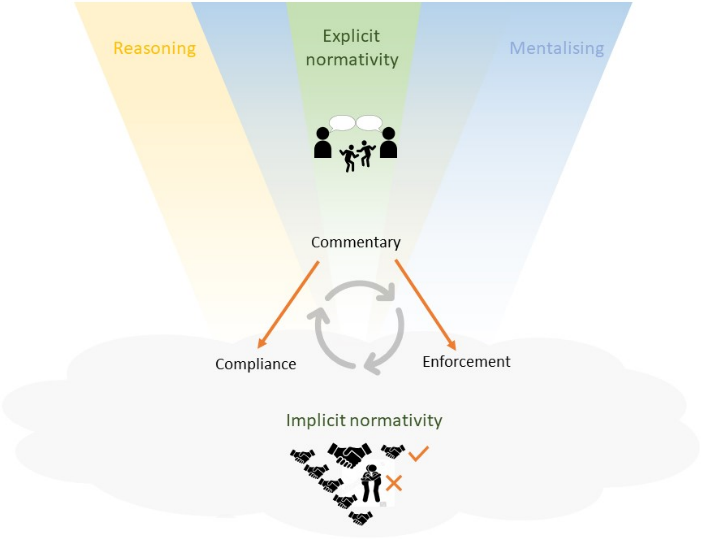

Figure 2. A cultural evolutionary framework for norm psychology. Normativity (centre) - Normativity consists of three kinds of behaviour: compliance, enforcement, and commentary. The actions that constitute compliance and enforcement within a society are defined by commentary within that society (straight arrows), not by their dependence on psychological processes specialised for norm processing. Implicit normativity (lower panel) - In early childhood, commentary is absent, and compliance and enforcement depend exclusively on implicit, domain-general, psychological processes found in a wide range of animals. Implicit acquisition processes are sensitivity to frequencies and outcomes of observed and executed actions in various settings (greeting icons). Explicit normativity (upper panel) - The development of explicit, domain-specific normative processes occurs gradually and is not complete until adulthood. Rooted in implicit compliance and enforcement, and driven by cultural learning - primarily by adult and peer commentary (speech bubble icons) – the developmental process fashions the capacity for 'ought thought' (Westra & Andrews, forthcoming), for thought about what is appropriate, allowed, required, and forbidden (green) from reasoning (yellow) and mentalising (blue). At the population level, the developmental process and the psychological mechanism it produces are shaped by cultural selection. Implicit-explicit relations (centre). Explicit normativity enables commentary and influences compliance and enforcement by modulating the behavioural outputs of implicit processes (curved arrows).

The nativist view casts genetic evolution (natural selection acting on genetic variants), and the gadget view casts cultural evolution (natural selection acting on cultural variants), as the principal architect of norm psychology – the population-level process that has shaped norm psychology and made it broadly adaptive. The culture-gene coevolutionary hypothesis behind the nativist view, suggests that socially inherited norm content was a crucial source of selection pressure for the genetic selection of normative thinking. Socially inherited norm content played a role that is filled by climatic and ecological variables in other evolutionary trajectories. In contrast, the gadget view suggests that it was selection among socially inherited variants that 'designed' the special ways in which we think about norms. The cultural domain was not just the source of selection pressure, but the site of Darwinian selection. As a corollary at the individual- rather than the population-level, the nativist view suggests that humans genetically inherit Big Special psychological resources for norm processing, such as mechanisms specialised for the detection of norm-implicating behaviour, or to infer rules from such behaviour (see section 2.3). In contrast, the gadget view suggests that the human-specific, genetically inherited psychological resources that contribute to the development of norm processing are Small Ordinary. Genetic evolution has expanded our domain-general capacities to learn and remember (Fagot & Cook 2006; Holland 1992) and tweaked our attentional and motivational processes - for example, giving us biases to attend to faces and voices (Johnson et al. 1991; Reid et al. 2017; Vouloumanos & Werker 2007) and greater sensitivity to social rewards (Floccia et al. 1997) - making us much more receptive to social information than our hominin and other primate ancestors (Heyes 2018a, Chapter 3). But, on the gadget view, this genetic 'starter kit' is not norm specific (Frith 2001; Heyes 2018a). It helps us to learn from others about everything, not just about norms.

In a complementary way, the gadget view assigns a weightier role to social inheritance, to culture, in the development of normativity. In both the nativist and gadget frameworks, norm content is learned from other people. We learn that cousin marriage is prohibited, or that hat wearing is required, by verbal instruction, observing the behaviour of others in our social group, and through exposure to the rewards and punishments they deliver when we comply with or violate the group's norms. However, on the gadget view, we also learn via these social routes how to think and feel about norms in general. We learn from other people to represent norms as a distinctive kind of regularity in nature, and emotions - including guilt, shame, and righteous indignation - that motivate compliance with and enforcement of norms (Barrett 2017; Hoemann et al. 2020).

To make comparison easier, this section, outlining the cultural evolutionary account of norm psychology, mirrors the structure of section 2, outlining the nativist evolutionary account. It starts (section 4.1) with a new characterisation of norm psychology, focussed on behaviour rather than rules, which is designed to overcome the explanatory target problem (section 3.1). The second part (4.2) addresses the problem of evidential relations (3.2) with a 'poverty-wealth' scheme for assessing evidence that a psychological capacity has been shaped by nature (or is 'innate'), nurture, and culture. The final part (4.3) sketches a new psychological model of normativity.

4.1 Characterisation - behaviour

Viewed as a natural phenomenon, norm psychology is the set of psychological processes responsible for normative behaviour. Viewed as a research project, norm psychology aims to discover the psychological processes responsible for normative behaviour. There are three types of normative behaviour: 1) Compliance – Normative agents tend to behave in ways that are approved by other members of their social group, and to avoid behaving in ways that are forbidden. 2) Enforcement – Normative agents tend to react positively to others when they behave in ways that are approved, and to react negatively when they behave in ways that are forbidden. 3) Commentary – Spontaneously and when questioned, human normative agents make judgements – they say things - about the types of behaviour that are appropriate, allowed, required, and forbidden; about what ought and ought not to be done. These three kinds of normative behaviour, or 'normativity', are closely related. For example, an agent's behaviour counts as compliant, rather than merely conformist, to the extent that it is consistent with their group's enforcement behaviour and/or commentary (House et al. 2020).

This behaviour-based characterisation of norm psychology foregrounds the traditional explanatory target of psychology – behaviour – without moving too far away from the conceptions of norms used in disciplines such as anthropology, history, philosophy, and sociology. The behaviour-based approach leaves room for us to think of norms as rules – spoken or written statements – or 'standards' that have been inferred from compliance, enforcement, and commentary behaviour by people within a society, or by scholarly observers from outside. It also leaves room for evidence that mental rules – sentences in the head – generate compliance, enforcement, and commentary behaviour, but the behaviour-based approach does not prejudge this issue. It encourages us to think freely about the psychological mechanisms that could generate normative behaviour, not only in adult humans, but in children and nonhuman animals. In principle, they could be domain-specific or domain-general or a mixture of both; they could function as a system, with high levels of interdependence, or in a more piecemeal fashion. These are empirical questions.

4.2 Evidence of nature, nurture, and culture

There is compelling evidence that normative behaviour is culturally universal, important, early developing, and intrinsically motivated (section 2.2). However, it is not clear why nativist norm psychologists take this evidence to indicate that the processes responsible for normative behaviour are innate and domain-specific (section 3.2). To create a healthy research environment, in which alternative hypotheses can be tested against one another by evidence and argument, it would be helpful to have a scheme indicating the kinds of observations that count in favour of (and against) innateness, learning and cultural inheritance of psychological characteristics.

I have argued that an expanded version of Chomsky's (1965) poverty of the stimulus argument provides such a scheme (Heyes 2018a; 2019a; Ray & Heyes 2011). This approach asks whether the developmental environment provides too little (poverty) or at least enough (wealth) usable information to explain the characteristics of a target psychological mechanism. Poverty is a sign that the development of a psychological trait depends on genetically inherited information (nature), whereas wealth is a sign that development depends on learning (nurture), and/or on culturally inherited information (culture). Both nurture and culture require learning, but in the nurture case individuals learn through their own efforts about social or asocial features of their world, whereas in the culture case they learn from others; what I learn by interaction with you depends, not just on what you do, but on what you know (Heyes 2019b). Consequently, where there is evidence of wealth, nurture is indicated if development varies with features of the environment in which development is actually occurring, i.e., with information that can be acquired by asocial learning, and by the kinds of social learning found in a broad range of animals. Culture is indicated when development varies with longer-term features of the environment; features that may not be present when a particular individual is developing or that can be acquired only via the kinds of distinctively human social learning known as cultural learning. Training studies can help to distinguish the roles of nature, nurture, and culture (e.g., De Klerk et al. 2015; Lohmann & Tomasello 2003), but most of the empirical methods with the power to parse development in this way examine patterns of spontaneous covariation. They relate differences in cognitive ability to opportunities for learning, social learning, and cultural learning across (1) time points in development, (2) groups or individuals within a human society, (3) human societies, or (4) species (Heyes, 2018a; 2019a).

Within this poverty-wealth scheme, evidence that a component of norm psychology is culturally universal (comparison across human societies) constitutes evidence of poverty of the stimulus, and therefore innateness, only if there is ancillary evidence that the universality could *not* be due to convergent learning and cultural evolution. Similarly, signs of early development (comparison across time points in development) counts as evidence of innateness only if studies of the developmental environment indicate that children do not have the opportunity to acquire the normative characteristic via domain-general mechanisms or cultural learning prior to its emergence. The other two pillars of the nativist case, importance and intrinsic motivation, do not have an obvious place within the poverty-wealth evidential framework. Given that cultural evolution can produce important psychological characteristics (e.g., literacy), and that intrinsic motivation can be learned (see section 3.2.2), it is not clear why these characteristics of normativity have been seen as evidence of innate, domain-specific normative processing.

