Theory before the test: How to build high-verisimilitude explanatory theories in psychological science

Iris van Rooij
Donders Institute for Brain, Cognition and Behaviour,
Radboud University, Nijmegen, The Netherlands

Giosuè Baggio
Norwegian University of Science and Technology, Trondheim, Norway

This paper has been accepted for publication

Please cite as:

van Rooij, I. & Baggio, G. (in press). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*.

Abstract. Drawing on the philosophy of psychological explanation (Cummins, 1983; 2000), we suggest that psychological science, by focusing on effects, may lose sight of its primary explananda: psychological capacities. We revisit Marr's (1982) levels-of-analysis framework, which has been remarkably productive and useful for cognitive psychological explanation. We discuss ways in which Marr's framework may be extended to other areas of psychology, such as social, developmental, and evolutionary psychology, bringing new benefits to these fields. Next, we show how theoretical analyses can endow a theory with minimal plausibility even prior to contact with empirical data: we call this the *theoretical cycle*. Finally, we explain how our proposal may contribute to addressing critical issues in psychological science, including how to leverage effects to understand capacities better.

Keywords: theory development; formal modeling; computational analysis; psychological explanation; levels of explanation; computational-level theory; theoretical cycle

1. Introduction

> "(…) a substantial proportion of research effort in experimental psychology isn't expended directly in the explanation business; it is expended in the business of discovering and confirming effects" – Cummins (2000).

Psychological science has a preoccupation with "effects". Yet, effects are explananda (things to be explained), not explanations. The Stroop effect, for instance, does not explain why naming the color of the word 'RED' written in green takes longer than naming the color of a green patch. That just *is* the Stroop effect.[1] The effect itself is in need of explanation. Moreover, effects such as we experimentally test in the laboratory are *secondary* explananda for psychology. Ideally, we do not construct theories *just* to explain effects.[2] The Stroop effect, the McGurk effect, the primacy and recency effects, visual illusions etc. rather serve to arbitrate between competing explanations of the capacities for cognitive control, speech perception, memory, and vision, respectively.

Primary explananda are key phenomena defining a field of study. They are derived from observations that span far beyond, and often even precede, testing of effects in the lab. Cognitive psychology's primary explananda are the cognitive capacities that humans and other animals possess. These include, besides those already mentioned, capacities for learning, language, perception, concept formation, decision making, planning, problem solving, reasoning etc.[3] It is

---

[1] Cummins (2000) uses the McGurk effect to make the same point. We mention the Stroop effect because it is among the least contested effects in psychology, easily replicable in a live in-class demonstration. Yet, our point is that even uncontested, highly replicable effects are not primary explananda in psychology.

[2] This is not to say that in practice this never happens (Newell, 1973). But we believe *good* theoretical practice has a different aim and starting position, as we explain shortly.

[3] One can reasonably debate whether these capacities "carve up" the mind in the right way; and indeed this is a topic of dispute between, e.g., cognitivists and enactivists (van Dijk et al., 2008). Still, few if any cognitive psychologists would maintain that the primary explananda are laboratory effects, instead of

only by the way in which we postulate that such capacities are exercised, that our explanations of capacities come to imply effects. An example is given by Cummins (2000):

> Consider two multipliers, *M1* and *M2*. *M1* uses the standard partial products algorithm (...). *M2* uses successive addition. Both systems have the capacity to multiply (...). But *M2* also exhibits the "linearity effect": computation is, roughly, a linear function of the size of the multiplier. It takes twice as long to compute $24 \times N$ as it does to compute $12 \times N$. *M1* does not exhibit the linearity effect. Its complexity profile is, roughly, a step function of the number of digits in the multiplier.

This example illustrates two points. First, many of the effects studied in our labs are by-products of how capacities are exercised. They may be used to test different explanations of how a system works: e.g., by giving someone different pairs of numerals and by measuring response times, one can test whether or not their timing profile fits *M1* or *M2*, or any different *M'*. Second, candidate explanations of capacities (multiplication) come in the form of different algorithms (e.g., partial products or repeated addition) computing a particular function (i.e., the product of two numbers). Such algorithms aren't devised to explain effects, but posited as *a priori* candidate procedures for realizing the target capacity.

While effects are usually discovered empirically through intricate experiments, capacities (primary explananda) do not need to be discovered in the same way (Cummins, 2000). Just like we knew that apples fall straight from the trees (rather than move upward or sideways) before we had an explanation in terms of Newton's theory of gravity,[4] so too we already know that humans can learn languages, interpret complex visual and social scenes, and navigate dynamic, uncertain, culturally complex social worlds. These capacities are so complex to explain computationally or mechanistically that we do not know yet how to emulate them in artificial systems at human levels of sophistication. The priority should be the discovery not of experimentally constructed effects, but of plausible explanations of real-world capacities. Such explanations may then

---

cognitive capacities, regardless of how one "carves up" the latter. We see the "carving up" as part of the activity of theory development.

[4] We thank Ivan Toni for this analogy.

provide a theoretical vantage point from which to also explain known effects (secondary explananda) and perhaps to guide inquiry into the discovery of new informative ones.

This approach is not the one psychological science has been pursuing the last decades, nor is it what the contemporary methodological reform movement in psychological science has been recommending. Methodological reform so far seems to follow the tradition of focussing on establishing statistical effects, and arguably, the reform has even been entrenching this bias. The reform movement has aimed primarily at improving methods for determining which statistical effects are replicable (cf. debates on preregistration; Nosek et al., 2019; Szolosi et al., 2020), and there has been relatively little concern for improving methods for generating and formalizing scientific explanations (for notable exceptions, see Guest & Martin, 2020; Muthukrishna & Henrich, 2019; Smaldino, 2019; van Rooij, 2019). But if we are already "overwhelmed with things to explain, and somewhat underwhelmed by things to explain them with" (Cummins, 2000), why do psychological scientists expend so much energy hunting for more and more effects? We see two reasons besides tradition and habit.

One is that psychological scientists may believe we first need to build large collections of robust, replicable, uncontested effects before we can even think about starting to build theories. The hope is that, by collecting many reliable effects, the empirical foundations are laid *on which* to build theories of mind and behavior. As reasonable as this seems, without a prior theoretical framework to guide the way, collected effects are unlikely to add up and contribute to the growth of knowledge (Newell, 1973; Anderson, 1990; Cummins, 2000). An analogy may serve to bring the point home. In a sense, trying to build theories on collections of effects is much like trying to write novels by collecting sentences from randomly generated letter strings. Indeed, each novel ultimately consists of strings of letters, and theories should ultimately be compatible with effects. Still, the majority of the (infinite possible) effects are irrelevant for the aims of theory building, just as the majority of (infinite possible) sentences are irrelevant for writing a novel.[5] Also, many of the *relevant* effects (sentences) we may never happen upon by chance, given the vast space of

[5] See Meehl (1997) on the "*crud factor*" (Lykken, 1991): in complex systems often "everything correlates with almost everything else" (p. 393); so the precise *null* hypothesis for statistical effects is seldom if ever true, but most effects may be of little informative value about the key principles of operation of a complex mechanism; many will be by-effects of idiosyncratic conditions.

possibilities.[6] How can we know which effects are relevant and informative, and which ones are not? For this, we first need to build candidate theories and determine which effects they imply.

Another reason, not incompatible with the first, may be that psychological scientists are unsure how to even start to construct theories if not somehow based on effects. After all, building theories of capacities is a daunting task. The space of possible theories is, prima facie, at least as large as the space of effects: for any finite set of (naturalistic or controlled) observations about capacities, there exist (in principle) infinitely many theories consistent with those observations. However, we argue that theories may be built following a *constructive strategy*, while meeting key *plausibility constraints* to rule out from the start theories that are least likely to be true: this we refer to as the *theoretical cycle*. An added benefit of this cycle is that theories, so constructed, already have (minimal) verisimilitude *before* their predictions are tested: this may increase the likelihood that confirmed predicted effects turn out replicable (Bird 2018). The assumptions that have to be added to theories to meet those plausibility constraints (i) provide means for making more rigorous tests of theory possible, and (ii) restrict the number and types of theories considered for testing, channelling empirical research toward testing effects that are most likely to be relevant (more on this later).

This paper aims to make accessible ideas for doing exactly this. We present an approach for building theories of capacities, drawing on a framework that has been highly successful for this purpose in cognitive science: Marr's levels of analysis.

