Author accepted manuscript, 10 November 2021. To appear in *Current Biology*

Building better theories

Clare Press[1,2], Daniel Yon[1] and Cecilia Heyes[3,4]

In Brief: In this My Word, Press *et al.* tackle the 'theory crisis' in cognitive science. Using examples of good and not-so-good theoretical practice, they distinguish theories from effects, predictions, hypotheses, typologies, and frameworks in a self-help checklist of seven questions to guide theory construction, evaluation, and testing.

Science is made of ideas and data, theory and observation. Theories can tell us which data to collect, summarise masses of observations, and explore the space of potential explanations. Some theories, the ones that describe mechanisms, can even turn straw into gold — transforming data into explanations of the world around us. Given the central importance of theory, there are good reasons to worry about the lack of ideas in biology[1], and the 'theory crisis' in the cognitive and behavioural sciences[2,3]. It has been argued that this lack of 'novel ideas' is slowing scientific progress[4], but what can be done?

In the long term, we need to look closely at how scientists are trained. Typically, students are taught a great deal about how to collect data and next to nothing about how to formulate and evaluate theory. We force feed techniques and statistics, but, like parental prudes, leave the kids to work out for themselves where theories come from. In the medium term, there is no shortage of detailed, technical advice for those with the freedom and inclination to make radical changes to their science — to get into formal modelling or interdisciplinary collaboration[5–7] — but even these changes can leave the core theoretical problems unresolved[8]. In the short term, we can all up our theoretical game by thinking a little harder about what theories are for, how they should be constructed, and which tests are worth the effort.

In this My Word, we offer a checklist of seven questions we find useful when formulating and testing our own theories, and when thinking through the work of others in our research fields. Theory has several functions[9] but, like many scientists, we want our work to contribute to explanation. Therefore, whenever we say ‘theory’ we mean mechanistic theory — the kind that has the potential to be explanatory[10]. Otherwise, we are broad-minded — a mechanistic theory may be labelled as such or as a ‘model’; it can be formal, for example computational / mathematical, or informal, for example graphical / narrative. Without naming and shaming, we use examples from our own fields of research — social cognition, memory, perception, and action — areas where cognitive science meets neuroscience and evolution.

What is the target?

A theory needs a target or ‘explanandum’ — a phenomenon that exists or occurs in the natural world that the theory is designed to explain. There are very few constraints on the size of the target — classics include magnetism, oxidation, fermentation, and lava flow — but, whatever the size or aspect of the natural world, the target should show some signs of unity, of being a ‘thing’. Say you want to explain imitation. If you define imitation loosely, to encompass all sorts of vaguely social learning, you are unlikely to come up with a successful theory because the cases are probably not produced by the same causal structure. It is unlikely that the same psychological or neurological mechanism enables snails to find food by following the slime trails of other snails and enables people to learn calculus by reading textbooks. On the other hand, if you define imitation tightly — for example, as copying behavioural topography, the way that parts of the body move relative to one another — there is a decent chance all cases are due to the same causal structure; that one mechanism allows imitation of scowls, arabesques and Fosbury Flops[11].

Obviously, the more you know about your target the better, and plenty of important research aims to characterise the target rather than to test theory. What are the triggers, modulators, and inhibitors? What happens when the system is damaged or operating under unusual conditions? How does the system function at different points in development and across species? Not-so-obviously, it is important to keep all this information in mind when formulating or assessing a theory. If a theory works for deliberate but not automatic imitation, or for imitation in primates but not in birds, we need good reasons to regard these as distinct targets rather than areas where the theory fails.

Scientists sometimes treat an effect — an impact of an independent variable on a dependent variable – as an explanatory target. This is risky because, although effects have significant functions in science[12], an effect rarely captures all and only the manifestations of a single causal structure. For example, many psychologists and neuroscientists are interested in the finding that perception is often attenuated during action; for example, we cannot tickle ourselves. However, it is risky to use this 'sensory attenuation' effect as an explanatory target because it can arise via many different routes[13,14]. For instance, self-produced tickle might be perceptually attenuated by general gating mechanisms in the spine, predictive processes that 'cancel out' expected action outcomes, or because our attention is directed differently when we have to act and perceive simultaneously. The causes of sensory attenuation do not have the unity and specificity needed to make it a good explanatory target. It may be better to focus theorising on understanding the targets (for example, prediction) that cause the effect, rather than the effect itself.