The poverty-wealth scheme suggests that two types of evidence have been under-exploited by norm psychologists. Variation among groups and individuals within a society (2 above), and across species (4 above), could provide important clues to the origins of norm psychology, but is rarely mentioned by those who identify as norm psychologists (Westra & Andrews, forthcoming).

4.3 Model

The cultural evolutionary model suggests that two types of psychological mechanism support normative behaviour, implicit and explicit. Research on decision-making and cognitive control indicates that implicit processes are typically fast, automatic, associative, effortless, and non-conscious. In contrast, explicit processes are typically slower, controlled, rule-based, effortful, and conscious. Explicit processes do, and implicit processes do not, tend to interfere with one another (Evans & Stanovich 2013a; 2013b; Norman & Shallice 1986). The implicit processes supporting early development of normative behaviour are domain-general and taxon-general; they are perceptual, attentional, learning, memory, motoric, and motivation processes that support normative and non-normative behaviour via the same computations in a wide range of animal species. These implicit processes, although domain-general, could be described as innate. Their development is canalized and depends on genetically inherited information and experience (nature and nurture) but minimally or not at all on culturally inherited information (culture). The explicit processes are domain-specific. They represent the expectations of others in concepts – recombinable elements of conscious, deliberate thinking (Shea 2018) – and language-like mental rules about what is appropriate, allowed, required, or forbidden in different contexts for various members of a social group. Most of these rules, and the capacity to reason with them in distinctive ways, are learned from other people via language and other forms of cultural learning. The innate, domain-general implicit processes get normative behaviour off the ground, phylogenetically and ontogenetically. They are solely responsible for compliance and enforcement in early childhood and continue to play important roles throughout adult life. However, explicit processes become increasingly influential in middle childhood. They support commentary behaviour, and influence both compliance and enforcement by modulating the behavioural outputs of implicit processing. Commentary behaviour – statements about what is appropriate, allowed, required, and forbidden – has synchronic and diachronic functions. Within generations, commentary enables negotiation of norm content (e.g., attitudes to gay marriage), and debate about what constitutes norm violation (e.g., through gossip and legal processes; Jolly & Chang 2021). Between generations, commentary facilitates cultural learning of norm content and, recursively, of the explicit normative processes that make commentary possible.

'Two-system' or 'dual-process' theories, postulating implicit and explicit or 'automatic' and 'controlled processes, are common in moral psychology (Cowell, Calma-Birling & Decety 2018, Crockett 2013, Cushman 2013, Greene 2017, Rhodes & Wellman 2017) but not in norm psychology (see Kelly 2020 for an interesting exception). The following sections outline the ways that implicit (4.3.1) and explicit (4.3.2) processes give rise to normative behaviour – compliance, enforcement, and commentary. They refer to psychological processes that are modelled in a variety of ways in contemporary cognitive science. For example, as associative or statistical learning; model-free or model-based reinforcement learning; mental models or pragmatic schemas; assuming inverse or forward ('predictive') processing. The cultural evolutionary account of normativity is not committed to particular modelling strategies. Rather, it is committed to the view that mature human normativity is 1) rooted in implicit psychological processes that, however they are modelled, do many jobs in many species, and 2) distinguished from nonhuman normativity by cognitive and motivational features that are culturally learned.

4.3.1 Implicit processes

Compliance. In infancy and early childhood, compliance depends on domain-general processes of categorisation and reinforcement learning (Ayub & Wagner 2020; Foster-Hanson et al. 2021; Morris & Cushman 2018; Schlegelmilch et al. 2021; Wellman, Kushnir & Brink 2016). For example, in adult commentary 'giving' is more normatively loaded than 'reaching', but young children learn to categorise a variety of different body movements as (what an adult would call) *giving*, and that giving has positive outcomes in many contexts (e.g., hugs), in the same way that they learn to treat a variety of different body movements as *reaching*, and that reaching has positive outcomes in many contexts (e.g., toys; Pulverman et al. 2006). Direct social rewards, like receiving a hug or a smile, are especially important when learning compliance, but direct experience is augmented by observation of others' behaviour in both normative and non-normative cases (Haaker et al. 2021). Young children can learn when to give by watching another child give a toy and get a hug, and they can learn via the same process to reach by watching another child reach and secure a toy. Social learning of this kind, which also occurs in rats (Saggerson & Honey 2006) and birds (Heyes & Saggerson 2002; Saggerson et al. 2005), is made possible by 'conditioned' or ‘secondary’ reinforcement (Williams 1994). The sight of another agent’s action acquires positive or negative value for the observer when it has been positively correlated with similar, direct experience – for example, when the learner has eaten while seeing others eat, or felt distress while observing the distress of others (Heyes, 2018b).

Once established, compliant behaviour may be maintained by habit. For example, children in Victorian society may have learned to stand whenever an adult entered a room by tracking social rewards and punishments, but, with repetition, standing became a reflex response to adult arrival (Adams 1982; Pool et al. 2021). It would have been difficult for them to stop if they had been told not to do it. Compliant behaviour can also be maintained by a type of conditioned reinforcement that amounts to intrinsic motivation (Kruglanski et al. 2018). During the acquisition of a compliant behaviour, the feeling-of-doing the action – for example, the sensations from her own body and from the world that a child feels when she rises from a seated to a standing position – are repeatedly correlated with extrinsic rewards, attention and smiles from others, in a distinctive setting, the arrival of an adult. Consequently, the feeling-of-doing – the stimulus change resulting from action execution - becomes rewarding in that setting (Osborne 1977). The child is no longer motivated by the expectation that she will get attention or other extrinsic rewards. She enjoys performing the action for its own sake[5].

Enforcement. Early enforcement builds on early compliance and, like early compliance, depends on domain-general processes (Seymour, Singer & Dolan 2007). We know from research on the conflation of descriptive and prescriptive norms (section 3.3.1) that behaviours receiving normative approval in adult commentary are more common, and therefore on average more familiar and predictable, than behaviours that are not approved. Decades of research on the 'mere exposure effect' shows that humans and other animals like familiar stimuli – images, sounds, tastes, and smells they have encountered without reinforcement – more than unfamiliar stimuli (Bornstein & Craver-Lemley 2016). Conversely, prediction error, produced by the occurrence of unexpected events, is associated with increased arousal, which is typically aversive (Theriault et al. 2021). Therefore, much enforcement behaviour, in children and throughout life, is likely to be very simple indeed. An individual observing familiar, normative behaviour feels good and is therefore more likely to be friendly towards the actor, whereas an individual observing unfamiliar, non-normative behaviour feels bad and is therefore more likely to be unfriendly. Evidence of this pattern comes from research showing that, when emotional states are induced by unrelated videos, happy people donate more and punish less than angry people (Drouvelis & Grosskopf 2016). In addition, studies of aversion-induced aggression indicate that humans, like a wide range of other animals (e.g., hamsters, gophers, monkeys, cats, chickens, snakes, and turtles), respond aggressively, not only when provoked by the behaviour of another agent (Clutton-Brock & Parker 1995), but also – underlining the domain-generality of the underlying mechanisms – when in pain, stressed, or uncomfortably warm (Groves & Anderson 2018). This kind of 'reactive aggression' – in contrast with goal-directed, 'proactive aggression' - is a highly conserved trait (Wrangham 2018).

The effect of familiarity-based enforcement behaviour is to reward the normative behaviour of others and to punish their counter-normative behaviour. That is what makes it 'enforcement'. However, the implicit processes mediating familiarity-based enforcement do not categorise observed action as normative or non-normative. They just register how often and how recently the behaviour has been observed under similar circumstances in the past. Furthermore, these implicit processes do not represent the potential effects of the enforcer's behaviour on the target. The enforcer does not act with the intention of rewarding, punishing, or having any effect at all on the target's behaviour. In this sense, familiarity-based enforcement is intrinsically rather than instrumentally motivated.

Of course, reinforcement learning, in which outcomes are important, also contributes to early enforcement. It would be strange if infants could learn via reinforcement mechanisms to control the movements of toys (Kenward et al. 2009), and videos (Klossek, Russell & Dickinson 2008), but not the movements of other people. It would be odd if, for example, infants could not learn that creating a physical obstruction or saying 'Stop!', can make another actor stop doing something unpleasantly novel, and start doing something that is pleasingly familiar, or an action that is associated in the enforcer's mind with positive extrinsic rewards. When adult commentary would class the unpleasantly novel behaviour as 'bad' or 'wrong' and the alternative as 'good' or 'right', the enforcing child may also get a warm response (social reward) from an observing adult, making her yet more inclined to enforce familiarity in future.

Recent research on 'normative protest' can be used to illustrate how implicit processes yield early norm enforcement. This fascinating work suggests that children begin to show enforcement when they at 18 months (Schmidt, Rakoczy & Tomasello 2019) or 3 years of age (Schmidt et al. 2016). Children at these ages observe an adult model performing an action on an object and are then either given access to the object themselves or allowed to observe a puppet moving the object in the same way as the adult or in a different way.

When children manipulate the object themselves, they typically repeat the adult's action. When the puppet moves the object in a different way, the children often intervene by putting the object in the location where it was placed by the adult, or saying things like 'No!', 'Stop!' or 'Must go in there!' (Schmidt, Rakoczy & Tomasello 2019). Implicit processes explain this pattern of results: 1) The children copy the model because that yields a more familiar outcome. They may have seen the action more than once, but it was just seconds or minutes ago. Also, they are inclined to copy an adult because copying has been rewarded in similar contexts in the past – for example, when an adult behaves without hesitation in full view of the child. In other words, they copy in the expectation of a good outcome – social approval, or a warm feeling due to conditioned reinforcement (Miller & Dollard 1941; Schein 1954). 2) When the puppet moves the object in a different way, the children experience aversive prediction error, and anticipate omission of the expected rewarded - that they will miss out on the approval or warm feeling. Consequently, the protestors behave in an unfriendly way towards the puppet and intervene to restore familiarity.