---

[6] How vast is the space of possible effects? We can in principle define an unlimited number of conditions, and compare them to each other. If a 'condition' is some combination of values for situational variables, even if we assume only binary values ('yes' *vs* 'no', 'presence' *vs* 'absence', '>' *vs* '<'), then there are 2^k distincts conditions that we can, in principle, define. For the number of situation variables (from low-level properties of the world, like lighting conditions, to higher-order properties, like day of the week), $k \geq 100$ is a conservative lower bound. Then there are at least $2^{100} \times (2^{100} - 1) > 10^{59}$ distinct comparisons we can make to test for an "effect" of condition; cf. the number of seconds since the birth of the universe ($<10^{18}$) and the world population ($<10^{10}$). Sampling this vast space to discover "effects", without any guidance of substantive theory, we are likely to "discover" many meaningless effects (*crud factor*) and fail to discover actually informative effects.

2. What are theories of capacities?

A capacity is a dispositional property of a system at one of its levels of organization: e.g., single neurons have capacities (firing, exciting, inhibiting) and so do minds and brains (vision, learning, reasoning) and groups of people (coordination, competition, polarization). A capacity is a more or less reliable ability (or disposition or tendency) to transform some initial state (or 'input') into a resulting state ('output').

Marr (1982) proposed that, to explain a system's capacities, we should answer three kinds of questions: (1) what is the nature of the function defining the capacity? (the input-output mapping); (2) what is the process by which the function is computed? (the algorithms computing or approximating the mapping); (3) how is that process physically realized? (e.g., the machinery running the algorithms). Marr called (1) the computational-level theory, (2) the algorithmic-level theory, and (3) the implementational-level theory. Marr's scheme has been occasionally criticized (e.g., McClamrock, 1991) and variously adjusted (e.g., Newell, 1982; Pylyshyn, 1984; Anderson, 1990; Horgan & Thienson, 1996; Poggio 2012; Griffiths et al. 2015), but its gist has been widely adopted in cognitive (neuro)science, where it has supported critical analysis of research practices and theory building (see Baggio et al., 2012a, 2015; Isaac et al., 2014; Krakauer et al., 2017). We see much untapped potential for it also in areas of psychology outside cognitive science.

Following Marr's views, we (propose to) adopt a top-down strategy for building theories of capacities, starting at the computational level. A top-down or function-first approach (Griffiths et al. 2010) has several benefits. First, a function-first approach is useful, if the goal is to 'reverse engineer' a system (Dennett, 1994; Zednik & Jäkel, 2014; 2016):

> an algorithm is (...) understood more readily by understanding the nature of the problem being solved than by examining the mechanism (...) in which it is embodied (Marr, 1982, p. 27; Marr, 1977).

Knowing a functional target ("what" a system does) may facilitate the generation of algorithmic- and implementational-level hypotheses (i.e., how the system "works" computing that function). Reconsider for instance, the multiplication example from the Introduction: by first expressing the

function characterizing the capacity to multiply ($f(x,y) = xy$), one can devise different algorithms realizing this computation (*M1* or *M2*). Lacking any functional target it is difficult or impossible to come up with ways of computing that target. This relates to a second benefit of a function-first approach: it allows one to assess candidate algorithmic or implementational theories for whether they indeed compute or implement that capacity as formalized (Cummins, 2000; Blokpoel, 2018). A third benefit, beyond cognitive psychology, is that social, developmental or evolutionary psychologists may be more interested in using theories of capacities as explanations of patterns of behaviour of agents or groups over time, *in the world*, rather than in the internal mechanisms of those capacities, say, *in the brain or mind*, which is more the realm of cognitive (neuro)science.

Generally, psychological theories of capacities should be (i) mathematically specified and (ii) independent of details of implementation. The strategy is precisely to try to produce theories of capacities meeting these two requirements, unless evidence is available that this is impossible, e.g., that the capacity cannot be modeled in terms of functions mapping inputs to outputs (Marr, 1977; Gigerenzer, 2019). A computational-level theory of a capacity is a specification of input states, let us denote that set I, and output states, O, and the theorized mapping, $f : I \rightarrow O$. For the multiplication example, the input would be the set of pairs of numbers ($\mathbb{N} \times \mathbb{N}$), the range would be the set of numbers ($\mathbb{N}$), and the function $f : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$, would be defined such that $f(a,b) = ab$, for each $a, b \in \mathbb{N}$. The mapping f need not be numerical. It can also be qualitative, structural, or logical. For instance, a computational-level theory of coherence-based belief updating could specify the input as a network $N = (P, C)$ of possible beliefs (modelled by a set of propositions P), where beliefs in the network may cohere or incohere with each other (modelled by positive or negative connections C in the network), a set of currently held beliefs (modeled as truth assignment T: $P \rightarrow$ {*believed to be true*, *believed to be false),* and new information that contradicts, conflicts or is otherwise incoherent with one or more of the held beliefs, D. The output could be a belief revision (modelled as a new truth assignment T') that maintains internal coherence as much as possible while accommodating new information (i.e., $f(N, T, D) = T'$) (for applications in the domain of moral, social, legal, practical and emotional judgements and decision making, see Thagard, 2000, 2006; Thagard & Verbeurgt, 1998).

Historically, Marr's computational level has been often applied to information-processing capacities as studied by cognitive psychologists. Yet, Marr's framework can be extended beyond its traditional domains. First, to the extent that *cognitive* capacities also figure in explanations in other subfields of psychology, Marr's framework naturally extends to these domains, too. We list a few areas where the approach has been fruitfully pursued:

- In Social Cognition, for instance: social categorisation (Klapper et al. 2018), mentalizing or 'theory of mind' (Mitchell, 2006; Baker et al. 2009; Thagard & Kunda, 1998; Michael & MacLeod, 2018), causal attribution (de Houwer & Moors, 2015), moral cognition (Mikhail, 2008), signaling and communication (Moreno & Baggio, 2015; Frank & Goodman, 2012), social attachment (Chumbley & Steinhoff, 2019).
- In Cognitive Development, for instance: (development of) theory of mind (Goodman et al., 2006), probabilistic and causal learning (Bonawitz et al., 2014; Gopnik & Bonawitz, 2015), self-directed learning (Gureckis & Markant, 2012), pragmatic communication (Bohn & Frank, 2019), analogical processing (Gentner, 1983, 2010), and concept formation (Kinzler & Spelke, 2007; Carey, 2009).
- In Cognitive Evolution, for instance: (evolution of) cognitive structures and architectures that aim to account for language, social cognition, reasoning and decision-making (Barrett, 2005; Cosmides & Tooby, 1995; Carruthers, 2006; Fodor, 2000; Lieder & Griffiths, 2019; Marcus, 2006).

Second, and this is a less conventional and less explored idea, the framework can also be applied to non-cognitive or non-individual capacities of relevance to social, developmental, evolutionary psychology, and more. Preliminary explorations into computational level analyses of non-cognitive or non-individual capacities can be found in work by Krafft & Griffiths (2018) on distributed social processes, Huskey et al., 2020, on communication processes, Rich et al. (2020) on natural and/or cultural evolution processes, and van Rooij (2012) on self-organised processes.

In sections 3 and 4, we spell out our approach to theory building using examples. To encourage readers to envisage applications of the approach to their own domains of expertise, and to more complex phenomena than those we can cover here, we provide a stepwise account of what is involved in constructing theories of psychological capacities *in general*. Following

Marr's (1982) successful cash register example, we foresee that more abstract illustrations demonstrating general principles can encourage wider and more creative adoption of these ideas.

3. First steps: Building theories of capacities

We have proposed that theories of capacities may be formulated at Marr's computational level. A computational-level theory specifies a capacity as a property-satisfying computation *f*. This idea applies across domains in psychology, and for capacities at different levels of organisation. How does one build a computational-level theory *f* of some capacity *c*? Or better even, how does one build a *good* computational-level theory?

A first thought may be to derive the *f* from observations of the input-output behaviour of a system having the capacity under study. However, for anything but trivial capacities, where we can exhaustively observe (or sample) the full input domain[7], this is unlikely to work. The reason is that computational-level theories (or *any* substantive theories) are grossly underdetermined by data. The problem that we cannot deduce (or even straightforwardly induce) theories from data is a limitation, or perhaps an attribute, of all empirical science (cf. the Duhem-Quine thesis; Stam 1992; Meehl 1997). Still, one may *abduce* hypotheses, including computational-level analyses of psychological capacities. Abduction is reasoning from observations (not limited to experimental data; more below) to possible explanations (Thagard 1978, 1981; Niiniluoto, 1999; Haig, 2018). It consists of two steps: generating candidate hypotheses (abduction proper), and selecting the "best" explanatory one (inference to the best explanation, IBE). Now, IBE is only as good as the quality of the candidates: the "best" hypothesis might not be any good, if the set does not contain any "good" hypotheses (Blokpoel et al. 2018; van Fraassen, 1985; Kuipers, 2000). So, it is worth building a set of good candidate theories *before* selecting from the set.