Do I have a theory or just a prediction?

'Hypothesis': a neat word with messy consequences. Used as a synonym for both 'theory' and 'prediction', 'hypothesis' can hide the fact that theories and predictions are very different beasts[7]. Theories postulate entities and activities that produce and explain observable phenomena[10]. For example, Baddeley's theory of working memory[15] explains short-term retention of information (the target) with reference to cognitive entities including the 'central executive' and 'phonological loop', and activities performed by these entities such as 'updating', 'binding' and 'inhibiting'. A prediction, on the other hand, is what you expect when you do an experiment or perform a new analysis of existing data. Predictions are sometimes derived from theories, but having a prediction is no guarantee that you have got a theory. Say we give a test of working memory to 5- and 7-year-old children. We might expect the 7-year-olds to do better than the 5-year-olds by pure extrapolation — because previous work shows that 5-year-olds perform better than 3-year-olds on the very same test. These are empirically-based rather than theory-based predictions, like expecting to wake up in the morning because we have done so in the past. Most of us do not have a theory of what causes waking (or sleeping), but we still bet when we go to sleep tonight that we will wake up tomorrow.

Alternatively, we might derive our prediction for this experiment from Baddeley's theory of working memory, which assigns a specific role to executive function in short-term retention. If we have evidence from previous work using different tests — assessing vigilance or inhibitory control, rather than memory — that executive function improves between the ages of 5 and 7, then we could make the theory-based prediction that 7-year-olds will do better than 5-year-olds in our experiment.

Empirically-based and theory-based predictions can be hard to tell apart. For example, psychopharmacologists often investigate how drugs that act on neurotransmitters alter cognition and behaviour. A psychopharmacologist might predict that administering the drug methylphenidate — which enhances the synaptic availability of dopamine — will enhance a volunteer's ability to update working memory. Is this a theory-based prediction about the role of dopamine in working memory? It is tempting to think so, because working memory has been the focus of many theories, and regardless of the nature of the prediction the papers frequently refer to ‘dopaminergic mechanisms’. However, it is sometimes an empirically-based prediction — derived only from previous results, not from a theory about how the system works. Such a theory might postulate that increased synaptic availability of dopamine destabilises representations, making them more amenable to insertions of new information.

Is it a theory or a framework?

A theory is bigger than a prediction and smaller than a framework. A framework is a way of thinking about all or part of the natural world. Theories generate predictions and live inside frameworks. As we are using the terms, theories should be testable, but frameworks can be useful even when no data could show that they are wrong. A nice illustration of this distinction comes from recent ideas in cognitive neuroscience surrounding the ‘Bayesian brain’. Scientists working within this framework suggest that the brain is fundamentally in the game of modelling the outside world — with all aspects of thought and behaviour arising as the brain combines probabilistic top-down expectations with the bottom-up evidence arriving at our senses[16]. The kernel of this idea has proved incredibly fertile, sprouting many specific theories that aim to explain diverse aspects of cognition. For example, Bayesian principles directly inspired the ‘strong prior’ theory of hallucinations, which posits that abnormal experiences like voice-hearing in psychosis emerge because patients give an unusually strong weight to top-down expectations when perceiving their surroundings[17].

Theories of this kind — derived from or inspired by an explicit framework — do make testable predictions. For instance, the ‘strong prior’ theory predicts that patients who hallucinate should also show a stronger reliance on top-down knowledge in other perceptual tasks[18]. But the framework lurking in the background may not be amenable to testing in the same way. For example, proponents of the Bayesian brain framework have noted mathematical proofs which guarantee it is always possible to specify a set of prior beliefs that would make an observed thought or behaviour seem ‘Bayes optimal’ — that is, compliant with the overarching framework[19]. This means that *in principle,* there is no result the framework cannot accommodate, and no pattern of possible results that could disprove it. Thus, our experiments are more likely to be fruitful when they aim to test theories rather than frameworks.

