Noname manuscript No.
(will be inserted by the editor)

Modelling the truth of scientific beliefs with cultural evolutionary theory

Krist Vaesen & Wybo Houkes

accepted for publication in *Synthese*. Final version available at: http://link.springer.com/journal/11229

Abstract Evolutionary anthropologists and archaeologists have been considerably successful in modelling the cumulative evolution of culture, of technological skills and knowledge in particular. Recently, one of these models has been introduced in the philosophy of science by De Cruz and De Smedt (2012), in an attempt to demonstrate that scientists may *collectively* come to hold more truth-approximating beliefs, despite the cognitive biases which they *individually* are known to be subject to. Here we identify a major shortcoming in that attempt: De Cruz & De Smedt's mathematical model makes one particularly strong tractability assumption that causes the model to largely miss its target (namely, truth accumulation in science), and that moreover conflicts with empirical observations. The second, more constructive part of the paper presents an alternative, agent-based model, which allows one to much better examine the conditions for scientific progress and decline.

Corresponding author:

K. Vaesen

Philosophy & Ethics

Eindhoven University of Technology

P.O. Box 513

5600 MB Eindhoven

The Netherlands

E-mail: k.vaesen@tue.nl

http://home.ieis.tue.nl/kvaesen/

1 Introduction

Philosophers of science are slowly starting to model the collective aspects of scientific knowledge and research practices. This could take the form of modelling disagreement between experts and aggregation of their judgements (List, 2005; Hartmann and Sprenger, forthcoming) or of the division of scientific labour (Kitcher, 1990; Strevens, 2003; Weisberg and Muldoon, 2009). Alternatively, one might focus on epistemic processes that span generations of collectives of scientists—the population dynamics of beliefs. Here, one may draw on evolutionary anthropology, where compelling evolution-theoretic models of technical skills and knowledge have been proposed (Boyd and Richerson, 1985; Richerson and Boyd, 2005; Henrich, 2004). In this paper, we identify a major shortcoming in a recent attempt at applying one of these models to scientific knowledge; then, in a more constructive mode, we propose a more promising (but still imperfect) alternative for modelling the population-dynamics of scientific belief.

The application in question concerns a recent paper by (De Cruz and De Smedt, 2012, henceforth D & D), who make a strong case for incorporating the socio-cultural dynamics of belief transmission into naturalized epistemology—introducing a collective, generation-spanning dimension in the formation of scientific beliefs. Their aim is to show that despite the truth-distorting cognitive biases that individual scientists are subject to, scientific beliefs may over time become more truth-approximating in virtue of the interaction of scientists.

The interaction D & D have in mind is cultural transmission, i.e. the processes of social learning that allow individuals to pass on acquired knowledge to other individuals remote in space and time. The basic idea is that scientists pass on previous achievements to subsequent generations, which can improve these achievements in turn. Such cumulative effects, D & D believe, may be usefully examined with models devised in evolutionary anthropology to explain the gradual accumulation of culture, of technological innovations in particular. More specifically, D & D borrow a popular model by

Joseph Henrich (2004) to argue that cultural transmission among scientists can yield a progressive evolution towards more truth-approximating representations on two conditions: first, that the relevant community of scientists is sufficiently large and, second, that from copied ideas scientists infer a sufficiently diverse set of new ideas. If these conditions are met, the distorting effects of cognitive biases may be filtered out over time—or so the argument goes.

D & D adopt from the original anthropological model a central assumption made for reasons of analytical tractability, namely that scientists are biased towards and try to copy the single best scientific theory available. This implicates D & D's account in two respects. First, as argued in Section 3, the tractability assumption reduces the explanatory scope of D & D's model to the extent that the model largely misses its supposed target, namely truth-accumulation in science. The second problem, discussed in Section 4, is the empirical inadequacy of the assumption—there is good reason to think that scientists are *not* biased towards the single best belief in the pool of scientific beliefs.

An alternative model, we conclude, should allow one to relax the tractability assumption. In Section 5, we describe such a model. In particular, we present an agent-based implementation of D & D's (and of Henrich's) model that can be used to examine the effects of varying scientists' copying biases. These effects, described in Section 6, will turn out large: in the limiting case, where scientists exhibit strong conformity biases, gradual truth-approximation does not get off the ground at all. Because it is largely an open empirical question which of the modelled biases characterize the biases of actual scientists best, the exercise does not allow us to settle the question whether processes of cultural transmission *actually* do give rise to progression towards truth. Still, the model *does* allow us to derive a set of in principle testable hypotheses about the circumstances under which such progress can be expected to occur and not to occur.

2 The cultural transmission of scientific knowledge: D & D's model

Before discussing its problems,[1] let us review what D & D's model is constructed to do, and how it is supposed to do this. As said above, the model is aimed at showing how, and under which conditions, the cultural transmission of scientific beliefs results in these beliefs approximating the truth, despite the cognitive biases of individual scientists. D & D do not give a general characterization of such biases, but provide a set of examples instead, such as the human disposition to conceptualize and reason about the world in terms of a limited number of intuitive ontologies. Humans, for instance, are inclined to reason about themselves and conspecifics as if they are not animals (Waxman, 2005). This intuitive ontology, D & D argue with Foley (2001), has hampered archaeologists and paleoanthropologists in their theorizing about the evolution of hominids. Until the mid 1970s, scientists believed that human evolution was linear and that only one hominid species existed at any one time, ideas that contrasted sharply with the bush-like evolutionary models posited for all other organisms, but that were acceptable given the special-status bias regarding human beings. Likewise, the intuitive idea that species have essences has resulted in 2000 years of stasis in taxonomic theory—or so argues Hull (1964). Finally, psychological evidence suggests that humans have a natural inclination for teleological reasoning (Kelemen, 2004); that inclination too can be expected to work against a correct causal understanding of the evolution of species.

Let us assume that human beings are indeed hampered by a set of specific biases in their search for scientific knowledge. How is it then, D & D ask, that science can still produce true beliefs? To answer that question, they consider a supposedly (more on this below) pessimistic scenario in which cognitive biases are very influential in science; and next, they model how these biases are, over time, overcome through processes of cultural transmission.

[1] Throughout, we assume that Henrich (2004) provides the best existing model of cultural transmission of technological knowledge. Shortcomings shared by the original model and its application by D & D are, as much as possible, left aside, in order to focus on problems specific to modelling the transmission of scientific knowledge.

The model D & D deploy was originally developed by evolutionary anthropologist Joseph Henrich (2004), who wanted to examine conditions for the acquisition and loss of complex skills and technologies. Henrich found that population size is one factor that determines whether complex skills may be retained over time.