---

[7] Even then, coding the input-output mapping as a look-up table isn't explanatory, even if it is descriptive and possibly predictive (within limits). As, for instance, Cummins (2000) notes, the tide tables predict the tides well, but do not explain them. One could make a list of (input, output) pairs ({1, 2}, 2), ({3,4},12), ({12, 3}, 36)), but that is hardly an explanation. Moreover, the list does not allow predictions beyond the observed domain: unless one hypothesizes that the capacity one is observing is 'multiplication' one would not be able to know the value of *x* in ({112, 3.5}, *x*). This is all the more pressing since any observations we would make in a laboratory task setting are typically a *very* small subset of all possible capacity inputs and outputs, and the functions that we are trying to abduce are much more complex than 'multiplication' (e.g., compositionality; Martin & Baggio, 2020; Baroni, 2020).

Abduction is sensitive to background knowledge. We cannot determine which hypotheses are good (verisimilar) by only considering locally-obtained data (e.g., data for a toy scenario in a laboratory task). We should interpret any data in the context of our larger "web of beliefs", which may contain anything we know or believe about the world, including scientific or common sense knowledge. One does not posit a function *f* in a vacuum. What we already know or believe about the world may be used to create a first *rough* set of *candidate* hypotheses on the *f*s for any and all capacities of interest. One can either cast the net wide to capture intuitive phenomena, then chisel away, make precise, and formalize the idea in a well defined *f* (Blokpoel, 2018; van Rooij, 2008). Or alternatively, one can make a first guess, then adjust it gradually based on constraints that one later imposes: the first sketch of an *f* need not be the final one; what matters is how the initial *f* is constrained and refined, and how the rectification process can actually drive the theory forward. Theory building is a *creative* process involving a dialectic of divergent and convergent thinking, informal and formal thinking.

What are the first steps in the process of theory building? Often, the theorist starts with an initial intuitive verbal theory (e.g., that decisions are based on maximizing utilities; that people tend towards internally coherent beliefs, that meaning in language has systematic structure). Next, one should formally define the concepts used in this informal theory (e.g., utilities are numbers, beliefs are propositions, meanings of linguistic expressions can be formalized in terms of functions and arguments; 'numbers', 'propositions', 'functions', and 'arguments' are all well-defined mathematical concepts). The aim of formalization is to cast initial ideas using mathematical expressions (again, of any kind, not just quantitative), so that one ends up with a well-defined function *f*, or at least a *sketch* of *f*. Once this is achieved, follow-up questions can be asked: Does *f* capture one's initial intuitions? Is *f* well-defined (no informal notions are left undefined)? Does *f* have all the requisite properties and no undesirable properties (e.g., inconsistencies)? If inconsistencies are uncovered between intuitions and formalization, the theorist must ask themselves if they are to change their intuitions or the formalisation, or both (van Rooij et al., 2019, Ch. 1; or see van Rooij & Blokpoel, 2020, for a tutorial). In practice, it always takes several iterations to arrive at a complete, unambiguously formalised *f*, given the initial sketch.

Let us illustrate the first steps of theory construction with an example—compositionality, a property of the meanings that can be expressed through language. Speakers of a language know intuitively that the meaning of a phrase, sentence, or discourse is co-determined by the meanings of its constituents: e.g., 'broken glass' means what it does by virtue of the meanings of 'broken' and 'glass', plus the fact that composing an adjective ('broken') and a noun ('glass') in that order is licensed by the rules of English syntax. Compositionality is the notion that people can interpret the meaning of a complex linguistic expression (a sentence etc.) as a function of the meanings of its constituents and of the way those are syntactically combined (Partee, 1995). It is the task of a computational theory of syntax and semantics to formalize this intuition. Like other properties of the outputs of psychological capacities, compositionality comes in degrees (Box 1): a system has that capacity just in case it can produce outputs (meanings) showing a higher-than-chance degree of compositionality, but not necessarily perfect compositionality.

Compositionality is a useful example in this context because it holds across cognitive and non-cognitive domains, and has important social, cultural, and evolutionary ramifications (Table 1), as may be expected from a core property of the human language capacity. Compositionality is therefore used here to illustrate the applicability of Marr's framework across areas of cognitive, developmental, social, cultural, and evolutionary psychology. So, cognitive psychologists may be interested in explaining a person's capacity to assign compositional meaning to a given linguistic expression, like a vision scientist may be interested in explaining how perceptual representations of visual objects arise from representations of features or parts (the 'binding problem'; Table 1, top row). In all cases covered in Table 1, a 'sketch' of a computational theory can be provided as a first step in theory building. A 'sketch' requires that the capacity of interest, the explanandum, is identified with sufficient precision to allow the specification of the inputs, or initial states, and the outputs, or resulting states, of the function *f* to be characterized in full detail in the theoretical cycle. At this stage, we need not say much about the *f* itself, the algorithms that compute it, and the physical systems that implement the algorithms. Also, the 'sketch' need not assume anything about the *goals* (if any) that the capacity serves in each case (Box 1). A discussion of the goals of compositionality, for example, would rather *require* that a 'sketch' is in place. Are compositional languages easier to learn or to use (Kirby et al., 2008; Nowak &

Baggio, 2016)? Is compositional processing one computational resource among others, harnessed only in particular circumstances (Baggio et al., 2012b; Baggio, 2018)? These questions on compositionality's goals are easier to address when a sketch of *f* is in place. In general, questions about the goals and purposes of the capacity need not affect how either *f* or the output property are defined (Table 1; Box 1).

Table 1. Sketches of computational-level analyses of explananda involving compositionality in different domains of psychological science.

Psychological domain	Example explanandum (compositionality)	Computational-level theory (sketch) *f*	Example explananda from other sub-domains
Cognitive	The capacity to assign a compositional meaning to a linguistic expression	*Input*: Complex linguistic expression $u_1,...,u_n$, with elementary parts u_i *Output*: Meaning of input $\mu(u_1,...,u_n)$, such that $\mu(u_1,...,u_n) = \boldsymbol{c}(\mu(u_1),...,\mu(u_n))$, where $\boldsymbol{c}$ is a composition operation	The capacity to recognize complex perceptual objects with parts (binding problem)
Development	The capacity to develop comprehension and production skills for a compositional language	*Input*: Basic sensori-motor, cognitive capacities (e.g., memory, precursors of theory of mind), a linguistic environment *Output*: A cognitive capacity f_c for processing compositional language	The capacity to develop, e.g., fine motor control, abstract arithmetic and geometric skills etc.
Learning	The capacity to learn a (second or additional) compositional language	*Input*: Basic sensori-motor, cognitive capacities; a linguistic environment; a cognitive capacity f_c for compositional language understanding and production *Output*: A new cognitive capacity f_c' that is also compositional	The capacity to learn a new motor skill related to one already mastered, e.g., from ice skating to skiing (skill transfer)
Biological evolution	The capacity to evolve comprehension and production skills for a compositional language	*Input*: A capacity for assigning natural or conventional meanings to signals *Output*: A cognitive capacity f_c for compositional language	The capacity to evolve, e.g., fine motor control, spatial representation, navigation etc.
Social interaction; cultural evolution	The capacity of groups and populations to jointly create new compositional communication codes	*Input*: An arbitrary assignment of meanings to strings *Output*: A compositional assignment of meanings to strings	The capacity of groups or populations to jointly create structured norms and rituals ('culture'); division of labour

4. Further steps: Assessing theories in the theoretical cycle

Once an initial characterization of *f* is in place, one must ask follow-up questions that probe the *verisimilitude* of *f*. This leads to a crucial series of steps in theory development, often overlooked in psychological theorizing. Even if one's intuitive ideas are on the right track and *f* is formalized and internally consistent, it might still lack verisimilitude. A traditional way of testing a theory's verisimilitude is by deriving predictions from *f* and investigating whether or not they are borne out when put to an empirical test. Using empirical tests to update and fine tune a theory is the modus operandi of the *empirical cycle*. We argue that even prior to (and interlaced with) putting computational-level theories to empirical tests, they can be put to *theoretical tests*, in what we call the *theoretical cycle* (Figure 3). Here, one assesses if one's formalization of intuitive, verbal theories satisfies certain theoretical constraints on *a priori* plausibility.