What should we make of the protestors' utterances? I am interpreting them as instrumental responses; assuming that, by 18-36 months, some children have discovered via reinforcement learning that 'No!', 'Stop!' and 'must', are 'pushy words', words that are apt to change the behaviour of another agent. In principle, these utterances could instead be commentary behaviour, rooted in explicit normative rules, reflecting some understanding of what is expected within the social group and approved by its members. That is possible but unlikely given that the children in these experiments are given no evidence that the adult's behaviour is descriptively or prescriptively normal. They see the behaviour performed once, by one person in one context, and without affirmation. If their utterances are commentary, their normativity is perverted rather than ‘promiscuous’ (Schmidt et al., 2016), untethered from its proper domain.

This interpretation of early normative protest as the product of implicit processing could be tested against the standard interpretation by behavioural experiments equating the familiarity of the puppet’s ‘normative’ and ‘counter-normative’ actions and varying the children’s experience of reward for action copying. In the meantime, the implicit interpretation is consistent with electroencephalographic (EEG) data from a study in which preschool children observed an adult performing a counter-normative behaviour (ripping a page out of a library book). When looking at pictures of harming, the children who protested the adult’s transgression differed from those who did not protest on an early ERP component (central P2) that indexes perceptual sensitivity and sustain attention, described as ‘implicit moral evaluation’ (Kim, Decety, Wu, Baek & Sankey 2021).

Thus, implicit normativity is intrinsically as well as extrinsically (or instrumentally) motivated, and its development depends in part on simple forms of social learning – on vicarious as well as direct experience of rewards and punishments delivered by other agents. The implicit processes outlined above are thoroughly normative in that they typically bring the agent’s behaviour (compliance), and the behaviour of their social partners (enforcement), into conformity with one another and with what is approved by normative commentary in the agent’s society. But they are not dedicated to that function. In other contexts, the same implicit processes do completely different jobs – for example, enabling the development of food preferences and the acquisition of motor skills. Furthermore, implicit processes are not rule-based in any interesting sense. They are not necessarily or exclusively based on stored exemplars of behaviour (Sripada and Stich’s criterion), and the behavioural regularities they produce can be described by rules, such as ‘Children should stand when an adult enters a room’, but these are very low bars for rule-hood. Even the learning of simple categories, such as ‘bird’, does not depend necessarily or exclusively on stored exemplars (Schlegelmilch, Wills & Helversen 2021), and any regularity in nature, anything that is not random, can be described from the outside by a rule. Implicit normativity does not have rules on the inside; it is not produced by rule-like mental structures, by sentences in the head.

4.3.2 Explicit processes

Explicit normativity begins to augment and interact with implicit normativity when children start to represent, not only what others do*,* but what others expect to be done (Bicchieri 2006; Theriault et al. 2021). As explicit processes emerge, a child who has learned to share toys in anticipation that sharing will make other people do rewarding things, such as smile, begins to appreciate that the people around them expect sharing in some contexts, and that their reactions to the child’s behaviour in these contexts – smiling, grabbing, reprimanding – depend on their expectations. Similarly, a child who previously repeated adult actions only because the repetition yielded pleasant feelings of familiarity, or because it made other people smile, begins to see repetition as expected and social rewards and punishments as contingent on this expectation.

The explicit processes that contribute to normative behaviour are often labelled ‘deontic reasoning’, a distinctive type of deductive reasoning in which information about what is appropriate, allowed, required, or forbidden is represented by proposition-like ‘mental models’ (Ragni, Kola & Johnson-Laird 2018) or ‘pragmatic schemas’ (Holyoak and Cheng

1995). In contemporary cognitive science, models of deontic reasoning are pitched at a high level of abstraction (Beller 2010). They capture core semantic and syntactic features of the explicit processes, but rarely acknowledge that the rules of deontic reasoning represent social facts, the expectations of others, and therefore depend on 'social understanding' (Carpendale & Lewis 2015).

Some processes of social understanding, known as 'explicit mentalising', 'mindreading' or 'theory of mind', represent mental states. They represent an expectation as something inside an individual's head, that arises from the individual's experience and influences their behaviour. Other explicit processes of social understanding encode behavioural rules, such as 'Drive on the right' and 'Help members of your group', and the situations in which these rules apply. They represent an expectation as something that resides in a group, situation, or institution rather than in the minds of individual agents. Cross-cultural evidence suggests that both kinds of explicit process contribute to normativity in all human societies, but cultures vary widely in the extent to which they rely on mental state attribution rather than behavioural rules (Taumoepeau 2019). For example, 38% of 12- to 14-year-old ni-Vanuatu children from Nguna Island 'fail' a classic false-belief test of mentalising that is typically 'passed' by 4- to 5-year-olds in the United States (Dixson et al. 2018). In a complementary way, Japanese children given the same test - in which children are asked to predict where a protagonist will look for an object that was moved in their absence – are more likely than children in Europe and North America to explain their prediction with reference to behavioural rules and situational factors rather than mental states (Naito & Koyama 2006).

The greater reliance of 'interdependent' or 'relational' cultures on behavioural rules suggests that, in these societies, explicit normativity depends more heavily on rules than in 'individualistic' cultures. For example, the helping behaviour of a person of European heritage may be more likely than that of a Japanese person to spring from an explicit but amorphous belief that others expect helping – a belief rooted in their prior experience in similar circumstances, but not encoded in an explicit, sentence-like mental structure specifying what is appropriate, allowed, required, or forbidden. This is consistent with evidence from Western samples that a good deal of our normative commentary, of our statements about what is and is not right, is post hoc rationalisation. We can usually formulate rules and engage in deontic reasoning when questioned, but our normative behaviour is generated by implicit processes (e.g., Haidt 2001; 2012).

There is ample evidence that explicit normativity is not 'in our genes'. This evidence indicates wealth of the stimulus and implicates cultural learning – human-specific forms of social learning - in the development of explicit normativity (see section 4.2; Dunn 1994; Dunn & Hughes 2014; Grusec 2014; Grusec et al. 2014; Wright & Bartsch 2008). For example, in Western samples there is a positive correlation between the number of auxiliary modal verbs used by children and the number used by their parents (Wells 1979). Six-year-old children, who usually cannot extract prescriptive messages from stories, become able to do so when they explain the story to an adult (Walker & Lombrozo 2017). The normative development of 10- to 15-year-olds, measured by their comments on a series of moral dilemmas, is predicted by the quality of their spontaneous conversation with parents and peers about normative questions. A 'Socratic style of eliciting the other's opinion and checking for understanding', and conversational focus on the child's experience of moral conflict, are especially effective in promoting normative development (Walker, Hennig & Krettenauer 2000, p.1045). From 5- to 86-years-of-age, normative development measured by Kohlberg's tests of moral reasoning (Colby et al. 1987) is linearly related to number of years of formal education (Dawson 2002).

Evidence of this kind suggests that 'ought thought' (Westra & Andrews, forthcoming) - explicit psychological processes specialised for normativity – is concocted from more domain-general mentalising and reasoning processes through social interaction in the course of development (Cushman et al. 2013; Hawkins et al. 2019; Ho, MacGlashan, Littman & Cushman 2017; Imuta et al. 2016; Rhodes & Wellman 2017; Sterelny 2021). The explicit processes of deontic reasoning used by some minorities, for example lawyers, ethicists, and moral philosophers in Western societies, have been constructed by 'intelligent design' (Dennett 2009); like algebra or calculus, they have been fashioned deliberately by generations of scholars with the express purpose of finding the right way of thinking about what is right. The rest of us use explicit normative processes that have been cobbled together primarily by cultural selection. Insofar as these processes do the jobs traditionally associated with normativity - promoting cooperation and true values – it is because groups with an explicit norm psychology that did those jobs relatively well passed on their explicit norm psychology, via cultural learning, to a larger number of descendants (Birch & Heyes 2021).

4.3.3 Implicit – explicit relations

Implicit normativity provides a foundation for the development of explicit normativity (Rhodes & Wellman 2017). A child without a repertoire of compliance and enforcement behaviour, acquired by implicit processes, would struggle to get the message about others' normative expectations. Even if she developed mentalising and reasoning, it is unlikely they would coalesce into culture-typical explicit normativity because there would be little of personal relevance to be explained by the expectations of others. A rule about helping makes sense, not only of what a child sees others doing, but also of her own implicit feelings and behaviour. It explains (or rationalises) why she sometimes helps when she does not want to, the uneasy feeling when she fails to help, and, pleasingly, casts her feelings when others fail to help as righteous indignation. Explicit processes interpret interoceptive experiences. They transform feelings of variable intensity (arousal), that are merely good or bad (valence), into full-blown emotions such as shame, guilt, and moral rage – the intrinsic motivators of explicit normativity (Barrett 2017; Theriault et al. 2021). Guilt and shame are forms of self-punishment that make my compliance less dependent others' enforcement (Frith & Frith forthcoming).

In some cases, the acquisition of specialised, explicit cognitive processes has profound effects on domain-general implicit processes. For example, learning algebra changes perception (Marghetis, Landy & Goldstone 2016). The extent to which explicit normativity transforms implicit normativity remains a contentious empirical question. In principle, explicit processes could take over entirely, or do nothing more than support normative commentary, providing post hoc justification for actions and judgements caused by implicit processes (Haidt 2001; 2012). Not surprisingly, current evidence suggests an intermediate reality in which explicit normative processes augment and regulate, but do not replace, implicit normative processes in generating compliance and enforcement behaviour. Evidence of regulation comes from the behaviour of adults playing economic trust games, where explicit processes, activated by instructions about another player's reputation or moral character, suppress subsequent reinforcement learning (Delgado et al. 2005; Fouragnan et al. 2013). Similarly, developmental research indicates that, from 6 years-of-age in Western samples, explicit or 'controlled' normative processes can mediate sharing (compliance), costly punishment (enforcement) and intent-based normative judgement (commentary) (Chernyak & Kushner 2018; Cushman et al. 2013; McAuliffe, Jordan & Warneken 2015; Steinbeis 2018).