In D & D's application of the model, any scientific belief i is associated with a value z_i, which denotes how well the belief captures observer-independent reality; the higher z_i, the more truth-approximating i. For instance, z's may capture the representational accuracy of beliefs concerning the structure of the atom. Dalton's early nineteenth-century idea of atoms as hard billiard balls would have a lower z than Thomson's plum pudding model; and the latter would have a lower z than Rutherford's early twentieth-century miniature solar system model. $\Delta\overline{z}$ denotes the average change in representational accuracy of beliefs between two generations. In case of models of the atom, $\Delta\overline{z}$ was positive for the early nineteenth-to-twentieth century interval. In case $\Delta\overline{z} = 0$, truth-approximation would have stabilized; in case $\Delta\overline{z} < 0$, accuracy would have decreased.

For modelling $\Delta\overline{z}$, the Price equation is used (Price, 1972):

$$\Delta\overline{z} = \underbrace{Cov(f, z)}_{\text{selective transmission}} + \underbrace{E(f\Delta z)}_{\text{noisy interference}} \tag{1}$$

As one can see, the equation introduces a new variable f. Any scientific belief i has a value f_i which represents the frequency with which the belief will be copied and passed on to the next generation. In case the z-value of a belief i is high, and assuming that individuals are more likely to copy beliefs with a high z-value, i will be copied more frequently, and f_i will be high.

The Price equation separates $\Delta\overline{z}$ into two factors. The first is $Cov(f, z)$, which represents how cultural success (or frequency of copying) and representational accuracy co-vary. Suppose we have two beliefs i and j, with $z_i > z_j$. Suppose moreover that representational accuracy and cultural success are highly correlated—better beliefs are more likely to be copied. If so, we should expect belief i to be copied more abundantly than belief j (or $f_i > f_j$), so that the next generation of beliefs will contain more copies of i than of j. Given that for this generation too $z_i > z_j$, it will have a higher average $z$value than the previous generation. $Cov(f, z)$ just tells us how much z increases because of this differential reproduction (i.e. different z-values leading to different numbers of descendants in the next generation).

For understanding the second factor in the Price equation, first let z_i' denote the z-value of the *copy* of belief i. If transmission between generations is perfect, $z_i' = z_i$. Often, however, copies will be different from the original, and $z_i \neq z_i'$. Let Δz_i be the difference between z_i' and z_i. Thus, Δz_i captures transmission fidelity; it is a measure of how faithfully copying proceeds. If $\Delta z_i = 0$, copies of i are identical to the original belief i; if $\Delta z_i \neq 0$, there is a transmission error or more systematic bias.

Evidently, whenever present, such transmission biases should affect $\Delta \overline{z}$. Suppose, for example, that copying is highly unfaithful, and that copying errors always lead to inferior copies. Under these circumstances one would intuitively expect average representational accuracy to decrease between generations. Conversely, in case copying always lead to superior copies (i.e., there is a positive bias or constructive source of error), $\Delta \overline{z}$ should *ceteris paribus* increase.

$E(f\Delta z)$ captures the effect of transmission fidelity. It is the expected value of the product of copying frequency (i.e. f) and transmission fidelity (i.e. Δz). f is included because the effect of transmission fidelity needs to be weighted by the number of copies affected by it. For example, even if a belief i has a high Δz_i, its impact on $\Delta \overline{z}$ might remain low if it has few offspring (a low f-value).

In his application of the Price equation, Henrich represents imperfections in imitation with the transmission-fidelity factor. In imitating the skills or achievements of their mentors, students may be prone to errors and biases. In some cases, these imperfections are harmful, leading to inferior skills with lower z-values; in other cases, imperfections may be beneficial, leading to increased z-values. D & D follow the same line in representing cognitive biases, such as intuitive ontologies. Scientists with essentialist dispositions, for instance, may be likely to copy evolutionary theories imperfectly. In some cases, this imperfection is harmful, decreasing the z-values of their theories; in others, it may be benign. Below, we discuss in some detail how the resulting changes are calculated.

As it stands, the Price equation is not helpful for studying the general conditions for $\Delta\overline{z}$ to increase (or not decrease). In particular, the possibility that students choose different, or even multiple, mentors makes the model analytically intractable. Therefore, D & D follow Henrich in assuming that all students copy the best (in their case: the most accurate) model (in their case: belief) in the population, namely h (having a z-value z_h). So $f_h = 1$, whereas $f_{not\,h} = 0$. This tractability assumption, the main target of our criticism in Sections 3 and 4, considerably simplifies the Price equation:

$$\Delta\overline{z} = \underbrace{z_h - \overline{z}}_{\text{selective transmission}} + \underbrace{\Delta z_h}_{\text{noisy interference}} \tag{2}$$

The first part of Equation (2) takes the difference between the z-value of the subsequent generation (which equals z_h, because all individuals are assumed to copy h) and the z-value of the earlier generation (which equals the average of z's of that generation, i.e. $\overline{z}$). The second part of Equation (2), namely Δz_h, captures the transmission bias associated with copying h.

D & D also assume (again following Henrich) that z-values of copied beliefs are drawn randomly from a Gumbel distribution. The justification for this is that a wide range of distributions, including the Normal and Gumbel distribution, yield a Gumbel distribution if the highest value is repeatedly taken from samples of size N (Henrich, 2004, p. 210).[2]

A relevant feature of Henrich's model is that students only rarely outperform their mentor h: their skill level typically decreases by an amount α (greater than or equal to 0). This represents a structural imperfection in imitation, which we shall call "inaccuracy". β (also greater than or equal to 0) represents the dispersion of the Gumbel probability

[2] For a discussion of the problems associated with assuming Gumbel rather than Normal distributions, see Vaesen (2012).

distribution, and indicates the variety among the individual students of h. If h is hard to copy, for instance, but students all make the same mistake, α would be high, and β low. In case both α and β are zero, replication is perfect. In general, the probability that students outperform their mentors is given by the area under the curve to the right of the dashed line in Figure 1, which represents z_h. By choosing α and β values, the model represents how inaccurate and variable imitation is for a particular technology or domain of knowledge.