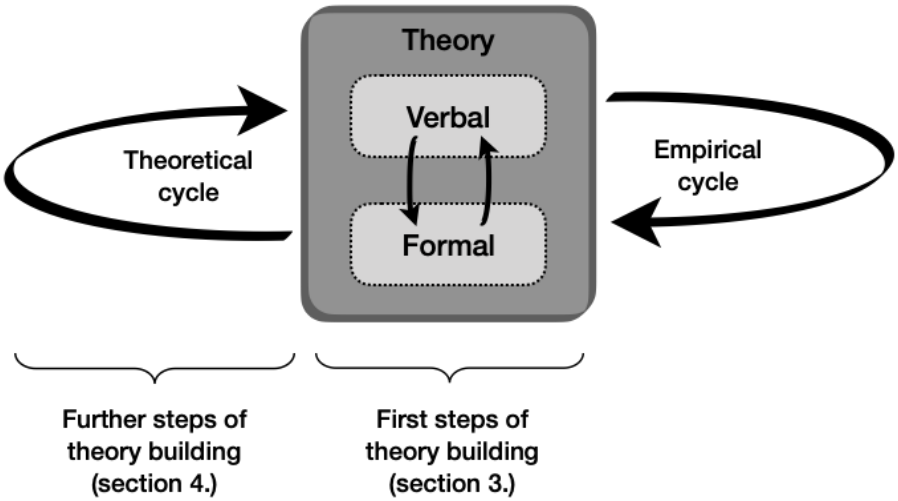

Figure 1. The empirical cycle is familiar to most psychological scientists: the received view is that our science progresses by postulating explanatory hypotheses, empirically testing their predictions (including, but not limited to, effects), and revising and refining the hypotheses in the process. Often in psychological research practice, explanatory hypotheses remain verbal. As explained in section 3, the first steps of (formal) theory building include making such verbal theories formally explicit. In the process of formalisation the verbal theory may be revised and refined. As explained in section 4, theory building does not need to proceed with empirical testing right away. Instead, theories can be subjected to rigorous theoretical tests in what we refer to as the theoretical cycle. This theoretical cycle is aimed at endowing the (revised) theory with greater a priori plausibility (verisimilitude), prior to assessing the theory's empirical adequacy in the empirical cycle. (For a dynamic version of the figure, see: https://redirect.is/x39eor5)

The hypothetical[8] example from the domain of action planning we are going to present appears simple, but as will become clear it will turn out already quite complex. One can think of an organism foraging as engaging in *ordering a set of sites to visit*, starting and ending at its "home base", such that the ordering has overall satisfactory value (e.g., the total cost of travel to the sites in that particular order yields a good trade-off between energy expended for travel and amount of food retrieved). This intuitive capacity can be formalized as follows:[9]

Foraging f

Input: A set of sites $S = \{s_0, s_1, s_2, \ldots, s_n\}$, each site $s_i \in S$ with $i > 0$ hosts a particular amount of food $g(s) \in \mathbb{N}$, and for each pair of sites $s_i, s_j \in S$ there is a cost of travel $c(s_i, s_j) \in \mathbb{N}$.

Output: An ordering $\pi(S) = [s^0, s^1, \ldots, s^n, s^0]$ of the elements in S such that $s^0 = s_0$ and the sum of foods collected at $s^1, \ldots, s^n$ exceeds the total cost of the travel, i.e.,

$$\sum_{s \in S} g(s) \geq c(s^n, s^0) + \sum_{s^i, s^{i+1} \in \pi(S)} c(s^i, s^{i+1})$$

Some arbitrary choices were made here, which might matter for the theory's explanatory value or verisimilitude. E.g., we could have formalized the notion of "good trade-off" by defining it either as (i) maximizing the amount of food collected, given an upper bound on the cost of travel, or as (ii) minimizing the amount of travel, given a lower bound on the amount of food collected, or as (iii) maximizing the difference between the total amount of food collected and the cost of travel. We could also have added free parameters, weighing differentially the importance of maximizing the amount of food and of minimizing the cost of travel.

---

[8] The idea is not fully hypothetical (see Lihoreau et al., 2012), but details here are for illustration only.

[9] The theory admits different orderings as long as they satisfy the output property (the constraint given by the inequality ≥). Formally, functions are always one-to-one or many-to-one, so strictly speaking we are dealing here with a *relation*, or a *computational problem*. This is fine for characterizing capacities, which usually involve abilities to produce outputs that are invariant with respect to some property (cf. "laws of qualitative structure", Simon, 1990; Newell & Simon, 1976). So too in our foraging example: depending on input details, there may be two or more routes of travel that meet the constraint; then producing at least one of them would be exercising the foraging capacity as defined above.

In the theoretical cycle, one explores the assumptions and consequences of a given set of formalization choices, thereby assessing if a computational-level theory is not making unrealistic assumptions or otherwise contradicts prior or background knowledge. As an example we will use a theoretical constraint called *tractability*, but others may be considered (more later). Tractability is the constraint that a theory *f* of a real-world capacity (say, foraging) must be realizable by the type of system under study, given minimal assumptions on its resource-limitations. Tractability is useful for illustrating the theoretical cycle, because it is a fundamental computational constraint; it is insensitive to the type of system implementing the computation; and it applies at all levels of organisation (given basic background assumptions: that computation takes time, that its speed is limited by an upper bound etc.). Tractability is a property of *f* that can be assessed independently of algorithmic- and implementational-level details (Frixione, 2001; van Rooij, 2008, van Rooij et al., 2019): e.g., an organism could solve the foraging problem by deciding on an ordering prior to travel (planning) or it could compute a solution implicitly as it arises from local decisions made while traveling (the same applies to sorting in Figure 2). An assessment of the (in)tractability of *f* is independent of this 'how' of the computation.

Vis-à-vis tractability, the foraging *f* (as stated) turns out *a priori* implausible. If an animal had the capacity *f* (as stated), then it would have a capacity for computing problems known to be intractable: the foraging *f* is equivalent to the known intractable (NP-hard) Traveling Salesperson Problem (Garey & Johnson, 1979). This problem is so hard, even to approximate (Orponen & Heikki, 1987; Ausiello et al., 1999), that all algorithms solving it require time that grows exponentially in the number of sites (n). For all but very small n such a computation is infeasible.

The intractability of an *f* does not necessarily mean that the computational-level theory is wholly misconceived, but it does signal it is underconstrained (van Rooij, 2015). Tractability can be achieved by introducing constraints on the input and/or output domains of *f*. For instance, one could assume that the animal's foraging capacity is limited to a small number of sites (say, $n \leq 5$) or that the animal has the general capacity, also for larger n, but only if the amount of food per site meets some minimum criterion (e.g., $g(s) \geq max(c(s,s')) + \frac{max(c(s,s'))}{n}$ for all s in S).[10] In both

[10] These constraints can be seen as hypothesized "normal conditions" for the proper exercise of a capacity. Intuitive argument for the $g(s) \geq max(c(s,s')) + max(c(s,s'))/n$ constraint: if the amount of food collected at each site exceeds $max(c(s,s'))$, the animal always collects more food than it expends energy

cases, the foraging f is tractable.[11] Theoretical considerations (e.g., tractability) can constrain the theory so as to rule out its most unrealistic versions, effectively endowing it with greater a priori verisimilitude. Moreover, theoretical considerations can yield new empirical consequences, such as predictions about the conditions under which performance breaks down (i.e., $n > 5$ versus $g(s) \leq max(c(s,s')) + \frac{max(c(s,s'))}{n}$; see Blokpoel et al. 2013; Bourgin et al., 2017, for further examples), and can constrain algorithmic-level theorizing (different algorithms exploit different tractability constraints; Blokpoel, 2018; Zednik & Jäkel, 2016). Thus, the theoretical cycle can improve both theory verisimilitude and theory testability.

(In)tractability analyses apply widely, not just to simple examples as above. The approach has been used to assess constraints that render (in)tractable computational accounts for various capacities relevant for psychological science, spanning across domains and levels (Table 1). To name a few areas (for an overview of more examples see Compendium C of van Rooij et al., 2019): coherence-based belief updating (van Rooij et al., 2019); action understanding and 'theory of mind' (Blokpoel et al., 2013; Zeppi and Blokpoel, 2017; van de Pol et al. 2018), analogical processing (Veale & Keane, 1997; van Rooij et al., 2008), problem solving (Wareham, 2017; Wareham, Evans, & van Rooij, 2011), decision-making (Bossaerts & Murawski, 2017; Bossaerts, Yadav, & Murawski, 2019), neural network learning (Judd, 1990), compositionality of language (Pagin, 2003; Pagin & Westerståhl, 2010), evolution, learning or development of heuristics for decision-making (Otworowska et al., 2018, Rich et al., 2019) and evolution of cognitive architectures generally (Rich et al., 2020). This existing research shows that tractability is a widespread concern for theories of capacities relevant for psychological science, and moreover that the techniques of tractability analysis can be fruitfully applied across psychological domains.

Building on other mathematical frameworks, and depending on the psychological domain of interest (Table 1) and on one's background assumptions, computational-level theories can also be assessed for other theoretical constraints, such as computability, physical realizability,

---

for traveling from s^0 to the n sites; to have enough food to cover traveling back from site s^n to s^0, it needs additionally $max(c(s,s'))/n$ at each site.