Further evidence of the top-down influence of explicit processes comes from studies manipulating belief in free will (Frith & Frith forthcoming), a normatively charged belief with marked cultural variation (Berniūnas et al. 2021). Laboratory studies with Western samples show that people who have been induced to doubt the existence of free will are less helpful and more aggressive (Baumeister et al. 2009), more likely to cheat in exams (Vohs & Schooler 2008), and, crucially, less likely to show post-error slowing in a simple reaction time task (Rigoni et al. 2013). The effects of the belief manipulation on helping, aggression and cheating could be due to explicit normative thinking alone, but post-error slowing is due to a paradigmatically implicit process (Dutilh et al. 2012).

There is also evidence that implicit, domain-general mechanisms of categorisation and learning continue to have an unregulated or minimally regulated effect in adult life (Bear & Knobe 2017; Burton-Chellew & Guerin 2021; see section 3.3). For example, people playing economic games are more likely to show prosocial giving behaviour when they have been exposed, in the lab or their ordinary lives, to institutions that are effective in punishing behaviour that deviates from giving norms. However, exposure to these institutions does not make people more likely to punish other players who violate giving norms. This suggests that institutions have narrow effects, the kind one would expect if they produce behaviour change via implicit reinforcement learning. If institutions were causing conceptual change – for example, development of an explicit belief that 'giving is right' – one would expect them to influence both giving behaviour and punishment of those who fail to give (Stagnaro, Arechar & Rand 2017).

Explicit processes can also become implicit. Like driving a car, patterns of normative thought that were once deliberative - conscious, effortful – can become automatic with intensive practice (Pizarro & Bloom 2003). Rules such as *Do not tamper with nature* and *Acts are worse than omissions* can 'go underground', becoming 'heuristics' or 'intuitions' with a pervasive influence on behaviour that is unexplained by ('moral dumbfounding'; Haidt 2001), or at odds with, the actor's normative commentary (Sunstein 2005). Unlike the implicit processes that initiate normative development, these heuristics are domain-specific – they apply only to what one ought and ought not to do – but their origins lie, not in the genes, but in a socially inherited apparatus of normative thinking.

5. Conclusion

5.1 Objections

A common and potentially powerful objection to cognitive gadgets says that they would have become cognitive instincts (e.g., Del Guidice 2019; Dor, Ginsburg & Jablonka 2019; Turner & Walmsley 2021). Even if distinctively human cognitive mechanisms were at first socially inherited and shaped by cultural selection, a process variously known as

‘Baldwinisation’ (Baldwin 1896), ‘canalization’ (Gottlieb 1991) and ‘genetic assimilation’ (Waddington 1953) would have favoured genetic variants that reduced the environmental input necessary for their development. In this way, the erstwhile gadget would be subjected to genetic selection and, over many generations, become innate - genetically rather than culturally inherited.

Plausibility arguments and modelling do not tell us whether Baldwinisation is likely to have affected gadgets in general or normativity in particular. It is plausible that normative behaviour has been important to human survival and reproduction for long enough, measured in biological generations, for the Baldwinisation of norm-specific cognitive mechanisms. However, it is also plausible that social inheritance of norm-specific mechanisms was cheap and reliable enough that genetic mutations with the potential to reduce environmental input conferred little or no selective advantage (Morgan et al. 2020). A more promising approach, pursued incisively by Turner and Walmsley (2021), looks to empirical research on ‘preparedness’ for evidence of Baldwinisation. This research is widely believed to show that taste aversion (Garcia & Koelling 1966) and fear learning (Seligman 1970) depend on genetically evolved, domain-specific learning mechanisms - processes that forge associations via distinctive computations when, for example, a ‘fear relevant’ object such as a snake, rather than a ‘fear-irrelevant’ object such as a flower, is experienced with an aversive stimulus (Barrett & Broesch 2012). In principle, any domain-specific features of prepared learning could be due to standard or ‘aplastic’ (Morgan et al. 2020) genetic evolution, but studies of experimental evolution in *Drosophila* indicate that preparedness is more likely to be due to Baldwinisation (Mery & Kawecki 2002).

Empirical work on preparedness is exactly the right place to look for evidence that cognitive gadgets would have been Baldwinised but it does not deliver. Careful scrutiny of 50 years of research on preparedness confirms that taste aversion and fear learning are 'special', but also indicates that genetic evolution has improved their efficiency only by tweaking perceptual and motor processes, and their attentional and motivational modulators (Heyes, Chater & Dwyer 2020). For example, fear-relevant stimuli such as snakes are genetically primed to attract attention (Gray & McNaughton 2003; Thrasher & LoBue 2016). This gives them privileged access to learning mechanisms, but those mechanisms use the same computations to learn about snakes and flowers. Contrary to early claims (Hugdahl & Ohman 1977), fear learning extinguishes at the same rate as learning about fear-irrelevant stimuli (Åhs et al. 2018), and is equally malleable by instructions (McNally 1987; 2016). Evidence of this kind - relating to fear learning, taste aversion learning, language and imitation – indicates Baldwinisation of input and output devices, analogues of scanner and printer interfaces, not of core inference processes. Cognitive gadgets are core inference processes. Therefore, research on preparedness does not support the objection that cognitive gadgets would have been Baldwinised. On the contrary, it supports the idea that genetic selection provided a starter kit for the evolution of human minds, not by fashioning Big Special cognitive mechanisms, such as dedicated processes for learning and implementing norms, but by tweaking input and output processes, such as social tolerance and motivation (see section 4; Heyes 2018a; 2019a).

5.2 Implications

There are three key contrasts between the cultural evolutionary and nativist accounts of norm psychology. On the gadget account: 1) The heavy lifting is done by domain-general rather than norm-specific processes. 2) These domain-general, implicit processes track the frequencies and outcomes of behaviour; they do not represent what others expect or what is allowed, as mental rules or in any other way. 3) Explicit processes are rule-based and norm-specific but culturally rather than genetically inherited.

To find out which account is closer to the truth, and in what ways, we need more empirical work of two kinds. First, to establish whether the early development of normative behaviour is guided by the maturation of domain-specific or domain-general processes, we need more experiments that test these alternative hypotheses against one another. At present, developmental studies by nativist norm psychologists typically test one rich domain-specific hypothesis against another, rather than against a leaner domain-general alternative. For example, they ask whether an infant wants to help an adult or to engage with the adult, not whether the child (or, more precisely, their neurocognitive system) is working to restore the predictability of their environment. On the other hand, domain-generalists often rely on modelling to show that normative jobs could be done efficiently by processes that are not specialised for the task. This kind of modelling is valuable but to move from 'how possibly' to 'how actually' we need data.

Second, to discover to what extent explicit normativity is culturally rather than genetically inherited, we need more research guided by the poverty-wealth scheme (section 4.2). While nativist norm psychology was the only evolutionary framework available, those pursuing a cognitive-evolutionary account of normativity hardly needed to appeal to the poverty of the stimulus. In most cases, it was obvious that no child would achieve mature normative competence if they had to work it all out for themselves. Now there is a cultural evolutionary alternative to the nativist view – a framework acknowledging the need for inheritance of normative thinking but arguing that the inheritance is social – it is apparent that evidence of universality, importance, early development, and intrinsic motivation (see section 2.2) is a blunt instrument. It implicates inheritance but it does not tell us what kind of inheritance, genetic or cultural, is important. Research guided by the poverty-wealth scheme would chart more fully variation in normative reasoning across time points in childhood development, groups or individuals within a society, human societies, and species, and, crucially, trace the sources of this variation to genetic factors and to opportunities for learning, social learning, and cultural learning. At present, there is a lingering assumption that normative reasoning shows quantitative but not qualitative variation; that some individuals and societies are better at it than others, but normative concepts, such as 'obligation', always have the same functional roles (Beller 2010). Similarly, it is assumed that across individuals and cultures normative reasoning depends on mental models *or* pragmatic schemas, and involves a standard blend of mentalising and behaviour rules. If we let go of these assumptions – rooted in nativist norm psychology, and the view that Western, educated, industrialised, rich and democratic societies (WEIRD; Henrich 2020) are psychologically representative of all humanity – the cultural evolutionary account predicts that we will find rich variation, not just in what people believe is right, but in the explicit processes they use to think about rightness. Tracing this variation to its sources, a crucial step in distinguishing genetic from cultural inheritance (Sterelny 2021), would involve a new kind of behavioural genetics that charts experiential inputs to development, opportunities for social and cultural learning, just as carefully as genetic inputs.

What about normativity in the wild? The cultural evolutionary account implies that normative behaviour is less constrained and less secure than the nativist view suggests. The implicit processes generating compliance and enforcement in early development and throughout our lives did not evolve specifically to promote cooperative behaviour. Therefore, they are not constrained genetically to produce behaviour that, on average and over long periods of time, has prosocial effects at the population level. Acting on implicit processes alone, a person will comply with and enforce any behaviour that is common and yields positive outcomes in their experience. It does not matter whether, in the long term, the behaviour promotes or interferes with the thriving of their narrow social group – the people with whom they interact – or of the wider society in which they live. Implicit processes also allow an individual's compliance and enforcement behaviour to change relatively rapidly with experience. Habit formation and intrinsic motivation create some inertia (4.3.1), but there is no equivalent among implicit processes of Sripada and Stich's (2006) 'norm data base' (Figure 1), where normative rules are safely insulated from change after acquisition. To the extent that normativity depends on implicit processes, acquisition and implementation are continuous in two respects: they are not distinct processes, and the representations mediating normative behaviour are continually subject to revision by new experience.