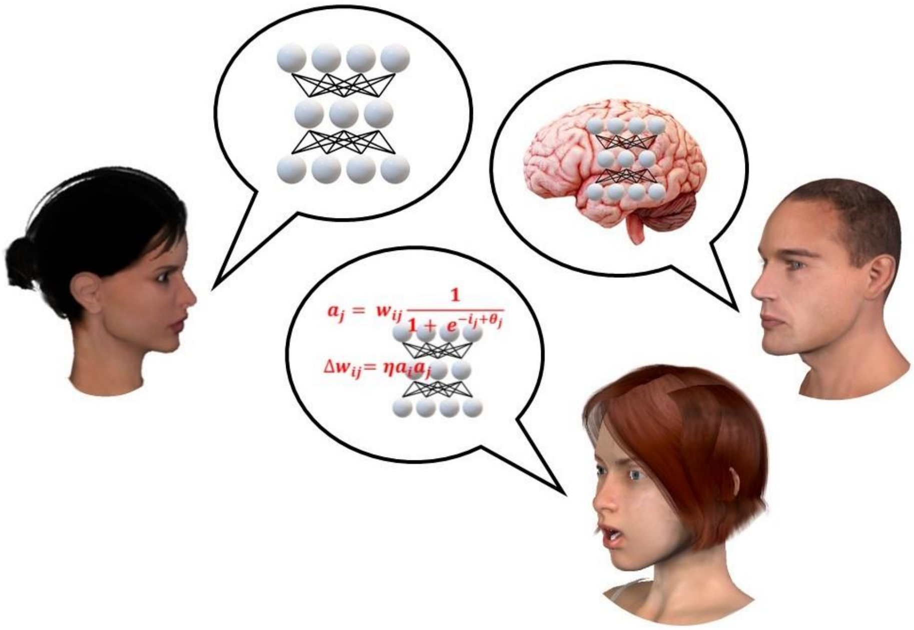
Is the theory pointing at the target?

There are good and not-so-good ways for a theory to fail. A theory fails in a good way when it makes plausible, testable predictions that are not confirmed by the evidence (see below)[6]. It fails in a not-so-good way when, regardless of the evidence, the theory lacks the potential to explain its target. This sometimes happens when typologies are mistaken for theories. Typologies, such as the Linnean classification of plants and animals, can help theory construction by organising and displaying the phenomena to be explained. However, by themselves typologies do not make predictions or offer explanations. Types of social learning — such as 'response facilitation' and 'emulation' — are routinely described as 'mechanisms of social learning', but they are defined purely by their stimulus inputs and behavioural outputs. They are names of effects rather than theories; they are not backed by accounts of the entities and activities in the mind/brain that produce the input-output relationships.

In other cases, a theory misses its target by postulating a causal structure that is as much in need of explanation and evidence as the target itself. When it is discovered, this problem is sometimes called 'begging the question'. An obvious example would be a claim that difficulty in recognising faces is caused by prosopagnosia — the name for difficulty in recognising faces. However, begging the question can be surprisingly difficult to spot because the theoretical entities are typically described in a different language from the target. For example, a central challenge for an imitation mechanism is the conversion of information in one modality (visual inputs hitting the retina) into another (motor commands driving execution of a 'corresponding' action). It was recently suggested that mirror neurons solve the correspondence problem for imitation, but this proposal begs the question. Unless the theory tells us, not just *that* but *how* mirror neurons convert visual input to motor output, the ‘explanation’ is as mysterious as the thing it is supposed to explain; the mirror neuron theory just moves the hard problem of imitation from behaviour into the brain.

Figure 1. Translating theories.

Scientific theories can be expressed in several ways. In our fields (psychology and neuroscience) theories are often expressed in the language of cognitive mechanisms, neural processes and formal equations (or a combination thereof). Often, researchers translate theories from one of these languages into another: describing a cognitive model in neural terms, or reformulating a cognitive model as a computational one. While reformulation can have benefits, there is a danger that this process of translation leads us to think we have created a new theory. This may not be the case if we cannot use it to generate any new predictions (see *Is the theory new?*)


$a_j = w_{ij}\frac{1}{1+e^{-i_j+\theta_j}}$
$\Delta w_{ij} = \eta a_i a_j$

Is the theory new?