[11] There may be other constraints that can achieve the same result; we invite interested readers to explore this as an exercise (for guidance, see van Rooij et al., 2019).

learnability, developability, evolvability etc. For instance, reconsider foraging. Above, we have discussed foraging only at the cognitive level (row 1 in Table 1), but one can also ask how a foraging capacity can be learned, developed, or evolved biologically and/or socially (rows 2-5 in Table 1). In some cases, these theoretical constraints can again be assessed by means of analyses analogous to the general form of the tractability analysis illustrated above (see e.g., Clark & Lappin 2013; Chater et al., 2015; Judd, 1990 for learnability; Kaznatcheev, 2019; Valiant, 2009 for evolvability; Rich et al., 2020; Otworowksa et al., for learnability, developability and evolvability). On the one hand, such analyses are all similar in spirit, as they assess the in-principle possibility of the existence of a computational process that yields the output states from initial states as characterised by the computational level theory. But on the other hand they may involve additional constraints that are specific to the real-world physical implementation of the computational process under study. For instance, a learning algorithm running on the brain's wetware needs to meet physical implementation constraints specific to neuronal processes (e.g., Lillicrap, Santoro, Marris, Akerman & Hinton, 2020; Martin, 2020); evolutionary algorithms realized by Darwinian natural selection are constrained to involve biological information that can be passed on genetically across generations (Barton & Partridge, 2000); and, cultural evolution is constrained to involve the social transmission of information across generations and through bottlenecks (Henrich & Boyd, 2001; Mesoudi, 2016; Kirby, 2001; Woensdregt, Cummins & Smith, 2020). Hence, brain processes and biological and cultural evolution are all amenable to computational analyses but may have their own characteristic physical realisation constraints.

By *combining* different theoretical constraints one can narrow down the space of possible functions to those describing real-world capacities (Figure 4). The theoretical cycle thus contributes to early theory validation, and advances knowledge even before putting theories to an empirical test. In practice, it serves as a timely safeguard system: it allows one to detect logical or conceptual errors as soon as possible (intractability of f being one example), and in any case before empirical tests are undertaken.

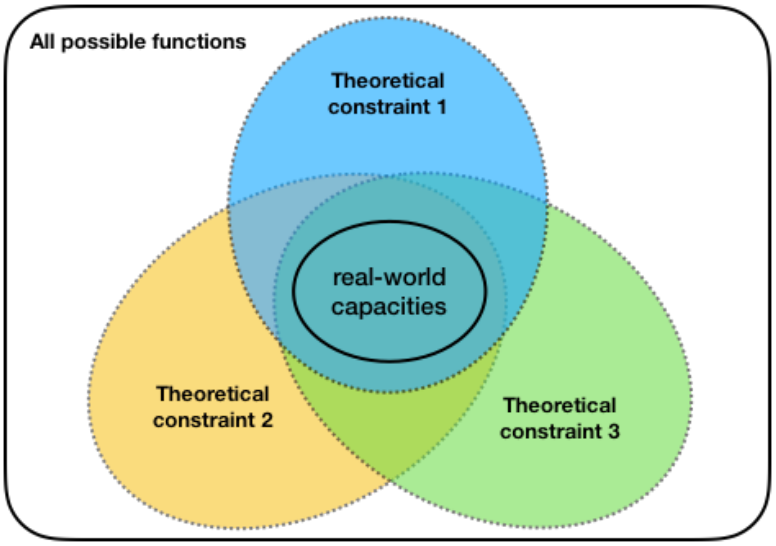

Figure 2. Applying several constraints jointly (e.g., tractability, learnability, evolvability) reduces the space of all possible functions to only those plausibly describing real-world capacities.

5. What effects can do for theories of capacities

We have argued that the primary aim of psychological theory is to explain capacities. But what is the role of *effects* in this endeavor? How are explanations of capacities (primary explananda) and explanations of effects (secondary explananda) related? Our position, in a nutshell, is that inquiry into effects should be pursued in the context of explanatory multi-level theories of capacities and in close continuity with theory development.[12] On this view, the main goal of empirical research, including experimentation (e.g., testing for effects) and computational modeling or simulation, is to further narrow down the space of possible functions, *after* relevant theoretical constraints have been applied. Specifically, good empirical inquiry assumes a set of a priori verisimilar theories of real-world capacities (the intersection in Figure 4), and then proceeds to partitioning this set into *n* subsets. Each subset will (a fortiori) contain high-verisimilitude theory of capacities. However, across subsets, theories may be *empirically different*: theories in subset A may predict effects

[12] This is implicit in our view of interactions between the theoretical and empirical cycles, and has been emphasized in recent philosophy of science: e.g., van Fraassen (2010) discusses "the joint evolution of experimental practice and theory", arguing that "experimentation is the continuation of theory construction by other means" (pp. 111-112).

that are not predicted by theories in subset B, and vice versa. Empirical research, including testing for effects, may allow one to adjudicate among competing theories A *vs* B, thereby eliminating some a priori verisimilar ones that turned out to be implausible a posteriori. To the extent that theories do predict effects, and that those effects are testable experimentally, or via models or simulations, psychology is already well equipped to test those predictions. Here, we are interested in situating effects conceptually in a broader view of inquiry which also encompasses the theoretical cycle: what can effects do for theories of capacities? To answer this question, we need to accept a simple premise: that finding out that a theory is empirically inadequate is more informative if the theory is deemed verisimilar a priori, than if it is already known to be implausible (e.g., the *f* intractable) before the test.

Consider again the multipliers example from Cummins (2000). Multiplication is tractable and the partial products (*M1*) and successive addition (*M2*) algorithms meet minimal constraints of learnability and physical realizability. *M1* and *M2* are plausible algorithmic-level accounts of the human capacity for multiplication. But depending on which algorithm is used on a particular occasion, performance (e.g., the time it takes for one to multiply two numbers) might show either the linearity effect predicted by *M2* or the step-function profile predicted by *M1*. Note that both *M1* and *M2* explain the capacity for multiplication. It is not the computational-level analysis that predicts different effects (the *f* is the same) but rather the algorithmic-level theory. In other cases, effects could follow from one's computational-level theory (for examples from the psychology of language and logical reasoning, see Geurts & van der Silk, 2005; Baggio et al., 2008; Baggio et al., 2015; from the psychology of action, see Blokpoel et al., 2012; Bourgin et al. 2017), or from limitations of resource usage (memory or time), from details of physical realization (some effects studied in neuroscience are of this kind) etc. So, one could classify effects depending on the level of analysis they follow from. This is not a rigid taxonomy but a stepping stone for thinking about the precise links between theory and data. For example, one should want to know *why* an a priori verisimilar theory of a capacity is found to be a posteriori implausible, which is essential in order to decide whether and how to repair the theory. A theory could fail empirically for many reasons, including because its algorithmic- or implementational analyses are incorrect (e.g., the capacity *f* is not realized as the theory says it is), or because the postulated *f* predicts

unattested effects, or it does not predict known effects, even though that *f* has passed relevant tests of tractability etc.

Another dimension in the soft taxonomy of effects suggested by our approach pertains to the degree to which effects are *relevant* for understanding a psychological capacity. Some effects may well be implied by the theory at some level of analysis, but may reflect more or less directly, or not at all, how a capacity is realized and exercised. For example, a brain at work may dissipate a certain amount of heat greater than the heat depleted by the same brain at rest; the chassis of an old vacuum tube computer may vibrate when it is performing its calculations. These effects (heat or vibration) can tell us something about resource use and physical constraints in these machines, but they do not illuminate how the brain or the computer process information. These may sound like extreme examples, but the continuum between clearly irrelevant effects like heat or vibration and the kind of effects studied by experimental psychologists is not partitioned for us in advance: the effects collected by psychological scientists cannot just be assumed to be all equally relevant for understanding capacities across levels of analysis. We may more prudently suppose that they sit on a continuum of relevance or informativeness vis-á-vis a capacity's theory: some can evince how the capacity is realized or exercised, but others are in the same relation to the target capacity as heat is to brain function.