Because the implicit processes are not constrained by nature, there is more pressure on specialised explicit normative processes to produce and maintain 'good' behaviour. However, the proper development of these processes is less assured on the cultural evolutionary than on the nativist account. Instead of being programmed in our genes, the development of explicit normative thought is dependent on social practices and institutions that can change rapidly and radically with political and economic conditions. Furthermore, the evidence from science and everyday life that moral judgements often depend on 'emotional' or 'Type 1' processes (e.g., Haidt 2001; 2012), suggests that, even when explicit normativity has developed in a group-typical way, it often struggles to get a grip on behaviour. In short, the culturally evolutionary account implies that contemporary individuals and societies – parents, peers, educators, elders, politicians, and lawyers – have more responsibility than the nativist view implies. Our actions don't just shape and transmit the rules, they create in each new generation mental processes that can grasp the rules and put them into action.

References

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 34B*(2), 77-98.

Åhs, F., Rosén, J., Kastrati, G., Fredrikson, M., Agren, T., & Lundström, J. N. (2018). Biological preparedness and resistance to extinction of skin conductance responses conditioned to fear relevant animal pictures: A systematic review. *Neuroscience & Biobehavioral Reviews*, *95*, 430-437.

Aleyasin, H., Flanigan, M. E., & Russo, S. J. (2018). Neurocircuitry of aggression and aggression seeking behavior: nose poking into brain circuitry controlling aggression. *Current opinion in neurobiology*, *49*, 184-191.

Andreoni, J., Nikiforakis, N., & Siegenthaler, S. (2021). Predicting social tipping and norm change in controlled experiments. *Proceedings of the National Academy of Sciences*, *118*(16).

Andrews, K. (2020). Naïve normativity: The social foundation of moral cognition. *Journal of the American Philosophical Association*, *6*(1), 36-56.

Ayub, A., & Wagner, A. R. (2020). What am I allowed to do here? Online Learning of Context-Specific Norms by Pepper. In *International Conference on Social Robotics* (pp. 220-231). Springer, Cham.

Baldwin, J.M. (1896) A new factor in evolution. *American Naturalist*, 30, 441–451.

Balliet, D., Molho, C., Columbus, S., & Cruz, T. D. D. (2021). Prosocial and punishment behaviors in everyday life. *Current Opinion in Psychology*.

Barrett, L. F. (2017). *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt.

Barrett, H. C., & Broesch, J. (2012). Prepared social learning about dangerous animals in children. *Evolution and Human Behavior*, *33*(5), 499-508.

Baumeister, R.F., Masicampo, E.J., and Dewall, C.N. 2009. Prosocial benefits of feeling free: disbelief in free will increases aggression and reduces helpfulness. *Personality & Social Psychology Bulletin,* 35, 260-268.

Batson, C. D. (2014). *The Altruism Question: Toward a Social-psychological Answer*. Psychology Press.

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition*, *167*, 25-37.

Beller, S. (2010). Deontic reasoning reviewed: psychological questions, empirical findings, and current theories. *Cognitive Processing*, *11*(2), 123-132.

Benton, D. T., & Lapan, C. (2021). Moral masters or moral apprentices? A connectionist account of sociomoral evaluation in preverbal infants. https://doi.org/10.31234/osf.io/mnh35

Berniūnas, R., Beinorius, A., Dranseika, V., Silius, V., & Rimkevičius, P. (2021). The weirdness of belief in free will. *Consciousness and Cognition*, *87*, 103054.

Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Bicchieri, C., & McNally, P. (2018). Shrieking sirens: Schemata, scripts, and social norms. How change occurs. *Social Philosophy and Policy*, *35*(1), 23-53.

Birch, J. (2017). *The Philosophy of Social Evolution*. Oxford University Press.

Birch, J. (2021). Toolmaking and the evolution of normative cognition. *Biology & Philosophy*, *36*(1), 1-26.

Birch, J., & Heyes, C. (2021). The cultural evolution of cultural evolution. *Philosophical Transactions of the Royal Society B*, *376*(1828), 20200051.

Blair, R. J. R. (2017). Emotion-based learning systems and the development of morality. *Cognition*, *167*, 38-45.

Bornstein, R. F., & Craver-Lemley, C. (2016). Mere exposure effect. In *Cognitive Illusions* (pp. 266-285). Psychology Press.

Boyd, R. (2016). *A different kind of animal: how culture made humans exceptionally adaptable and cooperative*. Princeton, NJ: Princeton University Press.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, *100*(6), 3531-3535.

Boyd, R., & Richerson, P. J. (2001). Norms and bounded rationality. *Bounded rationality: The adaptive toolbox*, 281-296.

Boyd, R., & Richerson, P. J. (2005). *The Origin and Evolution of Cultures*. Oxford University Press.

Brown, D. (1991). *Human Universals,* New York: McGraw-Hill.

Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, *15*(5), 655-661.

Burton-Chellew, M. N., & Guérin, C. (2021). Decoupling cooperation and punishment in humans shows that punishment is not an altruistic trait. *Proceedings of the Royal Society B*, *288*(1962), 20211611.

Buskell, A. (2017). What are cultural attractors? *Biology & Philosophy*, *32*(3), 377-394.

Campbell, D. T. (1965). Variation and selective retention in socio-cultural evolution. *Social Change in Developing Area*.

Carpendale, J. I., Kettner, V. A., & Audet, K. N. (2015). On the nature of toddlers' helping: Helping or interest in others' activity? *Social Development*, *24*(2), 357-366.

Carpendale, J. I., & Lewis, C. (2015). The development of social understanding.

Chalik, L., & Rhodes, M. (2015). The communication of naïve theories of the social world in parent–child conversation. *Journal of Cognition and Development*, *16*(5), 719-741.

Chellappoo, A. (2020). Rethinking prestige bias. *Synthese*, 1-22.

Chernyak, N., & Kushnir, T. (2018). The influence of understanding and having choice on children's prosocial behavior. *Current Opinion in Psychology*, *20*, 107-110.

Chomsky, N. (1965). *Aspects of the Theory of Syntax* (Vol. 11). MIT press.

Chudek, M., Heller, S., Birch, S., & Henrich, J. (2012). Prestige-biased cultural learning: Bystander's differential attention to potential models influences children's learning. *Evolution and Human Behavior*, *33*(1), 46-56.

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

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature*, *373*(6511), 209-216.

Colby, A., Kohlberg, L., Speicher, B., Hewer, A., Candee, D., Gibbs, J., & Power, C. (1987). *The measurement of moral judgement: Volume 2, Standard issue scoring manual* (Vol. 2). Cambridge university press.

Colombo, M. (2014). Two neurocomputational building blocks of social norm compliance. *Biology and Philosophy* 29 (1): 71–88.

Cosmides, L., & Tooby, J. (1989). Evolutionary psychology and the generation of culture, Part II. Case study: a computational theory of social exchange. *Ethology & Sociobiology*, 10, 51–97.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides & J. Tooby (Eds.), *The adapted mind: evolutionary psychology and the generation of culture* (pp. 163–228). New York: Oxford University Press.

Cowell, J. M., Calma-Birling, D., & Decety, J. (2018). Domain-general neural computations underlying prosociality during infancy and early childhood. *Current Opinion in Psychology*, *20*, 66-71.

Cowell, J. M., & Decety, J. (2015). The neuroscience of implicit moral evaluation and its relation to generosity in early childhood. *Current Biology*, *25*(1), 93-97.

Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, *17*(8), 363-366.

Crockett, M. J., Özdemir, Y., & Fehr, E. (2014). The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology: General*, *143*(6), 2279.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273-292.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6-21.

Dawson, T. L. (2002). New tools, new insights: Kohlberg's moral judgement stages revisited. *International Journal of Behavioral Development*, *26*(2), 154-166.

Decety, J., & Yoder, K. J. (2017). The emerging social neuroscience of justice motivation. *Trends in Cognitive Sciences*, *21*(1), 6-14.

de Klerk, C. C., Johnson, M. H., Heyes, C. M., & Southgate, V. (2015). Baby steps: investigating the development of perceptual–motor couplings in infancy. *Developmental Science*, *18*(2), 270-280.

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*(11), 1611-1618.

Del Giudice, M. (2019). Cognitive gadgets: A provocative but flawed manifesto. *Behavioral and Brain Sciences*, *42*, e174.

Dennett, D. (2009). Darwin's ‘strange inversion of reasoning’. *Proceedings of the National Academy of Sciences*, *106*(Supplement 1), 10061-10065.

Dixson, H. G., Komugabe-Dixson, A. F., Dixson, B. J., & Low, J. (2018). Scaling theory of mind in a small-scale society: A case study from Vanuatu. *Child Development*, *89*(6), 2157-2175.

Dor, D., Ginsburg, S., & Jablonka, E. (2019). The evolution of cultural gadgets. *Mind & Language*, *34*(4), 518-529.

Drouvelis, M., & Grosskopf, B. (2016). The effects of induced emotions on pro-social behaviour. *Journal of Public Economics*, *134*, 1-8.

Dunn, J. (1994). Changing minds and changing relationships. *Children's early understanding of mind: Origins and development*, 297-310.

Dunn, J., & Hughes, C. (2014). Family talk about moral issues: The toddler and preschool years. *Talking about right and wrong: Parent-child conversations as contexts for moral development*, 21-43.

Durkheim, E. (1912/1968). *The Elementary Forms of the Religious Life*. Allen & Unwin.

Dutilh, G., Vandekerckhove, J., Forstmann, B.U., Keuleers, E., Brysbaert, M. & Wagenmakers, E.-J. (2012). Testing theories of post-error slowing. *Attention, Perception, & Psychophysics* 74, 454-465.

Eriksson, K., & Coultas, J. C. (2014). Corpses, maggots, poodles and rats: emotional selection operating in three phases of cultural transmission of urban legends. *Journal of Cognition and Culture*, *14*(1-2), 1-26.

Eriksson, K., Vartanova, I., Ornstein, P., & Strimling, P. (2021). The common-is-moral association is stronger among less religious people. *Humanities and Social Sciences Communications*, *8*(1), 1-8.

Evans, J. S. B., & Stanovich, K. E. (2013a). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223-241.

Evans, J. S. B., & Stanovich, K. E. (2013a). Theory and metatheory in the study of dual processing: Reply to comments. *Perspectives on Psychological Science*, *8*(3), 263-271.