Sometimes what appears to be a new theory turns out to be an old theory expressed in a new way. For example, across the last century psychologists have theorised about mechanisms of associative learning — mechanisms that allow us to learn a ringing bell predicts an upcoming food pellet, or that stamping my foot on the left-most pedal tends to make the car stop. The classic Rescorla–Wagner theory argues that animals like us form mental associations between events, adjusting the strength of these links based on patterns we experience (foe xample, strengthening the link between 'bell ringing' and 'food delivery' if these tend to co-occur). This account remains influential because of its ability to explain several experimental phenomena — such as 'blocking', where learning about one predictive event is slowed in the presence of another competing predictor.

In more recent years, however, alternative 'Bayesian' learning accounts have been offered, which suggest that learners use samples of experience to adjust graded beliefs in a 'hypothesis space' that represents how different events in our environment are related. These newer models also predict phenomena like 'blocking' — and thus account for the same experimental results. While the associative and Bayesian theories are formulated using different mathematical frameworks, both suggest that learning depends on the same kind of information (for example, probabilistic co-occurrence). Are the Bayesian theories new? On the one hand, it seems not. Re-expressing an old idea (for example, adjusting associative links) in a new language (for example, updating probabilistic beliefs) does not in and of itself produce a new theory, if the new version can only predict (and account for) the same phenomena.

On the other hand, it is possible that the process of reformulation ultimately generates new predictions — meaning that a new theory is born. In our example, reformulating associative learning in Bayesian terms encouraged theorists to build ideas around variance and uncertainty into the learning process. This leads to distinctive empirical predictions. For example, Bayesian models predict that we should learn faster when the world is more volatile — a prediction, since confirmed, which could not have been derived from the original Rescorla–Wagner theory. Re-expressions are useful if, like this, they generate new empirical predictions. Without this predictive novelty, re-expressions offer identical ideas[20].

Unfortunately, incentive structures that reward novelty and primacy in science discourage clear labelling. They make it tempting to present any reformulation as a new theory — giving the false, time-wasting impression that it can be tested against the original. These cosmetic reformulations can disguise old knowledge and lead to decades spent on identical research programmes that masquerade under different labels. Buyers beware[21].

What shall I test?

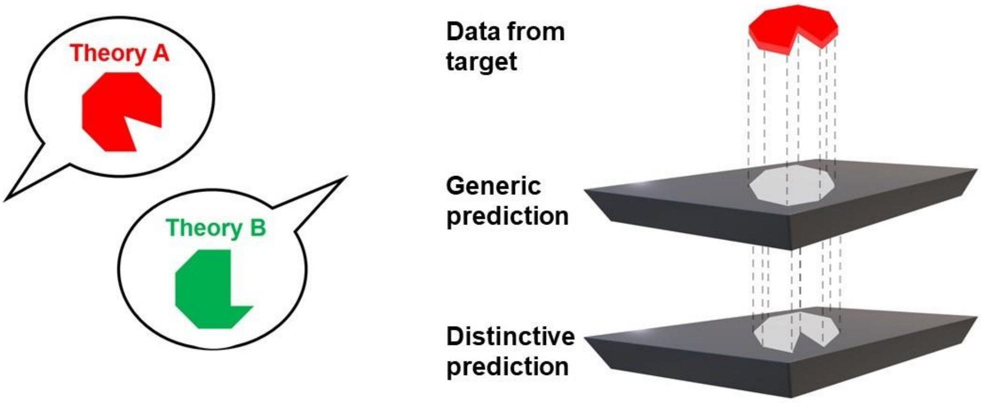
We use a theory to generate an empirical prediction, and when the data are collected we confirm or update the theory. Therefore, we should conduct those empirical tests that are most likely to inform the theories. What are they?

'Underdetermination' of theory by evidence means that a particular pattern of empirical results can always be explained by a range of actual or potential theories. To minimise this problem, we should test a theory's most surprising predictions — the predictions that, if fulfilled, will be inconsistent with the maximum number of alternative theories ('risky testing'[8]). That is, assuming that all theories are based upon some preceding evidence and plausible assumptions, the most surprising predictions are usually those where theories deviate. Pitting theories against each other versus looking for confirmation of only one should not be a pure matter of personal taste — as is often claimed. The latter is more likely to lead to the problem whereby the data do not shape the theoretical landscape.