There are no steadfast decision rules prescribing how or where to position specific effects on that continuum, but one may envisage some diagnostic tests. Consider again classical effects, like the Stroop effect, the McGurk effect, primacy and recency effects, visual illusions, priming etc. In each case, one may ask whether the effect reveals a *constitutive aspect* of the capacity and how it is realized and exercised. Diagnostic tests may be in the form of counterfactual questions: take effect X and suppose X was *not* the case; would that entail that the system lacks the capacity attributed to it? Would it entail that the capacity is realized or exercised differently than what the algorithmic or implementation theories hypothesize? For example, would a system *not* subject to visual illusions, i.e., lacking their characteristic phenomenology, also lack the human capacity for visual perception? Would a system that does *not* show semantic priming effects also thereby lack a capacity for accessing and retrieving stored lexical meanings? Our intent is not to force specific answers upon the reader, but to draw attention to the fact that addressing those questions

should enable us to make better informed decisions on *what effects we decide to leverage to understand capacities*. It also matters for whether we can expect effects to be stable across situations or tests. An effect that reveals a constitutive aspect of a capacity (one for which a counterfactual question gets an affirmative answer) may be expected to occur *whenever that capacity is exercised*, and so do effects that are direct manifestations of how the capacity is realized: such effects can therefore also be expected to replicate across experimental tests.

This brings us to a further point. Tests of effects can contribute to theories of capacities to the extent that the information they provide bears on the *structure of the system*, whether it is the form of the *f* it computes, or the virtual machines (algorithms) or physical machines it is running on. The contrast between qualitative (e.g., the direction of an effect) and quantitative predictions (e.g., numerical point predictions) cuts the space of effects in a way that may be useful in natural science (e.g., physics), but is not in psychology. Meehl (1990, 1997) rightly criticized the use of 'weak' tests for theory appraisal, but his call for 'strong' tests (i.e., tests of hypotheses with risky point predictions), if pursued systematically, would entrench the existing focus in psychology on effects, albeit requiring that effects be quantitative. The path to progress in psychological science lies not in a transition from weak qualitative to strong quantitative tests, but rather in *strong tests of qualitative structure*: i.e., tests for effects that directly tap into the workings of a system as it is exercising the capacity of interest. Computational-level theories of capacities are not quantitative but qualitative models: they reveal the internal formal structure of a system, or the *invariants* that allow it to exercise a particular capacity across situations, conditions etc. (Simon 1990). There is usually some flexibility due to free parameters but, as we have argued, principled constraints on those parameters (tractability etc.) may be established via analyses in the theoretical cycle and by explorations of qualitative structure (Navarro, 2019; Pitt et al., 2006; Roberts & Pashler, 2000).

6. Conclusion

Several recent proposals have emphasized the need to potentiate or improve theoretical practices in psychology (Szollosi & Donkin, 2019; Guest & Martin, 2020; Smaldino, 2019; Muthukrishna & Henrich, 2019; Cooper & Guest, 2014; Fried, 2020), while others have focused on clarifying the complex, nuanced relationship between theory and data in scientific inference (Devezer et al.,

2019, 2020; Navarro, 2019; Kellen, 2019; MacEachern & Van Zandt, 2019; Szollosi et al., 2020). Our proposal fits within this landscape, but aims at making a distinctive contribution through the idea of the *theoretical cycle*: before theories are even put to an empirical test, they can be assessed for their plausibility on formal, computational grounds; this requires that there is something to assess formally, i.e., a computational-level analysis of the capacity of interest. In a theoretical cycle, one addresses questions iteratively concerning, e.g., the tractability, learnability, physical realizability etc. of the capacity, as formalized in the theory. However, the theoretical cycle and the empirical cycle will need to always be interlaced: abduced theories can then both be explanatory and meet plausibility constraints (i.e., have minimal verisimilitude) upon testing; and conversely, relevant effects can be leveraged to understand capacities better. The net result is that empirical tests are informative and can play a direct role in improving our theories further.

Acknowledgements: The authors thank Mark Blokpoel, Marieke Woensdregt, Gillian Ramchand, Alexander Clark, Eirini Zormpa, Johannes Algermissen and participants of ReproduciliTea Nijmegen for useful discussions. We are grateful to Travis Proulx and two anonymous reviewers for their constructive comments that helped us improve an earlier version of this paper. Parts of this paper are based on a blogpost “Psychological science needs theory development before preregistration”, written by IvR for the Psychonomic Society.

Box 1 — Possible Objections

One could object that computational-level theorizing is only possible for *cognitive* (sub)systems, while other types of systems require fundamentally different ways of theorizing. We understand this objection in two ways: (a) there is something special about the bio-physical realization of cognition that makes Marr's computational level apply only to it (e.g., that brains are computational in ways other systems are not); (b) computational-level analyses intrinsically assume that capacities have functional purposes or serve goals, while non-cognitive systems (e.g. evolution or emergent group processes) do not.

To counter (a), we note that *multiple realizability* (Miłkowski, 2016; Chalmers, 2011; Dennett, 1995) is the bedrock of Marr's approach: any function *f* can be physically realized in different physical systems, even at different levels of organisation. Consider the capacity for *sorting*. Its inputs are items varying with respect to a quantity that can be ordered (say, the unordered list of numbers 89254631), and gives as output the items in order of magnitude (the ordered list 12345689). This ordering capacity may be physically implemented in several ways. One individual could perform the ordering; or a computer program could do it; or a group of people could implement the capacity in a distributed way. In the latter case, each individual need not have access to the input array, or need not even be aware that they are partaking in a collective behaviour that is *ordering* (see Figure A for an illustration).

We note that a system may produce outputs where the target property only comes in degrees: e.g., the network in Figure A may not always produce a perfect sorting by height, if the people entering the maze do not meet at every node; even then, (i) the outputs tend to show a greater degree of ordering than is expected by chance and (ii) under relevant idealizations (i.e., people meet at every node), the system can still produce a complete, correct ordering: (i)-(ii) together illustrate the system's *sorting capacity*. The fact that target properties come in degrees holds generally; see our discussion of compositionality in the text.

Functions, so conceived, may describe capacities at any level of organisation: we see no reason to reserve computational-level explanations only to levels of organisation traditionally considered in cognitive (neuro-)science. Even within cognitive psychology, the computational level may be (and has been) applied at different levels of organisation: from various neural levels (e.g., feature detection) to persons or groups (e.g., communication). A Marr-style analysis applies regardless of the level of organisation at which the primary explananda are situated, hence it need not be limited to the domain of cognitive psychology.

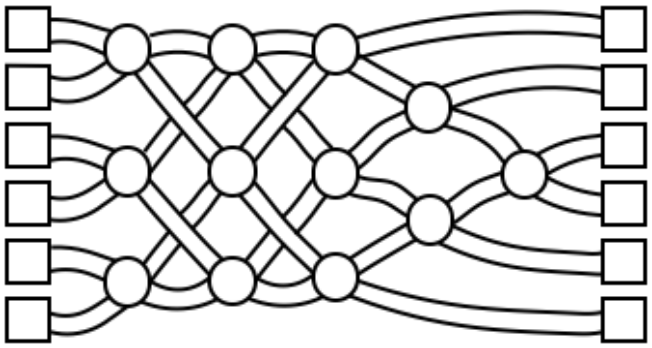

Figure A. A sorting network [adapted from https://www.csunplugged.org/, CC BY-SA 4.0]. Imagine a maze structured this way and six people, walking from left to right, each entering a square on the left. Every time two people meet at any node (circle) they compare their height. The shorter of the two turns left next, and the taller turns right. At the end of the maze, people end up sorted by height. This holds regardless of which six people enter the maze and of what order they enter the maze. Hence, the maze (combined with the sub-capacity of people for making pairwise comparisons) has the capacity for sorting people by height.

To counter objection (b), we note that computational-level theories are usually considered to be normative (e.g., rational or optimal, in a realist or instrumentalist "as if" sense; Chater & Oaksford, 2000; Chater et al., 2003; Danks, 2008, 2013; van Rooij et al., 2018), but that is not formally required. A computational-level analysis is a mathematical object, a function *f*, mapping some input domain to an output domain (Egan, 2010, 2017). Any normative interpretation of an *f* is just an add on—an additional, independent level of analysis (Danks, 2008; van Rooij et al., 2018). Marr did suggest that the theory "contains separate arguments about *what* is computed and *why*" (1982, p. 23), but the meaning of 'why' has been altered over time by (especially, Bayesian) modelers, as requiring that computational theories are idealized optimization functions serving rational goals (Anderson, 1990, 1991; Chater & Oaksford, 1999). Such explanatory strategy may have heuristic value for abducing computational-level theories (see section 3), but it is mistaken to see this strategy as a necessary feature of Marr's scheme.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences, 14* (3): 471-517.

Ausiello, G., Crescenzi, P., Gambosi, G., Kann, V., Marchetti-Spaccamela, A., & Protasi, M. (1999). *Complexity and approximation: Combinatorial optimization problems and their approximability properties*. Springer.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. Cognition, 113(3), 329– 349.

Baggio, G. (2018). *Meaning in the Brain*. Cambridge (MA): The MIT Press.

Baggio, G., Van Lambalgen, M., & Hagoort, P. (2008). Computing and recomputing discourse models: An ERP study. *Journal of Memory and Language*, 59(1), 36-53.

