12 May 2021, for *Analyse & Kritik*

Refining the Skill Hypothesis: Replies to Andrews & Westra, Tomasello, Sterelny, and Railton

Jonathan Birch
Department of Philosophy, Logic and Scientific Method,
London School of Economics and Political Science,
Houghton Street, London, WC2A 2AE, UK.
j.birch2@lse.ac.uk
http://personal.lse.ac.uk/birchj1

Abstract

I reflect on the commentaries on my "skill hypothesis" from Andrews & Westra, Tomasello, Sterelny, and Railton. I discuss the difference between normative cognition and the broader category of action-guiding representation, and I reflect on the relationship between joint intentionality and normative cognition. I then consider Sterelny and Railton's variants on the skill hypothesis, which highlight some important areas where future evidence could help us refine the account: the relative importance of on-the-fly skill execution vs. longer-term strategizing, the relative importance of toolmaking vs. collaborative foraging, and the question of whether norms are encoded in control models themselves or in the goals and ideals that our control models help us pursue.

I am very grateful to Kristin Andrews & Evan Westra (2021), Michael Tomasello (2021), Kim Sterelny (2021) and Peter Railton (2021) for their excellent responses to my "skill hypothesis" regarding the evolution of normative cognition (Birch 2021a, b). Their commentaries are so rich and interesting that I cannot respond to every point here. I will focus on a handful of points that struck me as particularly important.

Pushing outwards: is action-guiding representation already enough?

I may as well start by explaining why I doubt that normative cognition is present in bees, despite **Westra & Andrews**' interesting positive case. It is not because I doubt the cognitive sophistication of bees. The bee brain represents aspects of the environment, and those representations guide action. One intensively studied example is path integration, or dead reckoning, which appears to be achieved by means of an inner representation in the brain's central complex that resembles the traverse boards once used by mariners for the same task (Stone et al. 2017; the analogy with a traverse board is from a talk by Barbara Webb). These inner traverse boards are not cognitive control models, since they don't represent the causal structure of skill execution, but they are still action-guiding representations.

The inner traverse board guides flight, but does it encode a norm for correct navigation? Prediction errors will be registered, but are prediction errors *traced to their causes* in a way that allows for the registration of a difference between errors of performance and changes in the environment? The key issue here is whether the bee distinguishes between *two types* of prediction error: prediction errors that are attributable to a mismatch between *its own performance* and its predicted performance in a case where the environmental parameters were as predicted (e.g. a failure to correct for wind) and prediction errors that are attributable to an unpredicted *change in the environment* (e.g. a change in the wind). If it does, the next key issue is whether the bee can trace the error in performance to a specific aspect of technique (e.g. a specific miscalculation). If it can, the third key issue is whether the bee feels a distinctive kind of affective pressure (directed discontent) that motivates a change to technique in those cases where it traces the source of the prediction error to an aspect of technique. I don't know of any evidence that bees register a distinction between two types of error, but I do not rule it out either. All I can say is that I think some such distinction must be registered if the control model is to encode a norm of correct performance.

The same point can be made in relation to Andrews & Westra's fascinating example of the "stop signal". Do bees signal differently in cases where the dancer's error is one of *performance* (the dancing bee is failing to relay the properties of the nest correctly) and cases where the error is due to a *change in the environment* (the dancing bee is relaying out-of-date, superseded information correctly?). This is an empirical question, and it points to another context in which we could test whether bees can make this basic, all-important distinction between two types of error.

We humans do draw this distinction. We draw it even in cases of simple, basic motor actions: if a cup slips from my grasp, or if I slip when walking (an example discussed by **Sterelny**), I will experience the error as an error of performance, a change in the environment, or sometimes both. Since errors of performance by the agent's own lights (and those of others) are possible for even basic actions, I do think that even basic actions are norm-guided in humans. There are ways of grasping a cup, and ways of walking, that can attract normative approval or disapproval. Yet, to explain the evolution of this capacity, I think we need to look to more complex skills, such as craft skills, where internalizing a motivation to adhere to a specific style of execution would plausibly have yielded fitness benefits in ancestral environments.

Pulling inwards: is joint intentionality necessary?

Andrews & Westra think I underestimate other great apes. I suggest that ant-dipping probably doesn't require model-based cognitive control; they reply that I underestimate the complexity of this skill, which requires years of practice and involves fine-grained modification of technique to many aspects of the nest being probed. **Tomasello** makes a similar point (with a different critical aim) regarding the complexity of nutcracking, which requires careful calibration of performance in response to auditory feedback.