Fagot, J., & Cook, R. G. (2006). Evidence for large long-term memory capacities in baboons and pigeons and its implications for learning and the evolution of cognition. *Proceedings of the National Academy of Sciences*, *103*(46), 17564-17567.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63-87.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137-140.

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, *2*(7), 458-468.

Fitzpatrick, S. (2020). Chimpanzee normativity: evidence and objections. *Biology & Philosophy*, *35*(4), 1-28.

Floccia, C., Christophe, A., & Bertoncini, J. (1997). High-amplitude sucking and newborns: The quest for underlying mechanisms. *Journal of Experimental Child Psychology*, *64*(2), 175-198.

Foster-Hanson, E., Roberts, S. O., Gelman, S. A., & Rhodes, M. (2021). Categories convey prescriptive information across domains and development. *Journal of Experimental Child Psychology*, *212*, 105231.

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational priors magnify striatal responses to violations of trust. *Journal of Neuroscience*, *33*(8), 3602-3611.

Frank, R. (1988). *Passions Without Reason*. Norton.

Frith, U. (2001). Mind blindness and the brain in autism. *Neuron*, *32*(6), 969-979.

Frith, C. & Frith, U. (forthcoming) Chapter 10: Getting Along Together.

Fusaro, M., & Harris, P. L. (2008). Children assess informant reliability using bystanders' non-verbal cues. *Developmental Science*, *11*(5), 771-777.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science*, *4*(1), 123-124.

Gavrilets, S., & Richerson, P. (2017). Collective action and the evolution of social norm internalization. *Proceedings of the National Academy of Sciences*, 114 (23), 6068-6073.

Gelfand, M. J. (2018). *Rule makers, rule breakers: How tight and loose cultures wire our world*. Scribner.

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., ... & Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, *332*(6033), 1100-1104.

Germar, M., & Mojzisch, A. (2019). Learning of social norms can lead to a persistent perceptual bias: a diffusion model approach. *Journal of Experimental Social Psychology*, *84*, 103801.

Göckeritz, S., Schmidt, M. F., & Tomasello, M. (2014). Young children's creation and transmission of social norms. *Cognitive Development*, *30*, 81-95.

Gottlieb, G. (1991) Experiential canalization of behavioral development: theory. *Developmental Psychology*, 27, 4–13.

Gray, J.A. and McNaughton, N. (2003) *The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System*. Oxford University Press

Greene, J. D. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition*, *167*, 66-77.

Gross, J. & Vostroknutov, A. (2021). Why do people follow social norms? *Current Opinion in Psychology* https://doi.org/10.1016/j.copsyc.2021.08.016

Groves, C. L., & Anderson, C. A. (2018). Aversive events and aggression. *Current Opinion in Psychology*, *19*, 144-148.

Grusec, J. E. (2014). 14 Parent–child conversations from the perspective of socialization theory. *Talking about right and wrong: Parent-child conversations as contexts for moral development*, 334.

Grusec, J. E., Chaparro, M. P., Johnston, M., & Sherman, A. (2014). The development of moral behavior from a socialization perspective.

Haaker, J., Diaz-Mataix, L., Guillazo-Blanch, G., Stark, S. A., Kern, L., LeDoux, J. E., & Olsson, A. (2021). Observation of others' threat reactions recovers memories previously shaped by firsthand experiences. *Proceedings of the National Academy of Sciences*, *118*(30).

Haidt, J. (2001). The emotional dog and its rational tale: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Haidt, J. (2012). *The Righteous Mind: Why good people are divided by politics and religion*. Vintage.

Hamlin, J. K. (2015). The case for social evaluation in preverbal infants: gazing toward one's goal drives infants' preferences for Helpers over Hinderers in the hill paradigm. *Frontiers in Psychology*, *5*, 1563.

Hamlin, J. K., Wynn, K., & Bloom, P. (2010). Three-month-olds show a negativity bias in their social evaluations. *Developmental Science*, *13*(6), 923-929.

Hauser, M. (2006): *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. New York: HarperCollins.

Hawkins, R. X., Goodman, N. D., & Goldstone, R. L. (2019). The emergence of social norms and conventions. *Trends in Cognitive Sciences*, *23*(2), 158-169.

Henrich, J. (2015). *The Secret of Our Success*. Princeton University Press.

Henrich, J. (2020). *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous*. Penguin UK.

Henrich, J., & Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior*, *19*(4), 215-241.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., & Gintis, H. (Eds.). (2004). *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*. OUP Oxford.

Henrich, J., & Ensminger, J. (2014). Theoretical foundations: The coevolution of social norms, intrinsic motivation, markets, and the institutions of complex societies. In J. Ensminger & J. Henrich (Eds.), Experimenting with social norms: Fairness and punishment in cross-cultural perspective (p. 19–44). Russell Sage Foundation.

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, *22*(3), 165-196.

Henrich, J., & Henrich, N. (2010). The evolution of cultural adaptations: Fijian food taboos protect against dangerous marine toxins. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1701), 3715-3724.

Henrich, J., & Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annual Review of Psychology*, *72*, 207-240.

Hertz, U. (2021). Learning how to behave: cognitive learning processes account for asymmetries in adaptation to social norms. *Proceedings of the Royal Society B*, *288*(1952), 20210293.

Heyes, C. M. (2016a). Who knows? Metacognitive social learning strategies. *Trends in Cognitive Sciences,* 20, 204-213.

Heyes, C. M. (2016b). Blackboxing: social learning strategies and cultural evolution. *Philosophical Transactions of the Royal Society: B,* 371, 20150369.

Heyes, C. M. (2016c). Born pupils? Natural pedagogy and cultural pedagogy. *Perspectives on Psychological Science,* 11, 280-295.

Heyes, C. M. (2017a). When does social learning become cultural learning? *Developmental Science,* 20, e12350.

Heyes, C. M. (2017b). Rattling the cage and opening the door. *Developmental Science,* 20, e12416.

Heyes, C. M. (2018a). *Cognitive Gadgets: The Cultural Evolution of Thinking*. Harvard University Press.

Heyes, C. M. (2018b). Empathy is not in our genes. *Neuroscience & Biobehavioural Reviews*, 95, 499-507.

Heyes, C.M. (2019a) Précis of Cognitive Gadgets: The Cultural Evolution of Thinking. *Behavioral and Brain Sciences,* 42, e169:1-5.

Heyes, C.M. (2019b) Is morality a gadget? Nature, nurture and culture in moral development. *Synthese* https://doi.org/10.1007/s11229-019-02348-w

Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing ourselves together: The cultural origins of metacognition. *Trends in Cognitive Sciences*, *24*(5), 349-362.

Heyes, C. M. & Pearce, J. M. (2015). Not-so-social learning strategies. *Proceedings of the Royal Society of London:B., 282, 20141709.*

Heyes, C., & Saggerson, A. (2002). Testing for imitative and nonimitative social learning in the budgerigar using a two-object/two-action test. *Animal Behaviour*, *64*(6), 851-859.

Ho, M. K., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition*, *167*, 91-106.

Hoemann, K., Wu, R., LoBue, V., Oakes, L. M., Xu, F., & Barrett, L. F. (2020). Developing an understanding of emotion categories: Lessons from objects. *Trends in Cognitive Sciences*, *24*(1), 39-51.

Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral punishment in everyday life. *Personality and Social Psychology Bulletin*, *44*(12), 1697-1711.

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*(6202), 1340-1343.

Holland, P. C. (1992). Occasion setting in Pavlovian conditioning. In *Psychology of Learning and Motivation* (Vol. 28, pp. 69-125). Academic press.

Holyoak, K. J., & Cheng, P. W. (1995). Pragmatic reasoning with a point of view. *Thinking & Reasoning*, *1*(4), 289-313.

House, B. R. (2018). How do social norms influence prosocial development? *Current Opinion in Psychology*, *20*, 87-91.

House, B.R., Kanngiesser, P., Barrett, H.C., Broesch, T., Cebioglu, S., Crittenden, A.N., Erut, A., Lew-Levy, S., Sebastian-Enesco, C., Smith, A.M. and Yilmaz, S., 2020. Universal norm psychology leads to societal diversity in prosocial behaviour and development. *Nature Human Behaviour*, *4*(1), pp.36-44.

Hugdahl, K., & Öhman, A. (1977). Effects of instruction on acquisition and extinction of electrodermal responses to fear-relevant stimuli. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(5), 608.

Hume, D., 1748/1975, *An Enquiry Concerning the Principles of Morals*, Oxford: Clarendon Press.

Imuta, K., Henry, J. D., Slaughter, V., Selcuk, B., & Ruffman, T. (2016). Theory of mind and prosocial behavior in childhood: A meta-analytic review. *Developmental Psychology*, *52*(8), 1192.

Jensen, K. (2016). Prosociality. *Current Biology*, *26*(16), R748-R752.

Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, *40*(1-2), 1-19.

Jolly, E., & Chang, L. J. (2021). Gossip drives vicarious learning and facilitates social connection. *Current Biology*.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American Psychologist*, *58*(9), 697.

Kaufman, S. B., DeYoung, C. G., Reis, D. L., & Gray, J. R. (2011). General intelligence predicts reasoning ability even for evolutionarily familiar content. *Intelligence*, *39*(5), 311-322.

Kelly, D. (2020). Internalized Norms and Intrinsic Motivation: Are Normative Motivations Psychologically Primitive. *Emotion Review*, 36-45.

Kelly, D. (2020). Two ways to adopt a norm. *The Oxford Handbook of Moral Psychology. Oxford University Press, New York*.

Kelly, D., & Davis, T. (2018). Social norms and human normative psychology. *Social Philosophy and Policy*, *35*(1), 54-76.

Kelly, D., & Morar, N. (2020). Bioethical ideals, actual practice, and the double life of norms. *The American Journal of Bioethics*, *20*(4), 86-88.