Baggio, G., Van Lambalgen, M., & Hagoort, P. (2012a). Language, linguistics and cognition. *Handbook of the Philosophy of Science*, 14, 325-355.

Baggio, G., Van Lambalgen, M., & Hagoort, P. (2012b). The processing consequences of compositionality. In *The Oxford Handbook of Compositionality* (pp. 655-672). Oxford University Press.

Baggio, G., van Lambalgen, M., & Hagoort, P. (2015). Logic as Marr's computational level: Four case studies. *Topics in Cognitive Science*, 7(2), 287-298.

Baggio, G., Stenning, K., & van Lambalgen, M. (2016). Semantics and cognition. *The Cambridge Handbook of Formal Semantics*, 756-774.

Baroni, M. (2020). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791).

Barrett, H. C. (2005). Enzymatic computation and cognitive modularity. *Mind & Language, 20*(3), 259-287.

Barton, N. & Partridge, L. (2000). Limits to natural selection. *BioEssays, 22*(12), 1075-1084.

Bechtel, W. (2012). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. Psychology Press.

Bechtel, W., & Shagrir, O. (2015). The non‐redundant contributions of Marr's three levels of analysis for explaining information‐processing mechanisms. *Topics in Cognitive Science, 7*(2), 312-322.

Bird, A. (2018). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science,* https://doi.org/10.1093/bjps/axy051.

Blokpoel, M. (2018). Sculpting computational-level models. *Topics in Cognitive Science, 10*(3), 641-648.

Blokpoel, M., Kwisthout, J., van der Weide, T., Wareham, T., & van Rooij, I. (2013). A computational-level explanation of the speed of goal inference. *Journal of Mathematical Psychology*, 570(3-4), 117-133.

Blokpoel, M., Wareham, T., Haselager, P., Toni, I., van Rooij, I. (2018). Deep analogical inference as the origin of hypotheses. *Journal of Problem Solving, 11*(1).

Bohn, M. & Frank, M. C. (2019). The pervasive role of pragmatics in early language. *Annual Review of Developmental Psychology, 1*, 223-249 https://doi.org/10.1146/annurev-devpsych-121318-085037

Bonawitz, E., Denison, S., Griffiths, T., & Gopnik, A. (2014) Probabilistic Models, Learning Algorithms, Response Variability: Sampling in Cognitive Development. *Trends in Cognitive Sciences*, 18, 497-500. doi: 10.1016/j.tics.2014.06.006

Bossaerts, P. & Murawski, C. (2017). Computational complexity and human decision-making. *Trends in Cognitive Sciences 21*(12), 917-929.

Bossaerts, P., Yadav, N., & Murawski, C. (2019). Uncertainty and computational complexity. *Philosophical Transactions of the Royal Society B: Biological Sciences, 374*(1766),

Bourgin, D. & Lieder, F. & Reichman, D. & Talmon, N. & Griffiths, T. (2017). The structure of goal systems predicts human performance. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.

Carey, S. (2009). *The Origin of Concepts*. Oxford: Oxford University Press.

Carruthers, P. (2006). *The Architecture of the Mind*. Oxford: Oxford University Press.

Chalmers, D. J. (2011). A computational foundation for the study of cognition. *Journal of Cognitive Science, 12*, 325–359.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences, 3*(2), 57-65.

Chater, N., & Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese, 122*, 93–131.

Chater, N., Oaksford, M., Nakisa, R., & Redington, M. (2003). Fast, frugal, and rational: How rational norms explain behavior. *Organizational Behavior and Human Decision Processes, 90*, 63–86.

Chater, N., Clark, A., Goldsmith, J. A., & Perfors, A. (2015). Empiricism and language learnability. Oxford: Oxford University Press.

Chumbley, J., & Steinhoff, A. K. (2019). A computational perspective on social attachment. *Infant Behavior and Development, 54*, 85-98.

Clark, A. & Lappin, S. (2013). Complexity in language acquisition. *Topics in Cognitive Science 5*(1), 89-110.

Cooper, R. P. & Guest, O. Implementations are not specifications: Specification, replication and experimentation in computational cognitive modeling. *Cognitive Systems Research, 27*, 42–49 (2014).

Cosmides, L., & Tooby, J. (1995). Beyond intuition and instinct blindness: Toward an evolutionary rigorous cognitive science. *Cognition on cognition*, 69-105.

Cummins, R. (1985). *The Nature of Psychological Explanation*. Cambridge (MA): The MIT Press.

Cummins, R. (2000). "How does it work?" vs. "What are the laws?" Two conceptions of psychological explanation. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 117-145). Cambridge (MA): MIT Press.

Danks, D. (2008). Rational analyses, instrumentalism, and implementations. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for rational models of cognition* (pp. 59–75). Oxford: Oxford University Press.

Danks, D. (2013). Moving from levels and reduction to dimensions anand constraints. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 2124–2129). Oxford: Oxford University Press.

De Houwer, J., & Moors, A. (2015). Levels of analysis in social psychology. In B. Gawronski & G. Bodenhausen (Eds.), *Theory and explanation in social psychology*, New York Guilford (pp. 24-40).

Dennett, D. (1994). Cognitive science as reverse engineering: Several meanings of 'top-down' and 'bottom-up'. In D. Prawitz, B. Skyrms, & D. Westerstahl (Eds.), Logic, methodology & philosophy of science IX (pp. 679–689). Amsterdam: Elsevier Science.

Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meanings of life*. New York: Simon & Schuster.

Devezer, B. et al. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLOS ONE 14*(5): e0216125.

Devezer, B., Navarro, D. J., Vandekerckhove, J, & Buzbas, E. O. (2020). The case for formal methodology in scientific reform. doi 10.1101/2020.04.26.048306

Egan, F. (2010). Computational models: a modest role for content. *Studies in History and Philosophy of Science, 41*, 253-259.

Egan, F. (2017). Function-theoretic explanation and the search for neural mechanisms, in *Explanation and Integration in Mind and Brain Science*, David M. Kaplan (ed.), Oxford University Press (pp. 145-163).

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336*(6084), 998–998.

Fodor, J. (2000). *The mind doesn't work that way: The scope and limits of evolutionary psychology.* Cambridge (MA): The MIT Press.

Frixione, M. (2001). Tractable competence. *Minds and Machines, 11*(3), 379-397.

Fried, E. I. (2020). *Lack of theory building and testing impedes progress in the factor and network literature*. https://doi.org/10.31234/osf.io/zg84s

Garey, M. R., & Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-completeness.* San Francisco: Freeman.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science, 34*(5). 752-775.

Guest, O., & Martin, A. E. (2020). How computational modeling can force theory building in psychological science. https://doi.org/10.31234/osf.io/rybh9

Geurts, B., & van Der Slik, F. (2005). Monotonicity and processing load. *Journal of Semantics*, 22(1), 97-117.

Gigerenzer, G. (2019). How to explain behavior? *Topics in Cognitive Science.* https://doi.org/10.1111/tops.12480

Goodman, N.D., Baker, C.L, Bonawitz, E.B., Mansinghka, V.K., Gopnik, A., Wellman, H., Schulz, L.E., & Tenenbaum, J.B. (2006) Intuitive Theories of Mind: A Rational Approach to False Belief. *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*. Vancouver, Canada.

Gopnik, A., & Bonawitz, E. (2015). Bayesian models of child development. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*(2), 75-86.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8), 357-364.

Gureckis, T. M. & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives in Psychological Science, 7*(5), 464-481.

Haig, B. D. (2018). An abductive theory of scientific method. In *Method Matters in Psychology* (pp. 35-64). Springer.

Henrich, J. & Boyd, R. (2001). Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology, 208*(1), 79-89.

Horgan, T., & Tienson, J. (1996). *Connectionism and the Philosophy of Psychology*. Cambridge (MA): The MIT Press.

Huskey R., Bue, A. C., Eden, A., Grall, C., Meshi, D., Prena, K., Schmälzle, R., Scholz, C., Turner, B. O. & Wilcox, S. (2020). Marr's tri-level framework integrates biological explanation across communication subfields. *Journal of Communication, 70*, 356–378

Isaac, A. M., Szymanik, J., & Verbrugge, R. (2014). Logic and complexity in cognitive science. In Johan van Benthem on logic and information dynamics (pp. 787-824). Springer, Cham.

Jones, J. & Adamatzky, A. (2014). Computation of the travelling salesman problem by a shrinking blob. *Natural Computing*, 2(13).

Judd, J. S. (1990). *Neural network design and the complexity of learning*. Cambridge, MA: MIT Press.

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese, 183*(3), 339.

Kaznatcheev, A. (2019). Computational complexity as an ultimate constraint on evolution. *Genetics*, 212(1), 245-265.