As I emphasize in my (2021a) article, I am happy to see the evolutionary origin date for model-based cognitive control pushed backwards by new evidence. I may have overestimated its recency. Acheulean toolmaking is a striking example of a skill that plausibly requires this form of control, but we can easily underestimate the cognitive demands of skills that lead to less spectacular products. If new evidence also points towards social norms in chimpanzees (as Andrews & Westra believe it does), then this is no problem for a hypothesis that posits a close evolutionary link between social norms and model-based cognitive control.

In Tomasello's view, however, the evidence in fact points to a dissociation, whereby model-based cognitive control was already there in the last common ancestor of *Homo* and *Pan* and yet normative cognition is wholly absent in modern chimpanzees. If this is right, then we still face the challenge of explaining why normative cognition evolved only in the *Homo* lineage, but now with the additional problem that we can't appeal to inadequate model-based cognitive control in chimpanzees as part of the explanation.

Where else might the roadblock encountered by *Pan* have been located, if not in model-based cognitive control? I think it may also be possible to locate it in the expansion from self-regulation to other-regulation: chimpanzees may regulate their *own* skill execution as we do, yet lack either the ability or the motivation to monitor and correct the performance of *others*. Tomasello agrees in general terms, but objects to the idea that the extra social motivation we possess that chimpanzees lack is merely a strong enough form of kin-based or reputation-based motivation. He thinks a big, distinctive, cognitive difference-maker is also needed: joint intentionality.

I agree, of course, that there are joint skills, in which we exercise our joint know-how (Birch 2019). I am also sympathetic to the idea that joint skill is an important part of the story of hominin evolution. Joint skill is a widespread feature of human social life, and I have always been very struck by Tomasello's claim that chimpanzees, by contrast, never really act jointly: as he has put it in the past, "you will never see two chimpanzees carrying a log together" (Stix 2014). That said, I think there is a risk of overestimating the cognitive requirements of joint skill. Joint skill requires close other-monitoring and other-prediction: each agent must know how to monitor, predict and make adjustments in response to the actions of the other agent. It also requires acting in easily predictable and monitorable ways, with the intention of facilitating coordination: acting in what I call (in Birch 2019) "actively coordination-enabling" ways. I am not convinced it requires more than this.

Tomasello posits a weighty extra ingredient: the agents must "imagine together in common ground a model to guide their actions. Then, in addition, the collaboration required each individual to play her individual role in the ideal way that they knew together in common ground was necessary for joint success." I see this as requiring too much of the agents in relation to their understanding of the other agent's role. Think of a jazz band: each player needs to be able to listen, predict and respond to the other band members, while helping them to do likewise, but a saxophonist need not understand what is involved in excellent drumming, nor vice versa. The other agent's performance can be a black box. They just need to do their part in a way I can coordinate with. Neither the standard for excellent drumming nor the standard for excellent saxophony is a matter of common knowledge.

My more minimalist take on the cognitive demands of joint skill makes me sceptical of the idea that chimpanzees fail to meet the cognitive system requirements. If joint skill is nonetheless absent in chimpanzees, then I am drawn again towards an explanation based on a general lack of prosocial motivation and trust. However, there is also another possibility (highlighted in Birch 2019): it could be that chimpanzees, while capable of other-monitoring and other-regulation *and* motivated to do it, lack the mindreading abilities to make their own actions sufficiently transparent and predictable to others. The plausibility of this depends on the mindreading abilities of chimpanzees, which is another disputed area (e.g. Heyes 2017; Krupenye et al. 2019).[1]

Two variants of the skill hypothesis

Sterelny and **Railton** are sympathetic to a close psychological and evolutionary connection between skilled action and normative cognition (as I would have expected, given their prior work) but are sceptical of some aspects of my account of the nature of that connection. Both offer their own alternative variants of their skill hypothesis. This is welcome. I am doubtful that the evidence currently available can actually tell between these variants, but I think future evidence may do so.

[1] I further discuss Tomasello's work on joint intentionality in my review of *A Natural History of Human Morality* (Birch 2017).

Sterelny, like Tomasello, attaches great importance to collaborative hunting, but for a different reason. For Sterelny, the importance lies not in joint intentionality but in the affective salience of performance error. Hunts have the potential to go *disastrously* wrong if one person makes a technical error, so that it is not difficult to imagine intense emotions like anger, outrage, disgust, guilt and shame accompanying these errors.