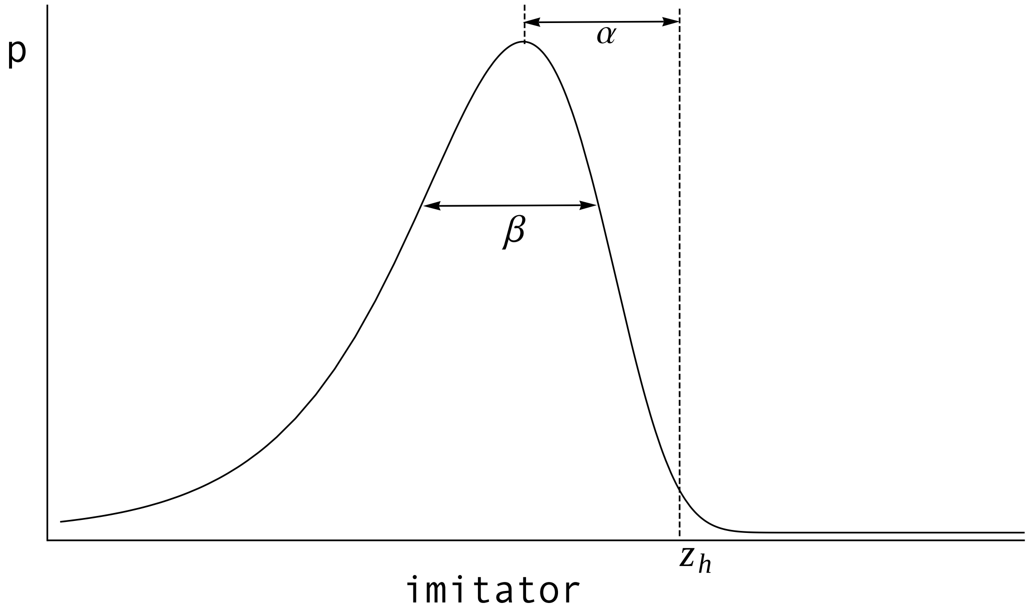

Fig. 1 Gumbel probability density function for imperfect imitation

With these assumptions, Henrich (2004) derives the following equation (for the technical details, see Appendix A):

$$\Delta\bar{z} = \underbrace{-\alpha}_{\text{degenerative}} + \underbrace{\beta(\gamma + \ln(N))}_{\text{positive}}, \tag{3}$$

where N is the population size;[3] and γ the Euler-Mascheroni constant (≈ 0.577). Clearly, the inaccuracy term α has a general degenerative effect on average skill level, which may be counterbalanced by the second term. This will happen when β and/or N are sufficiently large; in other words, when there is substantial variety among students who imitate h and/or there is a sizeable population of students. The idea behind this is not difficult to grasp: although inaccuracies in imitation have a structurally degenerative effect, the sheer number of imitations may, in combination with the possibility of beneficial errors or individual inventions, compensate for or even outweigh the constant downward pull of α.

Figure 2 presents these ideas graphically. The Y-axis gives different values of imitation inaccuracy α. All the three curves correspond to parameter combinations for which $\mathit{\Delta}\bar{z} = 0$; parameter combinations below a curve are associated with regimes of increasing representational accuracy (i.e. $\mathit{\Delta}\bar{z} > 0$), while those above a curve yield loss of accuracy (i.e. $\mathit{\Delta}\bar{z} < 0$). Each of the three curves, however, corresponds to a different β value; the higher the diversity β, it appears, the higher the imitation inaccuracy that can be compensated for. The same holds for larger population sizes N.

In D & D's application, α represents how transmission of scientific beliefs is structurally affected by cognitive biases. However, scientific progress can still occur on the condition that values of β and N are sufficiently high; in other words, when there is a sufficiently large number of replicas and/or sufficient variety between these replicas. There will be scientific progress when at least one result of imitation is superior to the currently best theory h—i.e. the z-value of that copied belief needs to be to the right of the dashed line in Figure 1. Then, that copy can go on to become the model for the next generation. When will that be more likely to happen? First, in case population sizes

[3] According to D & D, N refers to the size of the scientific community. But since $\mathit{\Delta}\bar{z}$ represents the representational accuracy of scientific *beliefs*, it is more accurate to say that N represents the population of available scientific beliefs (rather than of scientists). Nothing much depends on this as long as it is reasonable to assume that there is a 1:1 correlation between the two populations. In light of this, we will refer to these populations interchangeably.

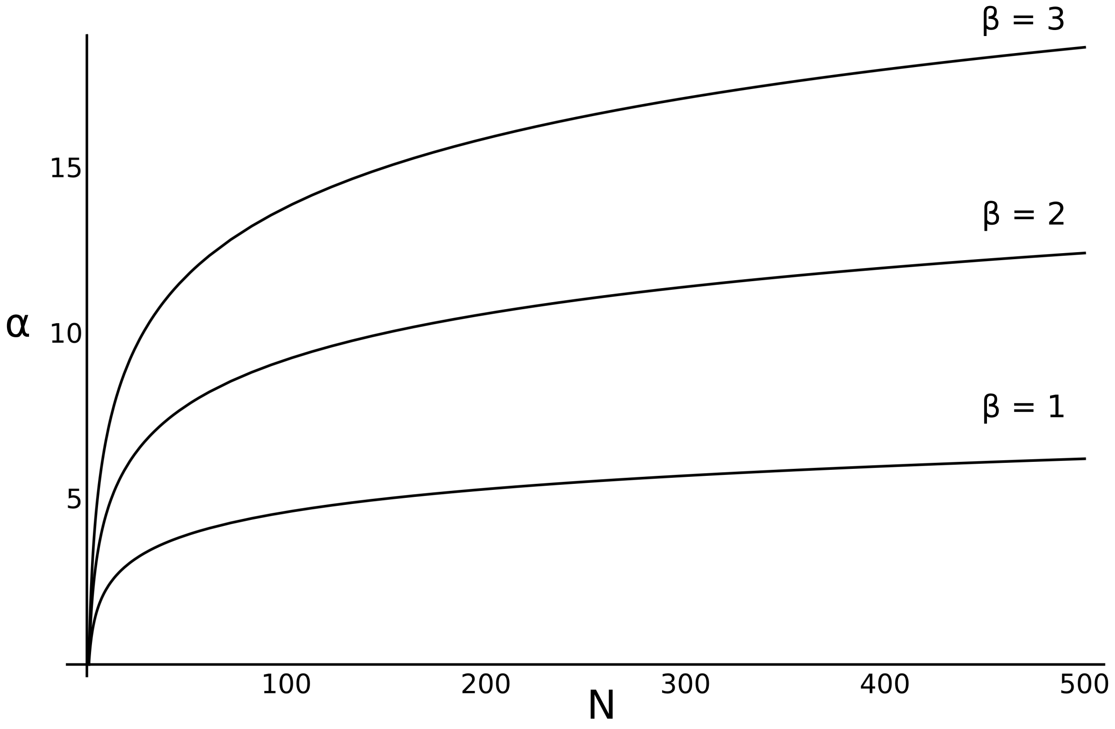

Fig. 2 Population size (N), inference diversity (β), and copying inaccuracy combinations for which $\Delta\bar{z} = 0$. Higher N's and β's can offset higher levels of copying inaccuracy.

are large; exceptionally high z-values will occur more readily when drawn from larger samples. Second, exceptionally high z-values will surface just in case there is variation in the kinds of copies made from h; in the extreme case where β is 0 (and $\alpha \neq 0$), no copy scores better than $z_h - \alpha$, and *a fortiori*, no copy scores better than z_h.