Kelly, D. & Setman, S. (2021). The Psychology of Normative Cognition, *The Stanford Encyclopedia of Philosophy,* Edward N. Zalta (ed.). https://plato.stanford.edu/archives/spr2021/entries/psychology-normative-cognition

Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social learning strategies: Bridge-building between fields. *Trends in Cognitive Sciences*, *22*(7), 651-665.

Kenward, B. (2012). Over-imitating preschoolers believe unnecessary actions are normative and enforce their performance by a third party. *Journal of Experimental Child Psychology*, *112*(2), 195-207.

Kenward, B., Folke, S., Holmberg, J., Johansson, A., & Gredebäck, G. (2009). Goal-directedness and decision making in infants. *Developmental Psychology, 45*(3), 809-819.

Kim, M., Decety, J., Wu, L., Baek, S., & Sankey, D. (2021). Neural computations in children's third-party interventions are modulated by their parents' moral values. *NPJ science of learning*, *6*(1), 1-13.

Klossek, U. M. H., Russell, J., & Dickinson, A. (2008). The control of instrumental action following outcome devaluation in young children aged between 1 and 4 years. *Journal of Experimental Psychology: General*, *137*(1), 39.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, *61*(1), 140-151.

Kohlberg, L. (1981). The philosophy of moral development moral stages and the idea of justice.

Kriss, P. H., Weber, R. A., & Xiao, E. (2016). Turning a blind eye, but not the other cheek: On the robustness of costly punishment. *Journal of Economic Behavior & Organization*, *128*, 159-177.

Kruglanski, A. W., Fishbach, A., Woolley, K., Bélanger, J. J., Chernikova, M., Molinario, E., & Pierro, A. (2018). A structural model of intrinsic motivation: On the psychology of means-ends fusion. *Psychological Review*, *125*(2), 165.

Lawrence, A. D., Calder, A. J., McGowan, S. W., & Grasby, P. M. (2002). Selective disruption of the recognition of facial expressions of anger. *Neuroreport*, *13*(6), 881-884.

Levine, S., A.M. Leslie and J. Mikhail. (2018). The mental representation of human action. *Cognitive Science*, 42: 1229–64.

Lewens, T. (2015). *Cultural evolution: Conceptual Challenges*. OUP Oxford.

Lew-Levy, S., Lavi, N., Reckin, R., Cristóbal-Azkarate, J., & Ellis-Davies, K. (2018). How do hunter-gatherer children learn social and gender norms? A meta-ethnographic review. *Cross-Cultural Research*, *52*(2), 213-255.

Lipton, P. (2003). *Inference to the Best Explanation*. London: Routledge.

Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. *Child Development*, *74*(4), 1130-1144.

Macchi, L., Caravona, L., Poli, F., Bagassi, M., & Franchella, M. A. (2020). Speak your mind and I will make it right: the case of 'selection task'. *Journal of Cognitive Psychology*, *32*(1), 93-107.

Manktelow, K. I., & Over, D. E. (1990). *Inference and understanding: A philosophical and psychological perspective*. Taylor & Frances/Routledge.

Marghetis, T., Landy, D., & Goldstone, R. L. (2016). Mastering algebra retrains the visual system to perceive hierarchical structure in equations. *Cognitive research: principles and implications*, *1*(1), 1-10.

Marshall, J., Yudkin, D. A., & Crockett, M. J. (2021). Children punish third parties to satisfy both consequentialist and retributive motives. *Nature Human Behaviour*, *5*(3), 361-368.

Marwell, G., & Ames, R. E. (1981). Economists free ride, does anyone else? Experiments on the provision of public goods, IV. *Journal of Public Economics*, *15*(3), 295-310.

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition*, *134*, 1-10.

McAuliffe, K., Raihani, N. J., & Dunham, Y. (2017). Children are sensitive to norms of giving. *Cognition*, *167*, 151-159.

McGuigan, N. (2013). The influence of model status on the tendency of young children to over-imitate. *Journal of Experimental Child Psychology*, *116*(4), 962-969.

McNally, R.J. (1987) Preparedness and phobias: a review. *Psychological Bulletin*, 101, 283–303.

McNally, R.J. (2016) The legacy of Seligman's 'Phobias and Preparedness' (1971). Behavior Therapy, 47, 585–594

Mery, F., & Kawecki, T. J. (2002). Experimental evolution of learning ability in fruit flies. *Proceedings of the National Academy of Sciences*, *99*(22), 14274-14279.

Mikhail, J. (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment.* Cambridge University Press.

Miller, N. E., & Dollard, J. (1941). *Social Learning and Imitation.* Yale University Press.

Molho, C., Tybur, J. M., Van Lange, P. A. M. & Balliet, D. (2020) Direct and indirect punishment of norm violations in daily life. *Nature Communications*, 11, 3432.

Morgan, T.J.H. et al. (2020) What the Baldwin effect affects depends on the nature of plasticity. *Cognition*, 197, 104165

Morris, A., & Cushman, F. (2018). A common framework for theories of norm compliance. *Social Philosophy and Policy*, *35*(1), 101-127.

Naito, M., & Koyama, K. (2006). The development of false-belief understanding in Japanese children: Delay and difference? *International Journal of Behavioral Development*, *30*(4), 290-304.

Nichols, S. (2021). The case for moral empiricism. *Analysis*, *81*(3), 549-567.

Nichols, S., Kumar, S., Lopez, T., Ayars, A., & Chan, H. Y. (2016). Rational learners and moral rules. *Mind & Language*, *31*(5), 530-554.

Norman, D. A., & Shallice, T. (1986). Attention to action. In *Consciousness and Self-regulation* (pp. 1-18). Springer, Boston, MA.

Nucci, L. P. (2001). *Education in the Moral Domain*. Cambridge University Press.

Oberliessen, L., & Kalenscher, T. (2019). Social and non-social mechanisms of inequity aversion in non-human animals. *Frontiers in Behavioral Neuroscience*, *13*, 133.

O'Neill, E., & Machery, E. (2018). The normative sense: What is universal? What Varies? In A. Zimmerson, K. Jones, & M. Timmons (Eds.), *The Routledge Handbook of Moral Epistemology.* Routledge Press.

Osborne, S. R. (1977). The free food (contra-freeloading) phenomenon: A review and analysis. *Animal Learning & Behavior*, *5*(3), 221-235.

Partington, S., Nichols, S., & Kushnir, T. (2021). Is children's norm learning rational? A meta-analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43, No. 43).

Pedersen, E. J., McAuliffe, W. H., & McCullough, M. E. (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. *Journal of Experimental Psychology: General*, *147*(4), 514.

Pedersen, E. J., McAuliffe, W. H., Shah, Y., Tanaka, H., Ohtsubo, Y., & McCullough, M. E. (2020). When and why do third parties punish outside of the lab? A cross-cultural recall study. *Social Psychological and Personality Science*, *11*(6), 846-853.

Pinker, S. (1995). *The Language Instinct: The New Science of Language and Mind* (Vol. 7529). Penguin UK.

Pizarro, D. A. & Bloom, P. (2003) The intelligence of the moral intuitions: Comment on Haidt. *Psychological Review,* 110(1), 193–98.

Pletti, C., Scheel, A., & Paulus, M. (2017). Intrinsic Altruism or Social Motivation—What Does Pupil Dilation Tell Us about Children's Helping Behavior? *Frontiers in Psychology*, *8*, 2089.

Pool, E., Gera, R., Fransen, A., Perez, O., Cremer, A., Aleksic, M., . . . O'Doherty, J. (2021). *Determining the effects of training duration on the behavioral expression of habitual control in humans: a multi-laboratory investigation*. PsyArXiv Preprints.

Press, C., Yon, D. & Heyes, C. (2022). Building better theories. *Current Biology*

Pulverman, R., Hirsh-Pasek, K., Golinkoff, R. M., Pruden, S., & Salkind, S. (2006). Conceptual foundations for verb learning: Celebrating the event. *Action meets word: How children learn verbs*, *2010*, 134-159.

Ragni, M., Kola, I., & Johnson-Laird, P. N. (2018). On selecting evidence to test hypotheses: A theory of selection tasks. *Psychological Bulletin*, *144*(8), 779.

Raihani, N. J., Thornton, A., & Bshary, R. (2012). Punishment and cooperation in nature. *Trends in Ecology & Evolution*, *27*(5), 288-295.

Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: young children's awareness of the normative structure of games. *Developmental Psychology*, *44*(3), 875.

Ray, E. D. & Heyes, C. M. (2011) Imitation in infancy: The wealth of the stimulus. *Developmental Science*, 14, 92-105.

Raymond, L., Kelly, D., & Hennes, E. P. (2021). Norm-Based Governance for Severe Collective Action Problems: Lessons from Climate Change and COVID-19. *Perspectives on Politics*, 1-14.

Rasa, O. A. E. (1976). Aggression: Appetite or aversion? An ethologist's viewpoint. *Aggressive Behavior*, *2*(3), 213-222.

Reid, V. M., Dunn, K., Young, R. J., Amu, J., Donovan, T., & Reissland, N. (2017). The human fetus preferentially engages with face-like visual stimuli. *Current Biology*, *27*(12), 1825-1828.

Rhodes, M., & Wellman, H. (2017). Moral learning as intuitive theory revision. *Cognition*, *167*, 191-200.

Richerson, P., Baldini, R., Bell, A. V., Demps, K., Frost, K., Hillis, V., ... & Zefferman, M. (2016). Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. *Behavioral and Brain Sciences*, *39*.

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2012). No third-party punishment in chimpanzees. *Proceedings of the National Academy of Sciences*, *109*(37), 14824-14829.

Riehl, C., & Frederickson, M. E. (2016). Cheating and punishment in cooperative animal societies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1687), 20150090.

Rigoni, D., Wilquin, H., Brass, M., and Burle, B. 2013. When errors do not matter: Weakening belief in intentional control impairs cognitive reaction to errors. *Cognition,* 127, 264-269.

Roberts, S. O., Gelman, S. A., & Ho, A. K. (2017). So it is, so it shall be: Group regularities license children's prescriptive judgments. *Cognitive Science*, *41*, 576-600.