I agree about this. I think the high stakes involved in collaborative hunting may well have driven what I have called the "repurposing of shame" (and other social emotions): their co-opting from earlier functions relating to the management of dominance hierarchies to the new function of signalling failures (in oneself or in others) to live up to shared standards of behaviour (Fessler 2004). In toolmaking, even when collaborative, the cost to me of your error is not so high, and weaker forms of affective approval and disapproval (such as directed discontent) seem more apt.

An unusual feature of my account is that I see the repurposing of these intense social emotions for norm-related functions, however achieved, as a relatively minor part of the overall story. What particularly interests me is not so much the *intensity* of the affective response to a performance error as its *directedness* towards a specific aspect of technique. In this I see the seeds of the ubiquitous micro-regulation of action by norms that is so distinctive of modern human behaviour. It is this micro-regulation (e.g. the fact that even walking the wrong way, or grasping a cup the wrong way, can set you up for social sanction), and not our ability to feel intense emotions, that is part of what "makes us odd", to use Cecilia Heyes' (2012) phrase.

I hypothesize that this directedness of affective responses towards aspects of technique began with relatively low-intensity emotions, and that it began in the context of regulating skilled action. Any learned, complex skill where the fine details of technique matter a great deal to success could be a skill of the relevant type. As I conceded in the last section, even ant-dipping and nutcracking may conceivably have the right features. So, although Acheulean toolmaking strikes me as a good case, complex hunting techniques may also fit the bill.

The second mutation in Sterelny's variant is that my emphasis on on-the-fly, in-the-moment modification of technique is replaced by planning out (and reviewing) sequences of actions over longer timescales. It is over these timescales, he suggests, that adjustments to technique in response to affective pressure are most likely to occur. I see this as mapping on to a debate in the philosophy of skill between Wayne Christensen and colleagues (2016) and David Papineau (2013). For Papineau, cognitive control matters to skill execution, but only in so far as an agent must choose among broad, coarse-grained strategies or styles of execution (e.g. defensive or aggressive batting), which are then executed automatically. Christensen et al. reply: no, cognitive control also matters in the moment, such as when a biker must adjust their technique in a very fine-grained way to avoid a specific obstacle. I think we need more evidence (e.g. evidence from structured interviews and questionnaires, or from neuroimaging) to resolve this disagreement about cognitive control. Whatever the future

evidence supports, our accounts of normative cognition should integrate with it. If it is true (as Christensen et al. propose) that cognitive control models do guide fine-grained adjustments on-the-fly, then it is plausible that subtle affective responses are involved in motivating those adjustments.

Railton's variant gives an important role to *evaluative* (but initially *non-normative*) guidance of skilled action—and takes out the idea that cognitive control models themselves encode norms. This proposal is motivated by an important criticism. I hypothesize that norms of correct performance are encoded implicitly in cognitive control models in the pattern of mismatches between predicted and experienced sensory feedback that generate affective pressure to adjust a specific aspect of technique. This leads to a specific sense in which the models are inflexible: they encode a *right way* of executing the skill that will always be the right way, regardless of the other goals an agent may have. Yet Railton points out that, when an agent possesses a skill, they are typically able to execute that skill in a variety of styles, depending on their higher-level goals. For example, a skilled biker may set herself the goal of riding wildly, taking unnecessary risks. Or she may set herself the goal of masking her true abilities, riding clumsily so as not to embarrass a less skilled companion (cf. similar examples discussed in Riley 2017).

A prediction of my view is that these deliberately poor performances will still feel wrong to a skilled agent. The prediction is that, as the biker rides intentionally slowly, making intentionally clumsy adjustments, these departures from their internalized norm of correct biking will still trigger affective pressure to change technique, regardless of the rider's higher-level performance goals. If the rider has a high-level goal of riding clumsily, they will just have to overcome that internal pressure to ride normally. The pressure may be *outweighed* by other motives, but it is present nonetheless. Otherwise, the norm cannot be encoded *in* the control model.

This is a prediction about the phenomenology of skill that could be tested using structured interviews and questionnaires. I take it Railton finds this prediction implausible. Suppose, then, that it does indeed prove false: the norms are *not* encoded in the control model. Rather, the control model supports the flexible pursuit of the agent's higher-level goals or ideals (such as executing a skill well, wildly or clumsily), with the norm of correct performance being represented in the goal or ideal, not in the control model itself. What the model does is monitor distance from the goal or ideal, creating positive affect when the agent moves closer and negative affect when it moves further away. As Railton points out, this would not undermine the idea of a deep psychological and evolutionary connection between norms and skills. It just offers a different perspective on the nature of that connection.