For example, contrasting models of perceptual expectation disagree as to whether top-down predictions should dampen or sharpen representations in the sensory brain. Intriguingly, both dampening and sharpening models can predict a reduction in overall brain activity when our expectations come true (a generic prediction), meaning it is not particularly informative to look at global signal changes when trying to compare these theories. However, the accounts diverge sharply on whether the information content of sensory brain areas should increase or decrease when inputs are expected (distinctive prediction), meaning that experiments which use information-based measures of brain activity are more powerful for distinguishing between accounts[14].

Figure 2. Risky testing.

Theories differ from each other when they make different claims about the nature of their targets. However, two different theories can share common assumptions about the 'shape' of their targets (generic predictions) while making divergent predictions about other features (distinctive predictions). Experiments designed to test these distinctive predictions tell us the most about which theories are likely to be true (see *What shall I test?*)



Have I listened to Mother Nature?

Many scientists love, and many philosophers loathe, Karl Popper's idea that the hallmark of science is falsification. Philosophers have turned against the principle of falsification because it suggests that a set of observations could logically imply that a theory is wrong. This is implausible because the results of an empirical study depend, not only on the characteristics of the target system, but on a mass of auxiliary assumptions about the validity of measurement and analysis techniques, the proper implementation of the experimental design, and the absence of extraneous influences on the target system. Say we have a theory, X, implying that human newborns can imitate facial expressions, and we run an experiment testing the prediction that newborns will open their mouths more often when they see an adult opening her mouth rather than sticking out her tongue. The results are disappointing; the babies' facial expressions do not vary with those of the adult. Does this mean theory X is wrong? It might, but it could also mean that our measure of facial movements was not subtle enough to pick up the difference between mouth opening and tongue protrusion in newborns, that we did not test enough babies, or that the babies were too uncomfortable or distracted to show us what they could do.

Uncertainty about the causes of a negative result makes it difficult to hear Mother Nature — to work out whether she is whispering that our theory is wrong, or whether her voice is being drowned out by the clanking of our own empirical machinery. Fortunately, the uncertainty can be reduced in several ways. Of course, it helps to use measures that are known to be valid and reliable, and statistical procedures with a good track record in estimating the likelihood of a hypothesis under particular patterns of data (for example, Bayesian statistics that signal whether there is support for the null or simply inconclusive evidence). But risky testing, a remedy that is often overlooked in discussions of 'reproducibility' and the 'replication crisis', is also crucial (see[6]). It is easier to hear Mother Nature in a result predicted by an alternative theory, and risky testing makes research into a conversation rather than a monologue. When research groups test their theories against each other, the members of group A will always be on hand to point out potential interpretive problems when group B declares a loss for theory A or a win for theory B. Initially and at a personal level, these pointers may be unwelcome, but they can inspire more rigorous and creative tests, and deep collegiality among 'rivals'. Healthy science involves 'competition among the cooperators'[22].

Persistent failure to listen when Mother Nature says 'no' can send science down blind alleys, wasting money and labour. An influential theory of cognitive development has survived for 40 years because frequent failures to find imitation in newborns were attributed to more than 20 extraneous factors, including inadequate sample size, inappropriate statistical tests, and the kind of seat in which infants were tested. A recent meta-analysis[23], finding no sign that these factors were modulating an underlying imitation effect, indicates the importance of letting go. It is tempting to protect one's own theory with special pleading and post hoc hypotheses, but in the long term and for the scientific community as a whole, it is better to allow a cherished theory to fail in a good way (see[4]) — to fall nobly in battle with the data.

Conclusion

If theories are like toothbrushes, with no one wanting to use someone else's[24], there are a lot of would-be theorists in science. Our hope is that this My Word will encourage aspirants to generate mechanistic theories that have unified explanatory targets (1), are bigger than predictions (2), more testable than frameworks (3), and do not beg the question (4) or imply that good old ideas are box fresh new ones (5). When theories with these characteristics are subjected to risky testing (6), and the results are interpreted by competing cooperators (7), the hay of venial science turns more rapidly into the gold of explanation.