Roberts, S. O., Guo, C., Ho, A. K., & Gelman, S. A. (2018). Children’s descriptive-to-prescriptive tendency replicates (and varies) cross-culturally: Evidence from China. *Journal of Experimental Child Psychology*, *165*, 148-160.

Roberts, S. O., Ho, A. K., & Gelman, S. A. (2019). The role of group norms in evaluating uncommon and negative behaviors. *Journal of Experimental Psychology: General*, *148*(2), 374.

Rozin, P. (1986). One-trial acquired likes and dislikes in humans: Disgust as a US, food predominance, and negative learning predominance. *Learning and Motivation*, *17*(2), 180-189.

Rybicki, A., Sowden, S., Schuster, B., & Cook, J. L. (2021). Dopaminergic challenge dissociates learning from primary versus secondary sources of information.

Saggerson, A. L., George, D. N., & Honey, R. C. (2005). Imitative learning of stimulus-response and response-outcome associations in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, *31*(3), 289.

Saggerson, A. L., & Honey, R. C. (2006). Observational learning of instrumental discriminations in the rat: the role of demonstrator type. *Quarterly Journal of Experimental Psychology*, *59*(11), 1909-1920.

Scarf, D., Imuta, K., Colombo, M., & Hayne, H. (2012). Social evaluation or simple association? Simple associations may explain moral reasoning in infants. *PLoS One*. 2012;7(8): e42698.

Schein, E. H. (1954). The effect of reward on adult imitative behavior. *The Journal of Abnormal and Social Psychology*, *49*(3), 389.

Scherer, K. R. (1997). The role of culture in emotion-antecedent appraisal. *Journal of Personality and Social Psychology*, *73*(5), 902.

Schlegelmilch, R., Wills, A. J., & von Helversen, B. (2021). A cognitive category-learning model of rule abstraction, attention learning, and contextual modulation. *Psychological Review*.

Schmidt, M. F., Butler, L. P., Heinz, J., & Tomasello, M. (2016). Young children see a single action and infer a social norm: Promiscuous normativity in 3-year-olds. *Psychological Science*, *27*(10), 1360-1370.

Schmidt, M. F. H., & Rakoczy, H. (forthcoming). On the uniqueness of human normative attitudes. In K. Bayertz & N. Roughley (Eds.), *The normative animal? On the anthropological significance of social, moral and linguistic norms.* Oxford University Press.

Schmidt, M. F., Rakoczy, H., & Tomasello, M. (2019). Eighteen-month-old infants correct non-conforming actions by others. *Infancy*, *24*(4), 613-635.

Schnuerch, R., & Gibbons, H. (2014). A review of neurocognitive mechanisms of social conformity. *Social Psychology*.

Seligman, M. E. (1970). On the generality of the laws of learning. *Psychological Review*, *77*(5), 406.

Seymour, B., Singer, T., & Dolan, R. (2007). The neurobiology of punishment. *Nature Reviews Neuroscience*, *8*(4), 300-311.

Shea, N. (2018). *Representation in Cognitive Science*. Oxford University Press.

Sperber, D. (1996). *Explaining Culture: A Naturalistic Approach*. Cambridge, MA: Cambridge, 1101.

Sperber, D., & Girotto, V. (2002). Use or misuse of the selection task? Rejoinder to Fiddick, Cosmides, and Tooby. *Cognition*, *85*(3), 277-290.

Sripada, C. S. and Stich, S. (2006). A Framework for the Psychology of Norms', *The Innate Mind, Volume 2: Culture and Cognition,* Peter Carruthers, Stephen Laurence, and Stephen Stich (eds.), New York: Oxford University Press, 280–301.

Stagnaro, M. N., Arechar, A. A., & Rand, D. G. (2017). From good institutions to generous citizens: Top-down incentives to cooperate promote subsequent prosociality but not norm enforcement. *Cognition*, *167*, 212-254.

Steinbeis, N. (2018). Taxing behavioral control diminishes sharing and costly punishment in childhood. *Developmental Science*, *21*(1), e12492.

Sterelny, K. (2010). Moral nativism: A sceptical response. *Mind & Language*, *25*(3), 279-297.

Sterelny, K. (2021). *The Pleistocene Social Contract: Culture and Cooperation in Human Evolution*. Oxford University Press.

Sunstein, C. R. (2005). Moral heuristics. *Behavioral and Brain Sciences*, *28*(4), 531-541.

Taumoepeau, M. (2019). Culture, communication and socio-cognitive development: understanding the minds of others. *Children's Social Worlds in Cultural Context*, 41-54.

Theriault, J. E., Young, L., & Barrett, L. F. (2021). The sense of should: A biologically-based framework for modeling social pressure. *Physics of Life Reviews*, *36*, 100-136.

Thrasher, C., & LoBue, V. (2016). Do infants find snakes aversive? Infants' physiological responses to "fear-relevant" stimuli. *Journal of Experimental Child Psychology*, *142*, 382-390.

Turner, C. R., & Walmsley, L. D. (2021). Preparedness in cultural learning. *Synthese*, *199*(1), 81-100.

Valiña, M. D., & Martín, M. (2016). The influence of semantic and pragmatic factors in Wason's selection task: State of the art. *Psychology*, *7*(06), 925.

Veissière, S. P., Constant, A., Ramstead, M. J., Friston, K. J., & Kirmayer, L. J. (2020). Thinking through other minds: A variational approach to cognition and culture. *Behavioral and brain sciences*, *43*.

Vohs, K.D., and Schooler, J.W. (2008). The value of believing in free will: encouraging a belief in determinism increases cheating. *Psychological Science,* 19, 49-54.

Vouloumanos, A., & Werker, J. F. (2007). Listening to language at birth: Evidence for a bias for speech in neonates. *Developmental Science*, *10*(2), 159-164.

Waddington, C.H. (1953) Genetic assimilation of an acquired character. *Evolution*, 7, 118–12

Walker, L. J., Hennig, K. H., & Krettenauer, T. (2000). Parent and peer contexts for children's moral reasoning development. *Child Development*, *71*(4), 1033-1048.

Walker, C. M., & Lombrozo, T. (2017). Explaining the moral of the story. *Cognition*, *167*, 266-281.

Wellman, H. M., Kushnir, T., Xu, F., & Brink, K. A. (2016). Infants use statistical sampling to understand the psychological world. *Infancy*, *21*(5), 668-676.

Wells G (1979). Learning and using the auxiliary verb in English. In: Lee V (ed) Language development. Croom Helm, London, pp 250–270.

Westra, E. & Andrews, K. (forthcoming). A new framework for the psychology of norms.

Williams, B. A. (1994). Conditioned reinforcement: Experimental and theoretical issues. *The Behavior Analyst*, 17, 261-285.

Wilson, D. S., & O’Gorman, R. (2003). Emotions and actions associated with norm-breaking events. *Human Nature*, *14*(3), 277-304.

Wilson, D. S., & Sober, E. (1998). *Unto Others*. Harvard University Press, Cambridge, 9, 95-112.

Wrangham, R. W. (2018). Two types of aggression in human evolution. *Proceedings of the National Academy of Sciences*, *115*(2), 245-253.

Wright, J. C., & Bartsch, K. (2008). Portraits of early moral sensibility in two children's everyday conversations. *Merrill-Palmer Quarterly (1982-)*, 56-85.

Wu, H., Luo, Y., & Feng, C. (2016). Neural signatures of social conformity: A coordinate-based activation likelihood estimation meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *71*, 101-111.

Zefferman, M. R. (2014). Direct reciprocity under uncertainty does not explain one-shot cooperation, but demonstrates the benefits of a norm psychology. *Evolution and Human Behavior*, *35*(5), 358-367.

Endnotes

[1] According to Sripada and Stich, the explicit system may be dedicated to moral rules of the sort that have preoccupied philosophers and some developmental psychologists (e.g., Kohlberg 1981), or it may be domain-general in the sense of dealing with factual as well as normative rules.

[2] To accommodate evidence of prescriptive-descriptive conflation, a nativist norm psychologist might argue that, in ancestral environments there were conditions in which it was adaptive for the norm acquisition mechanism to overshoot, making false positive judgements. However, without specification of what these conditions were, and evidence that they correspond with those in which contemporary actors 'mistake' descriptive for prescriptive norms, this would be a fudge (Lipton 2003). Alternatively, a norm psychologist might take prescriptive-descriptive conflation as a sign that domain-specific norm acquisition mechanisms make use of innate social learning strategies, such as conformist bias. Since conformist bias is thought to have a pervasive effect on social learning, this would be a significant retreat from the claim that norm processing depends on dedicated mechanisms. Given evidence that distinctively human social learning strategies are culturally inherited (see section 3.3.3.), it would also run contrary to the claim that norm psychology is innate.

[3] Long before 'norm psychology' was born, Cosmides and Tooby (1989; 1992) carved a nativist path to norms with their research on 'cheater detection' and 'social exchange reasoning'. Curiously, this work is rarely cited by those who now identify as norm psychologists. This may be because there is a broad consensus among psychologists who

study reasoning, ‘evolutionary psychologists’ and others, that an innate module for cheater detection cannot satisfactorily explain why people perform better on some versions of the Wason selection task than on others (Beller 2010; Kaufman et al. 2011; Macchi et al. 2020; Manktelow & Over 1990; Ragni et al. 2018; Valina & Martin 2016; Sperber & Girotto 2002).

[4] In contrast with other cognitive gadgets – including imitation, mentalising (also known as ‘mindreading’ and ‘theory of mind’), and metacognition (Heyes et al. 2021) - motivational processes are important ingredients of the normativity gadget. It could be called a ‘cognitive-motivational’ gadget.

[5] The idea that compliance in infancy and early childhood is due to implicit, domain-general processes is consistent with recent evidence that early helping is motivated by a desire to engage with other people rather than to promote their welfare (Carpendale, Kettner & Audet, 2015; Pletti et al. 2017).