In D & D's setup, it seems thus, processes of cultural transmission can offset the cognitive, truth-distorting biases which scientists are subject to. Even in the pessimistic scenario where initial theories are far from the truth and cognitive biases result in strong inaccuracies (i.e. low z, high α, respectively), science may progressively approximate truth as long as population size (N) and replica diversity (β) are sufficiently large. Below, however, we identify two flaws in D & D's argument.

3 Problem 1: Where is the problem?

According to D & D, the evolution of scientific belief is driven by the following two processes. First, scientists select a belief that they find imitation-worthy, and second, scientists make a copy of that belief. Let us call these two processes belief selection and belief imitation, respectively.

In the former, scientists do not select just any belief, but, per D & D's tractability assumption, the best (i.e. truest) belief, h, currently available in the scientific community. Consequently, *belief selection always favors increases of average truth value.* Progress in science is thus only worked against by the second process, that is, by belief imitation: due to cognitive biases, imitation may yield inaccurate copies of h, which typically are inferior to (but occasionally better than) h itself.

Now, the fact that D & D's model allows only belief imitation to interfere with the scientific aims of truth is not just empirically questionable (see Section 4), it seriously limits the scope of D & D's argument. The model can only show how processes of cultural transmission may offset structural *copying errors* occurring in science. By its tractability assumption, it cannot say anything about the conditions (if any) under which errors of belief selection may be cancelled out. This is remarkably unfortunate given the fact that belief selection, on D & D's own reading, comprises practices which make up the bulk of the scientific enterprise, such as experimentation, empirical validation, and so forth. For regarding belief selection, D & D *de facto* assume that scientists can and do properly 'assess competing scientific theories [e.g.] through epistemic values, intuition, experiment, evaluation of empirical adequacy or a combination of these factors (p. 12)'.

That the bulk of science is simply assumed to be free from the distortive effects of cognitive biases makes D & D's scenario much less pessimistic than advertised. The scenario is just as pessimistic as is credible the (to our ears highly incredible) claim that in science imitation is the only or chief source of error.

To fully appreciate how D & D's optimism reduces the explanatory scope of their model, let us contrast it with the original, namely that of Henrich. Although Henrich also assumes individuals to be able to select the best model in the population, this really is a pessimistic assumption for the phenomenon he wants to explain, namely the *loss* of complex technologies. In assuming that copiers are able to identify correctly the best mentor, Henrich might indeed overestimate a population's capacity for retaining skills through cultural transmission; but if loss occurs even under these favourable conditions, it will certainly occur in case model selection is less than perfect. In contrast, when the explanandum isn't so much cultural loss as cultural gain (as is the case for D & D), the same assumption assumes partly what is to be explained.[4] To bring out the contrast explicitly: the result of Henrich's model could be sloganized as "*Even if we would always learn from the best, complex skills may get lost in transmission*"—which is an informative claim. The counterpart for D & D's model would be: "*If we could always learn from the best, we would be more likely to come to believe the truth*"—which is a fairly trivial claim. Or more accurately, it is a trivial claim, *unless* D & D have convincing evidence showing that: (i) imitation is the most truth-distortive activity of science; and (ii) the forces that make imitation so error-prone do not affect the reliability of the other process at work in the evolution of belief, namely belief selection.

Before we question these two claims on empirical grounds in the next section, we would like to note that the first one seems implausible already in light of the model's own assumptions. In particular, the tractability assumption is really difficult to make fit with low-fidelity copying. It is hard to imagine how errors of imitation can be big if the prior process of belief selection was properly executed (as D & D assume). The ability to evaluate the empirical adequacy of someone else's theory T, say, through

[4] D & D are not alone in deploying Henrich's model to explain instances of cultural *gain*. Powell et al. (2009), for instance, present an agent- based version of Henrich's mathematical model to explain the emergence of modern behaviour in the Late Pleistocene. Everything we say here thus also applies to the results of Powell et al (in fact, this kind of criticism has been leveled at Powell et al by one of us before, see Vaesen (2012)).

a laborious process of experimentation, seems to require, minimally even, the ability to first reproduce a fairly accurate copy of that very theory T.[5] The assumption that belief selection proceeds perfectly thus limits how much error can be expected to occur subsequently during imitation. In light of this, if one accepts the tractability assumption, one has reasons to doubt the pertinence of the problem D & D have found a solution to.

4 Problem 2: What cognitive biases are we (not) talking about?

In the above, we have more or less bracketed the forces which may cause errors in science: cognitive biases. There are two problems in the way D & D conceive of them.

First, it is unclear how cognitive biases could leave belief selection unharmed (as, again, per D & D's tractability assumption). Cognitive biases stand in the way of seeing the truth; belief selection is a matter of identifying true belief; hence, belief selection is a natural target for cognitive biases. Consider human teleological biases again. These are portrayed as an impediment to a correct understanding of the evolution of species (p. 14). Now, if a teleological bias indeed makes it difficult to see the correctness of the Darwinian view, the bias should be expected to work also during belief selection, that is, while experimenting and empirically validating the view in attempt to verify whether it really is best; even more so perhaps than during belief imitation.

Second, a bit ironically, in the literature D & D borrow their model from discussions of cognitive biases virtually always concern selection biases, *not* biases affecting copying fidelity. Joseph Henrich (2003), in a paper co-authored with Richard McElreath, gives a useful overview of the kinds of biases affecting the likelihood of some individuals being copied more readily than others (see Figure 3).

At the highest level in Figure 3, one sees the first broad distinction, namely between so-called content and context biases. Content biases arise due to cues associated with

[5] Let us stress again that experimentation, on D & D's construal, happens before imitation. Scientists thus are assumed to engage in testing T, not, as seems more natural, in testing their inaccurate, internalized copy of it.