References

1. Nurse, P. (2021). Biology must generate ideas as well as data. Nature *597*, 305.

2. Oberauer, K. and Lewandowsky, S. (2019). Addressing the theory crisis in psychology. Psychon. Bull. Rev. *26*, 1596–1618.

3. Muthukrishna, M. and Henrich, J. (2019). A problem in theory. Nat. Hum. Behav. *3*, 221–229.

4. Chu, J.S.G. and Evans, J.A. (2021). Slowed canonical progress in large fields of science. Proc. Natl. Acad. Sci. *118*, e2021636118.

5. Robinaugh, D.J., Haslbeck, J.M.B., Ryan, O., Fried, E.I. and Waldorp, L.J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. Perspect. Psychol. Sci. *16*, 725–743.

6. van Rooij, I. and Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. Perspect. Psychol. Sci. *16*, 682–697.

7. Smaldino, P.E. (2020). How to build a strong theoretical foundation. Psychol. Inq. *31*, 297–301.

8. Maatman, F.O. (2021). Psychology’s theory crisis, and why formal modelling cannot solve it. PsyArXiv. doi:10.31234/osf.io/puqvs.

9. Downes, S.M. (2020). Models and modeling in the sciences: A philosophical introduction. (Routledge).

10. Craver, C.F. (2006). When mechanistic models explain. Synthese *153*, 355–376.

11. Heyes, C. (2021). Imitation primer. Curr. Biol. *31*.

12. Keyser, V. (2017). Experimental effects and causal representations. Synthese. doi:10.1007/s11229-017-1633-3.

13. Press, C. and Cook, R. (2015). Beyond action-specific simulation: domain-general motor contributions to perception. Trends Cogn. Sci. *19*, 176–178.

14. Press, C., Kok, P. and Yon, D. (2020). The perceptual prediction paradox. Trends Cogn. Sci. *24*, 13–24.

15. Baddeley, A. (2010). Working memory. Curr. Biol. *20*, R136–R140.

16. Friston, K. (2010). The free-energy principle: A unified brain theory? Nat. Rev. Neurosci. *11*, 127–138.

17. Corlett, P.R., Horga, G., Fletcher, P.C., Alderson-Day, B., Schmack, K., and Powers III, A.R. (2019). Hallucinations and strong priors. Trends Cogn. Sci. *23*, 114–127.

18. Teufel, C., Subramaniam, N., Dobler, V., Perez, J., Finnemann, J., Mehta, P.R., Goodyer, I.M., and Fletcher, P.C. (2015). Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. Proc. Natl. Acad. Sci. *112*, 13401–13406.

19. Friston, K.J. (2017). Precision psychiatry. Biol. Psychiatry Cogn. Neurosci. Neuroimaging *2*, 640–643.

20. Hommel, B., Chapman, C.S., Cisek, P., Neyedli, H.F., Song, J.H., and Welsh, T.N. (2019). No one knows what attention is. Atten. Percept. Psychophys. *81*, 2288–2303.

21. Cooper, R.P., Cook, R., Dickinson, A. and Heyes, C.M. (2013). Associative (not Hebbian) learning and the mirror neuron system. Neurosci. Lett. *540*, 28–36.

22. Campbell, D.T. (1975). On the conflicts between biological and social evolution and between psychology and moral tradition. Am. Psychol. *30*, 1103–1126.

23. Davis, J., Redshaw, J., Suddendorf, T., Nielsen, M., Kennedy-Costantini, S., Oostenbroek, J., and Slaughter, V. (2021). Does neonatal imitation exist? Insights from a meta-analysis of 336 effect sizes. Perspect. Psychol. Sci. *16*, 1373-1397.

24. Mischel, W. (2009). The toothbrush problem. APS Observer *21*.

[1]Department of Psychological Sciences, Birkbeck, University of London, Malet Street, London, WC1E 7HX, UK. [2]Wellcome Centre for Human Neuroimaging, University College London, 12 Queen Square, London, WC1N 3AR, UK. [3]All Souls College, University of Oxford, High Street, Oxford OX1 4AL, UK. [4]Department of Experimental Psychology, University of Oxford, Woodstock Road, OX2 6GG, UK. *Correspondence: c.press@bbk.ac.uk