Indeed, I think it would alter the connection in a way that reconciles the differences between my hypothesis and Tomasello's. Tomasello's concept of a role ideal would turn out to be of central importance after all, because norms as such would enter the story at the point where agents become able to set a role ideal as the objective of their evaluative control processes. In this story, the pursuit of Tomasello-style role ideals is made possible by a pre-existing

platform for the evaluative (but initially non-normative) control of skilled action—a platform that allows the agent to set an intended style of skill execution (not just a material objective, like finding food) and feel affective pressure to make progress towards it. This platform will be one precondition for normative guidance, but not the only precondition, since the ability to represent role ideals and the motivation to care about them is also needed.

This is an attractive overall picture, and I think it could well be right. It is, however, a variant of the skill hypothesis that takes away one of its bolder and more novel aspects: the idea that norms are literally encoded *in* control models, in the agent's robust dispositions to feel affective pressure to correct departures from a particular way of executing a skill. It also runs into a concern expressed earlier: the concern about the risk of overestimating the cognitive system requirements of early hominin joint action and the importance of role ideals. So I would like my original variant to stay on the table as a live option, alongside Railton's new variant. At the same time, I acknowledge that our current evidence points to an interesting relationship between skilled action and normative guidance without yet allowing us to triangulate the precise nature of that connection. For this, we will need more evidence.

Acknowledgements
Thanks to all the commentators and to Anton Leist for organizing this symposium. This work was supported by a Philip Leverhulme Prize from the Leverhulme Trust.

References

Birch, Jonathan. (2017). Review of Tomasello: A Natural History of Human Morality. *BJPS Review of Books*. Retrieved from: http://www.thebsps.org/reviewofbooks/michael-tomasello-a-natural-history-of-human-morality/

Birch, Jonathan. (2019). Joint know-how. *Philosophical Studies*, *176*, 3329-3352. https://doi.org/10.1007/s11098-018-1176-6

Birch, Jonathan. (2021a). Toolmaking and the evolution of normative cognition. *Biology and Philosophy*, *36*, 4 https://doi.org/10.1007/s10539-020-09777-9

Birch, Jonathan (2021b). The skilful origins of human normative cognition. *Analyse & Kritik*.

Christensen Wayne, John Sutton and Doris J. F. McIlwain. (2016). Cognition in skilled action: meshed control and the varieties of skill experience. *Mind and Language*, *31*, 37–66.

Fessler, Daniel M. T. (2004). Shame in two cultures: implications for evolutionary approaches. *Journal of Cognition and Culture*, *4*, 207–262.

Heyes, Cecilia M. (2012). Grist and mills: on the cultural origins of cultural learning. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*, 2181-2191.

Heyes, Cecilia M. (2017). Apes submentalize. *Trends in Cognitive Sciences*, *21*, 1-2. https://doi.org/10.1016/j.tics.2016.11.006

Krupenye, Christopher and Josep Call. (2019). Theory of mind in animals: current and future directions. *WIREs Cognitive Science*, *10*, e1503. https://doi.org/10.1002/wcs.1503

Levinson, Stephen C. (1995). Interactional biases in human thinking. In E. N. Goody (Ed.), *Social intelligence and interaction* (pp. 221–260). Cambridge: Cambridge University Press.

Papineau, David. (2013). In the zone. *Royal Institute of Philosophy Supplement*, *73*, 175–196

Railton, Peter. (2021). Normative guidance, evaluative guidance, and skill. *Analyse & Kritik*.

Riley, Evan. (2017). What skill is not. *Analysis*, *77*, 344-354. https://doi.org/10.1093/analys/anx059

Sterelny, Kim. (2021). The skill hypothesis: a variant. *Analyse & Kritik*.

Stix, Gary (2014). The 'it' factor. *Scientific American*, *311*, 72–79.

Stone, Thomas, Barbara Webb, Andrea Adden, Nicolai Ben Weddig, Anna Honkanen, Rachel Templin, William Wcislo, Luca Scimeca, Eric Warrant, Stanley Heinze. (2017). An anatomically constrained model for path integration in the bee brain. *Current Biology*, *27*, P3069-P3085. https://doi.org/10.1016/j.cub.2017.08.052

Tomasello, Michael. (2021). Norms require not just technical skill and learning, but real cooperation. *Analyse & Kritik*.

Andrews, Kristin and Evan Westra. (2021). If skill is normative, then norms are everywhere. *Analyse & Kritik*.