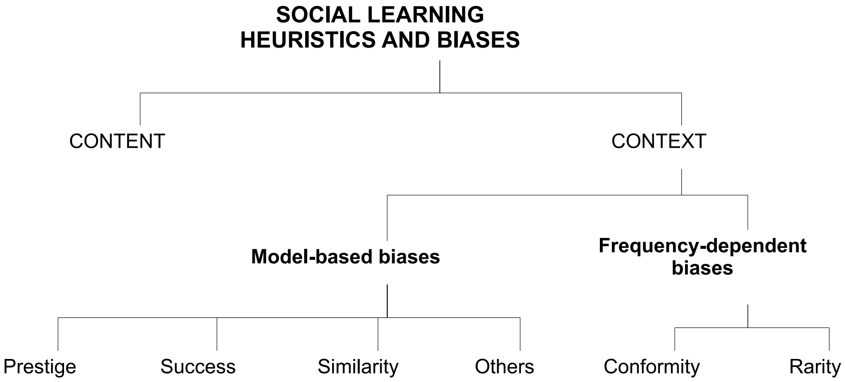

Fig. 3 Cognitive biases in social learning

the thing being transmitted. For example, the quality of a belief—defined in terms of its truth or utility or something else—affects the likelihood of its being copied. D & D's tractability assumption clearly falls under this heading; scientists are assumed to be strongly biased towards good beliefs, viz. beliefs that are true.

Importantly, however, content biases may also favour the transmission of less favourable beliefs. Think of the well-documented confirmation bias, which leads scientists to seek and copy information that confirms their false preconceptions (for a useful overview of research on confirmation biases, see Nickerson, 1998). Or think of the phenomenon of belief perseverance, which sustains the transmission of false beliefs even in the face of strong counter-evidence (for an early treatment of belief perseverance, see e.g. Ross et al., 1975).

D & D's assumption that scientists have the capacity to identify truth-approximating belief would thus imply that scientists are immune to the deleterious content biases just mentioned. The previous paragraph suggests that they are not.

Furthermore, D & D's ignore all biases on the right-hand side of Figure 3. These are the so-called *context* biases, related to the tendency of humans to copy others based on cues provided by the learning context. That is, one may copy other individuals not because their beliefs are particularly good, but because these individuals either have prestige; or are generally successful; or are similar to oneself; or are one's friends; or are one's competitors. If so, one would copy based on properties of the mentor rather than on any merits of the mentor's scientific beliefs. In a similar vein, belief selection may be informed by how widespread certain beliefs are in a community; one may copy the majority (conformity bias) or, conversely, beliefs that are rare (rarity bias). Although these contextual cues may sometimes be taken to be indicative of properties of content—e.g., the prestige of a scientist may sometimes really correlate with the truth of her scientific theory—such correlations cannot be taken for granted. For instance, the majority is arguably often wrong, so that conformity biases would promote the transmission of falsehoods rather than of truths.

Importantly, these context biases are not just hypothetical. There is ample empirical evidence showing that they are real and abundant (for a useful overview, see Mesoudi, 2009). So if D & D really think that scientists are liable to the same cognitive limitations as ordinary people (p. 6), they should incorporate the right-hand side of Figure 3. The operation and interaction of many of these biases have, of course, been brought out by sociological analyses that emphasize "mob psychology", "bandwagon effects" and peer pressure in scientific practice (Kuhn, 1962; Collins, 1985; Latour, 1987). This seems to create a dilemma for the modelling of learning processes in science: either one takes seriously all cognitive biases but ends up with an analytically intractable model (see Appendix B for why this is so), or one retains the strong tractability assumption but ends up with an empirically inadequate model.

5 An agent-based version of D & D's model

The dilemma between intractability and empirical inadequacy derived at the end of the previous section can be avoided by developing an agent-based version of D & D's mathematical model. Agent-based models have seen some use in the philosophy of science, to study effects of the structure of communication networks (Zollman, 2007; Grim, 2009), various types of social networks (Payette, 2011) and the distribution of different learning strategies (Weisberg and Muldoon, 2009) on the convergence of beliefs in communities of scientists. Our agent-based model studies the effects of a single learning strategy, imperfect intergenerational imitation. In particular, it is devised to examine the effects of relaxing the assumption of perfect mentor selection, shared by Henrich and De Cruz & De Smedt. Thus, our model represents a community of scientists who are imperfectly tuned towards the best beliefs currently in the pool of beliefs, and who imperfectly copy those beliefs. The model starts with N agents that act as the parent generation, and N agents that act as the offspring generation. In each time-step, the model goes through the following two steps:

1. *Transmission*: each offspring selects (*) one agent from the parent generation, the latter acting as a cultural parent for the former. The offspring individual takes the z-value of the parent, according to the transmission process described by D & D.
2. *Replacement*: the offspring generation replaces the parent generation, and the average z-level of the population, $\overline{z}$, is measured.

Different selection biases are implemented at (*) in step 1. In particular, largely following Vaesen (2012), [6] offspring select a parent according to one of the four following learning biases:

a. in case of *Extreme Success*, offspring select the single best parent in the population;

b. in case of *Modest Success*, offspring select a parent with probability proportional to the parent's z-value;

c. in case of *Conformity*, offspring select a parent with probability inversely proportional to the magnitude of the difference between the model's z-value and the mode, μ, of the distribution; and

d. in case of *Random Copying*, offspring select a parent at random.

Let us briefly explain these biases. *Extreme Success* is simply the kind of selectivity assumed by D & D.

Modest Success can be interpreted in two ways, either as a content bias or as reflecting a context bias. In the former case, scientists preferentially select good theories over bad theories in virtue of the goodness of the theories. They do so less perfectly than in case of *Extreme Success*, either due to inherent imperfections of the selection process (e.g., citation numbers may provide a good but imperfect indication of the goodness of a theory; scientific ideas may be unequally available; satisficing rather than optimizing strategies may be adopted for pragmatic reasons) or due to the fact that the selection process is affected by cognitive biases of the sort D & D are interested in (e.g., intuitive ontologies, essentialist preconceptions)—so that Problem 1 (Section 3) is addressed. In contrast, if construed as a context bias, *Modest Success* assumes scientists to be biased towards good theories because they are responsive to some (supposed) proxy for

[6] The main difference is that the model of Vaesen (2012) contains a third step, in which offspring also undergo vertical transmission (i.e. they learn from their biological parents). In our model, in contrast, cultural transmission is assumed to proceed only through oblique transmission (i.e. offspring learn from parents that are not necessarily their biological parents, as in step 2), simply because learning from one's biological parents arguably plays only a marginal (if any) in scientific practice.

the goodness of theories, such as the parent's prestige or the fact that her (empirical) success is more public than that of others. As such, *Modest Success* offers one way of addressing Problem 2 (Section 4).

A different context bias is *Conformity*, which refers to a disproportionate tendency to copy the most common behaviour in the parent population (in our case approximated by the mode, μ, of the distribution). How real is conformity in the scientific enterprise? There is good theoretical and empirical evidence for conformist transmission in general: models of Henrich and Boyd (1998) suggest that conformist biases are likely to evolve whenever social learning evolves; Henrich (2001) finds evidence of conformist transmission in field data on the diffusion of innovations; Henrich and Boyd (2002) and Mesoudi (2009) review a whole set of empirical studies demonstrating the powerful human propensity to conform. Consequently, following D & D's suggestion that scientists are liable to the same cognitive dispositions as ordinary people, and given the evidence for peer pressure and bandwagon effects provided by sociologists of science, there is good reason to treat textitConformity as, minimally, a real possibility.

The last condition, *Random Copying*, implies randomness in the strategies of offspring when selecting a parent. The offspring individual is assumed either to really select parents at random (i.e. she is insensitive to clues about such things as the parent's success or the popularity of the parent's beliefs), or to switch strategies seemingly at random (e.g, conformist under time pressure, success-based otherwise; success-based if cheap, conformist otherwise; conformist in matters of statistics, success-based in matters of biology). While there is accumulated evidence that contemporary humans do not just select cultural parents at random (for a recent report, see Mesoudi, 2011), and thus indeed *are* responsive to the features of the parents they choose, little is known about the precise conditions under which they are responsive to which clues, that is, about the conditions under which one of the various biases is dominant over the others. This holds also (especially?) for scientists: we simply do not know which heuristics scientists use when. In this light, it makes good sense to consider what happens in case no systematic trends in scientists' biases are assumed. The *Random Copying* condition does this; it acknowledges that we are unsure about what would qualify as a more realistic condition.

For given values of N, we simulated widely over α, to find the level of transmission inaccuracy that could be sustained by a population of size N without loss in average representational accuracy, $\Delta\overline{z}$. For further details, we refer the reader to Appendix C.

6 Results of the agent-based simulations

The results of the simulations are presented in Figure 4. Note first the profound difference between *Extreme Success* (which, recall, reproduces D & D's assumption of perfect belief selection) and the other three conditions. *Extreme success* is not only able to sustain much higher levels of transmission inaccuracy α, it is also the only condition for which there is a sustained population effect at higher N's.

These observations reinforce our previous point: D & D's model is extremely optimistic. It offers a credible explanation of scientific progress only on the unlikely condition that scientists always perfectly assess the epistemic standing of all theories and pick out the single best; and that they are immune to prestige, conformity and other biases. Once such considerations are taken into account, the conditions of scientific progress are much more limited. In case of *Conformity* or *Random Copying*, there is stasis throughout; some gain is achieved in *Modest Success*, namely for lower values of N.

The good news, however, is that D & D's targeted example—progress in biology from 1760 to 1860—is one that involves low values of N, namely a growth in the community of biologists from 60 in 1760 to 240 in 1860. So to have a case as regards this particular example, D & D would only need to establish the truth of *Modest Success*; that it is true that scientists, conformity and other biases notwithstanding, preferentially select good theories over bad ones. Although this may prove quite demanding, it is arguably less difficult than defending *Extreme Success.*

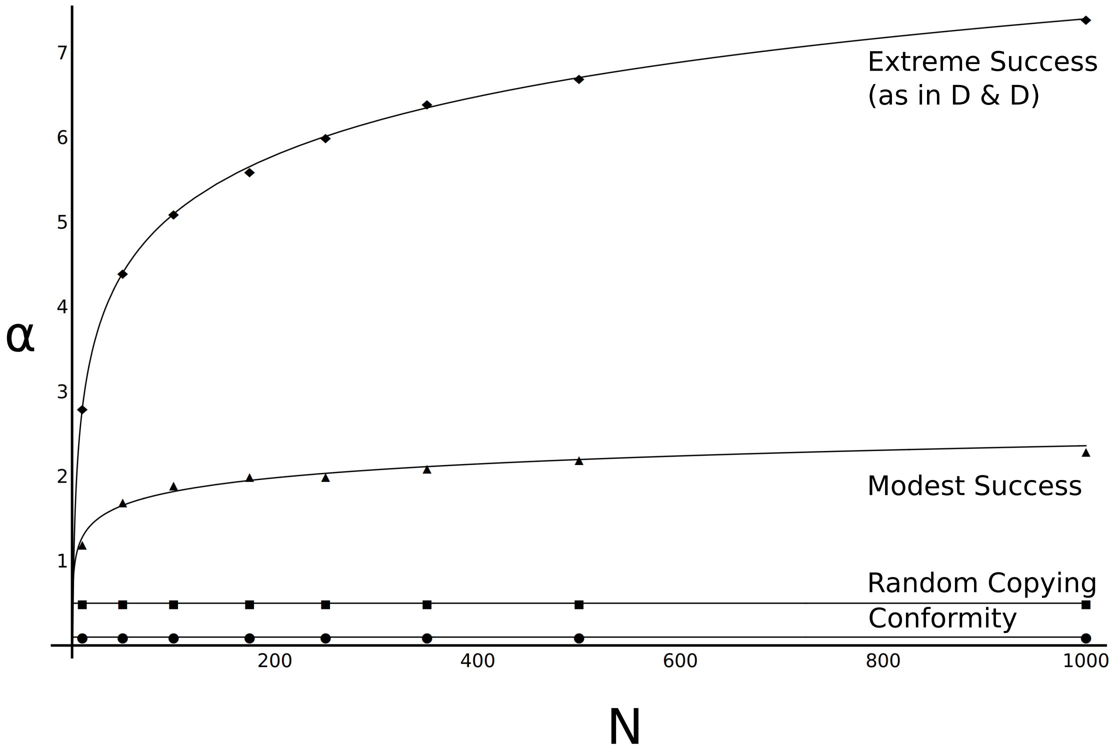

Fig. 4 Critical population size (N) versus transmission inaccuracy (α) for different selection biases. Parameter combinations above a curve correspond to a regime of loss (i.e. to decreases of average representational accuracy, $\Delta\overline{z} < 0$); those under a curve correspond to a regime of gain (i.e. to increases of average representational accuracy, $\Delta\overline{z} > 0$).

For examples including higher N's (examples from contemporary science perhaps) matters are different. Now one would need to show that scientists are biased towards good theories even more strongly than just with probability proportional to the parent's z-value (as in *Modest Success*). In this regard, one could try the idea that scientists engage in, what Mesoudi and Lycett (2009) call, *Frequency-dependent Trimming*, i.e. they focus only on reasonably good beliefs and ignore beliefs in the bottom B fraction of the frequency distribution.

The effects thereof are examined in Figure 5, which shows the results of implementing *Frequency-dependent Trimming* for values of B from 0.0 to 0.9. In case $B = 0.3$ (for instance) offspring discard the 30% percent worst parents; and choose a parent from the remaining 70%, with probability proportional to the parent's z-value.

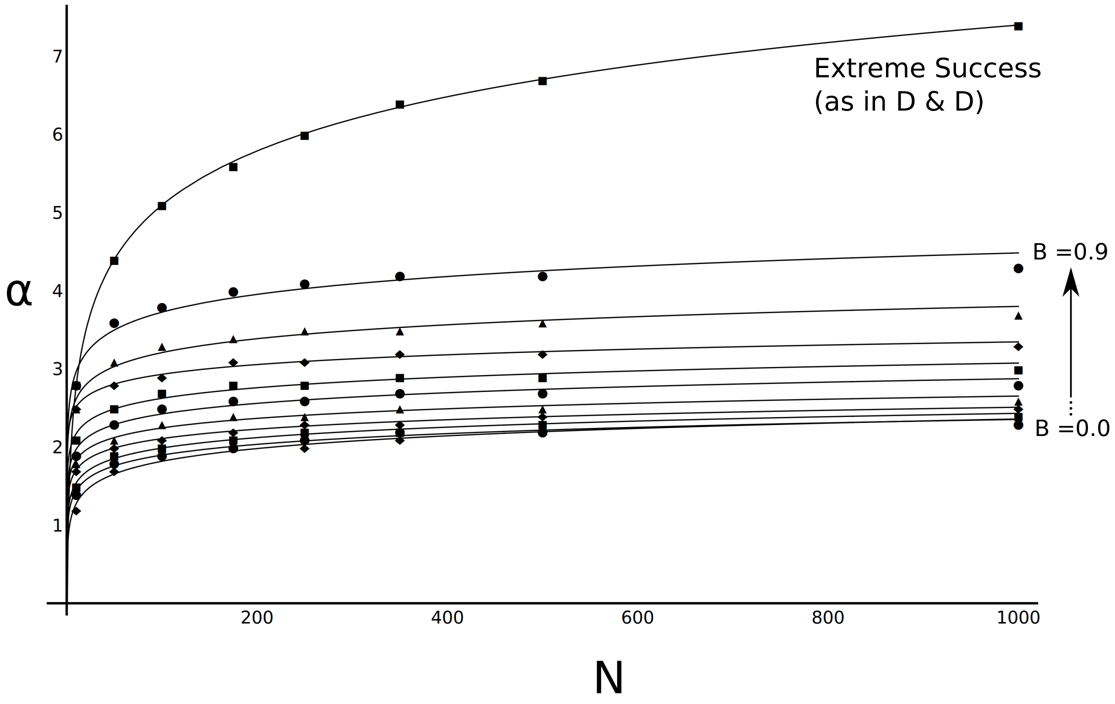

Fig. 5 Critical population size (N) versus transmission inaccuracy (α) for *Frequency-dependent Trimming*, discarding different bottom B fractions of the frequency distribution. For example, in case $B = 0.3$, the 30% lowest z-values are ignored; the offspring selects a parent from the 70% best parents, with a probability proportional to the parent's z-value. Note that the condition where $B = 0.0$ corresponds to *Modest Success*. For purposes of comparison, condition *Extreme Success* is plotted as well.

Surprisingly, the effects of weeding out bad theories are relatively small at the range we are interested in. Even when offspring make a selection just from the 10% percent best theories ($B = 0.9$), little is gained by introducing more individuals in the population once it has reached a threshold of around 350.[7] In fact, only D & D's original condition, i.e. *Extreme Success*, offers a persistent population effect, operative in small and large populations alike. That condition, however, is implausible in light of our earlier criticism (Section 3 and Section 4).

To summarize, our model supports the following conditional claims:

[7] Note that one shouldn't put too much weight on the exact number here. Different parameter settings will result in different thresholds. It is the qualitative point that matters: at a certain threshold, population effects cease to play.

C.1 If *Extreme Success* obtains (very unlikely), for any given initial population size, population growth can be expected to increase the population's average z-value.

C.2 If *Modest Success* or *Frequency-dependent Trimming* obtains (likely, or so we hope), for initial populations of limited size, population growth can be expected to increase the population's average z-value.

C.3 If *Modest Success* or *Frequency-dependent Trimming* obtains (likely, or so we hope), for initial populations of considerable size, population growth should not be expected to increase the population's average z-value.

C.4 If *Conformity* obtains (perhaps more likely than we hope), for any given initial population size, population growth should not be expected to increase the population's average z-value.

C.5 If *Random Copying* obtains (as likely as it is certain what actually is more likely), for any given initial population size, population growth should not be expected to increase the population's average z-value.

7 Conclusion

We have done two things in this paper. First, we have pointed out the shortcomings of a population-dynamic mathematical model deployed by De Cruz & De Smedt to address the collective dimensions of scientific knowledge. In particular, we have argued that their model makes one overly strong tractability assumption, which seriously limits the scope of their argument and which makes little sense empirically. Second, we have developed an agent-based version of De Cruz & De Smedt's model that allowed us to relax the tractability assumption in question. This enabled us to identify five conditional claims about when and when not one should expect growth of scientific communities to spur scientific progress.

Acknowledgements

Research by Krist Vaesen and Wybo Houkes was supported by the Netherlands Organisation for Scientific Research (NWO).

References

Boyd, R. and Richerson, P. (1985). *Culture and the evolutionary process.* University of Chicago Press.

Collins, H. (1985). *Changing order.* University of Chicago Press.

De Cruz, H. and De Smedt, J. (2012). Evolved cognitive biases and the epistemic status of scientific beliefs. *Philosophical Studies*, 157:411–429.

Foley, R. (2001). In the shadow of the modern synthesis? Alternative perspectives on the last fty years of paleoanthropology. *Evolutionary Anthropology*, 10:5–14.

Grim, P. (2009). Threshold phenomena in epistemic networks. In *Proceedings of the AAAI Symposium "Complex Adaptive Systems and the Threshold Effect: Views from the Natural and Social Sciences"*, pages 53–60.

Hartmann, S. and Sprenger, J. (forthcoming). Judgement aggregation and the problem of tracking the truth. *Synthese.*

Henrich, J. (2001). Cultural transmission and the diffusion of innovations. *American Anthropologist*, 103:1–23.

Henrich, J. (2004). Demography and cultural evolution: Why adaptive cultural processes produced maladaptive losses in Tasmania. *American Antiquity*, 69(2):197–21.

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

Henrich, J. and Boyd, R. (2002). On modeling cognition and culture. why cultural evolution does not require replication of representations. *Journal of Cognition and Culture*, 2:87–112.

Henrich, J. and McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology*, 12:123–135.

Hull, D. (1964). The effect of essentialism on taxonomytwo thousand years of stasis. *British Journal for the Philosophy of Science*, 15:314–326.

Kelemen, D. (2004). Are children "intuitive theists"? Reasoning about purpose and design in nature. *Psychological Science*, 15:295–301.

Kitcher, P. (1990). The division of cognitive labor. *Journal of Philosophy*, 87:5–22.

Kuhn, T. (1962). *The Structure of Scientific Revolutions.* University of Chicago Press.

Latour, B. (1987). *Science in action.* Harvard University Press.

List, C. (2005). Group knowledge and group rationality: A judgment aggregation perspective. *Episteme*, 2(1):25–38.

Mesoudi, A. (2009). How cultural evolutionary theory can inform social psychology, and vice versa. *Psychological Review*, 116:929–952.

Mesoudi, A. (2011). An experimental comparison of human social learning strategies: payoff-biased social learning is adaptive but underused. *Evolution and Human Behavior*, 32:334–342.

Mesoudi, A. and Lycett, S. (2009). Random copying, frequency-dependent copying and culture change. *Evolution and Human Behavior*, 30:41–48.

Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175–220.

Payette, N. (2011). For an integrated approach to agent-based modeling of science. *Journal of Artificial Societies and Social Simulation*, 14(4):9.

Powell, A., Shennan, S., and Thomas, M. (2009). Late pleistocene demography and the appearance of modern human behavior. *Science*, 324:1298–1301.

Price, G. (1972). Extension of covariance selection mathematics. *Annals of Human Genetics*, 35:485–490.

Richerson, P. and Boyd, R. (2005). *Not by Genes Alone: How Culture Transformed Human Evolution.* University of Chicago Press.

Ross, L., Lepper, M. R., and Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology*, 32:880–892.

Strevens, M. (2003). The role of the priority rule in science. *Journal of Philosophy*, 100:55–79.

Vaesen, K. (2012). Cumulative cultural evolution and demography. *PLoS ONE*, 7(7):e40989.

Waxman, S. (2005). Why is the concept ‘living thing’ so elusive? concepts, languages, and the development of folkbiology. In Ahn, W. K., Goldstone, R. L., Love, B. C., Markman, A. B., and Wolff, P., editors, *ategorization inside and outside the laboratory. Essays in honor of Douglas L. Medin*. American Psychological Association.

Weisberg, M. and Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, 76:225–252.

Zollman, K. (2007). The communication structure of epistemic communities. *Philosophy of Science*, 74:574–587.

Appendix A: Derivation of Equation (3)

According to Henrich, assuming that everyone copies the most skilled individual, population-level change in mean z-value, $\Delta\overline{z}$, is given by Equation 2, that is,

$$\Delta\overline{z} = \underbrace{z_h - \overline{z}}_{\text{selective transmission}} + \underbrace{\Delta z_h}_{\text{noisy interference}}$$

For a population with skill levels distributed according to a Gumbel distribution, we have

$$z_h = \mu + \beta(\gamma + \ln(N)),$$

where μ is the mode and β the spread of the distribution, and N represents population size. Further, for $\overline{z}$ we have

$$\overline{z} = \mu + \beta\gamma.$$

Finally, the transmission error is given by

$$\Delta z_h = -\alpha + \beta\gamma,$$

where α represents imitation inaccuracy, as represented in Figure 1. Adding all this yields Equation 3 in the main text.

Appendix B: Relaxing D & D's first tractability assumption

What happens if one were to relax D & D's first tractability assumption, and concede that belief selection is subject to the same cognitive biases as imitation? Put differently, what happens when imitators not only make inferior copies, but prior to that, also select inferior beliefs to copy from? Under these conditions, the relatively simple Equation 2 no longer applies. Instead, one would need to deploy the following equation:

$$\Delta\overline{z} = \frac{1}{n}\sum_{1}^{n}\frac{f_i}{\overline{f}}(z_i + \Delta z_i) - \frac{1}{n}\sum_{1}^{n} z_i \tag{4}$$

So instead of having only one best belief h, which is copied by all, one now has several beliefs $\{i, j, ..., n\}$, that all come with their own likelihood of being copied $\{f_i, f_j, ..., f_n\}$, and with their own transmission bias $\{\Delta z_i, \Delta z_j, ..., \Delta z_n\}$. Without making any further assumptions, there is no way of telling how $\Delta\overline{z}$ will evolve over time. In sum, the "best belief" assumption is indeed required for tractability. The only alternative for examining the effects of less selective forms of mentor selection is to switch to agent-based models (see Section 4).

Appendix C: Implementation of the simulations

Simulations were implemented in NetLogo (code available from the authors upon request). Simulations start with a population of $N = 10, 50, 100, 175, 250, 350, 500, 1000$ parents, each with an initial z-level randomly drawn from a Gumbel[10;1]-distribution, and a population of $N = 10, 50, 100, 175, 250, 350, 500, 1000$ offspring individuals, each with an initial z-level of 0.

In each run, models go through two stages: transmission and replacement.

During transmission, each offspring selects—according to the learning biases defined in the main text—one agent from the parent generation, the latter acting as a cultural parent for the former. In particular, the offspring's z-value is given by the parent's z-value, minus the structural transmission inaccuracy α, plus an individual error, randomly drawn from a Gumbel[0;1]-distribution.

During replacement, the offspring generation replaces the parent generation, and the average z-level of the population, $\overline{z}$, is measured.

Models go through 100 runs, after which the overall change in average z-level, $\Delta\overline{z}$, is measured. To account for stochastic variation in simulation outcomes, 1000 iterations were performed and results were averaged across these.