Kellen, D. (2019). A model hierarchy for psychological science. *Computational Brain & Behavior*, 2(3-4), 160-165.

Klapper, A., Dotsch, R., van Rooij, I., & Wigboldus, D. (2018). Social categorization in connectionist models: A conceptual integration. *Social Cognition, 36*(2), 221-246.

Kinzler, K. D., & Spelke, E. S. (2007). Core systems in human cognition. *Progress in Brain Research*, 164, 257-264.

Kirby S (2001) Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation 5*(2):102{110, DOI 10.1109/4235.918430

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences, 105*(31), 10681-10686.

Krafft, P. M., & Griffiths, T. L. (2018). Levels of analysis in computational social science. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, 93(3), 480-490.

Kuipers, T. A. F. (2000). *From instrumentalism to constructive realism*. Dordrecht: Kluwer Academic Publishers.

Lieder, F. & Griths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 1-85.

Lihoreau, M., Raine, N. E., Reynolds, A.M., Stelzer, R.J., Lim, K. S., Smith, A.D., Osborne, J.L., Chittka, L. (2012). Radar tracking and motion-sensitive cameras on flowers reveal the development of pollinator multi-destination routes over large spatial scales. *PLOS Biology, 10*(9).

Lykken, D. T. (1991). What's wrong with psychology anyway? In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology* (vol. 1). Minneapolis (MN): University of Minnesota Press.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C.J. & Hinton, G. (2020), Backpropagation and the brain. *Nature Reviews Neuroscience, 21*, 335–346.

MacEachern, S.N. & Van Zandt, T. (2019) Preregistration of modeling exercises may not be useful. *Comput. Brain Behav. 2*, 179–182

Marcus, G. F. (2006). Cognitive architecture and descent with modification. *Cognition*, 101(2), 443-465.

Marr, D. (1977). Artificial intelligence—a personal view. *Artificial Intelligence, 9*(1), 37-48.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman & Co. (MIT Press, 2010).

Martin, A. E. (2020). A compositional neural architecture for language. *Journal of Cognitive Neuroscience.* Advance online publication. doi:10.1162/jocn_a_01552.

Martin, A. E., & Baggio, G. (2020). Modelling meaning composition from formalism to mechanism. *Philosophical Transactions of the Royal Society B*, 375(1791).

McClamrock, R. (1991). Marr's three levels: A re-evaluation. *Minds and Machines, 1*(2), 185-196.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108-141.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What If There Were No Significance Tests*? (pp. 393-425) Mahwah, NJ : Erlbaum.

Mesoudi, A. (2016). Cultural evolution: A review of theory, findings and controversies. *Evolutionary Biology, 43*(4), 481-497.

Michael, J., & MacLeod, M. A. J. (2018). Computational approaches to social cognition. In *Routledge Handbook of the Computational Mind* (pp. 469-482). Routledge.

Mikhail, J. (2008). Moral cognition and computational theory. *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, 3, 81.

Miłkowski, M. (2016). Computation and multiple realizability. In *Fundamental Issues of Artificial Intelligence* (pp. 29-41). Springer, Cham.

Mitchell, J. P. (2006). Mentalizing and Marr: An information processing approach to the study of social cognition. *Brain Research, 1079*(1), 66-75.

Moreno, M., & Baggio, G. (2015). Role asymmetry and code transmission in signaling games: an experimental and computational investigation. *Cognitive Science*, 39(5), 918-943.

Muthukrishna, M. & Henrich, J. A problem in theory. *Nat. Hum. Behav. 1* (2019).

Navarro, D.J. (2018) Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Comput. Brain Behav. 2*, 28–34

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In Chase, W. G. (Ed.). *Visual Information Processing: Proceedings of the Eighth Annual Carnegie*

Symposium on Cognition, Held at the Carnegie-Mellon University, Pittsburgh, Pennsylvania, May 19, 1972. Academic Press.

Nosek, B. A. Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences, 23*(10), 815-818.

Newell, A. (1982). The knowledge level. *Artificial intelligence, 18*(1), 87-127.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM 19*(3), 113-126.

Niiniluoto, I. (1999). Defending abduction. *Philosophy of science, 66*, S436-S451.

Nowak, I., & Baggio, G. (2016). The emergence of word order and morphology in compositional languages via multigenerational signaling games. *Journal of Language Evolution*, 1(2), 137-150.

Orponen, P. & Heikki, M. (1987). On approximation preserving reductions: complete problems and robust measures. (1987). *Technical report.* [Online: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.64.7246]

Otworowska, M., Blokpoel, M., Sweers, N.,Wareham, T., & van Rooij, I. (2018). Demons of ecological rationality. *Cognitive Science, 42*, 1057-1065.

Pagin, P. (2003). Communication and strong compositionality. *Journal of Philosophical Logic*, 32(3), 287-322.

Pagin, P., & Westerståhl, D. (2010). Compositionality II: Arguments and problems. *Philosophy Compass*, 5(3), 265-282.

Partee, B. H. (1995). Lexical semantics and compositionality. In: An Invitation to Cognitive Science, Ed. D. Osherson, 311–360. Cambridge (MA): MIT Press.

Pelletier, F. J. (1994). The principle of semantic compositionality. *Topoi*, 13(1), 11-24.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, 113(1), 57. Doi: 10.31234/osf.io/43auj

Poggio, T. (2012). The levels of understanding framework, revised. *Perception, 41*(9), 1017-1023.

Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge (MA): MIT press.

Rich, P. & Blokpoel, M. & de Haan, R. & Otworowska, M. & Sweers, M. & Wareham, T. & van Rooij, I. (2019). Naturalism, tractability and the adaptive toolbox. *Synthese*. 10.1007/s11229-019-02431-2.

Rich, P., Blokpoel, M., de Haan, R., van Rooij, I. (2020). How intractability spans the cognitive and evolutionary levels of explanation. Preprint doi: 10.31234/osf.io/adx8j

Roberts, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107, 358-367.

Smaldino, P. (2019). Better methods can't make up for mediocre theory. *Nature, 575*, 9.

Simon. H. A. (1990). Invariants of Human Behavior. *Annual Review of Psychology, 41*, 1-20.

Stam, H. J. (1992). The demise of logical positivism: Implications of the Duhem-Quine thesis for psychology. In *Positivism in Psychology* (pp. 17-24). New York: Springer.

Szollosi, A. & Donkin, C. (2019) *Arrested theory development: The misguided distinction between exploratory and confirmatory research*. Preprint doi: 10.31234/osf.io/suzej

Szollosi, A., Kellen, D., Navarro, D.J., Shiffrin, R., van Rooij, I., Van Zandt, T. Donkin, C. (2020). Is Preregistration Worthwhile? *Trends in Cognitive Sciences, 24*(2), 94-95.

Thagard, P. R. (1978). The best explanation: Criteria for theory choice. *The Journal of Philosophy*, 75(2), 76-92.

Thagard, P. (1981). Peirce on hypothesis and abduction. In *Proceedings of the CS Peirce Bicentennial International Congress* (pp. 271-274). Lubbock, TX: Texas Tech University Press.

Thagard, P. (2006). *Hot thought: Mechanisms and applications of emotional coherence*. Bradford Book.

Thagard, P. & Kunda, Z. (1998) *Making sense of people: Coherence mechanism*s. In S. J. Read & L. C. Miller (Eds.), Connectionist models of social reasoning and social behavior (pp. 3-26)

Thagard, P. and Verbeurgt, K. (1998). Coherence as constraint satisfaction. *Cognitive Science, 22*, 1-24.

Valiant, L. G. (2009). Evolvability. *Journal of the ACM*, 56 (1), 3.

van de Pol, I., van Rooij, I., & Szymanik, J. (2018). Parameterized complexity of theory of mind reasoning in dynamic epistemic logic. Journal of Logic, Language and Information, 27(3). 255–294.

van Dijk, J., Kerkhofs, R., van Rooij, I., & Haselager, P. (2008). Can there be such a thing as embodied embedded cognitive neuroscience? *Theory & Psychology, 13*(8), 297-316.

van Fraassen, B. C. (1985). Empiricism in the philosophy of science. In P. Churchland & C. Hooker (Eds.), *Images of science: Essays on realism and empiricism* (p. 245). Chicago, IL: University of Chicago Press.

van Fraassen, B.C. (2010). *Scientific Representation: Paradoxes of Perspective*. Oxford University Press

van Rooij, I. (2008). The Tractable Cognition thesis. *Cognitive Science, 32*, 939-984.

van Rooij, I. (2015). How the curse of intractability can be cognitive science's blessing. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.) (2015). *Proceedings of the 37th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society*.

van Rooij (2019). Psychological science needs theory development before preregistration. Psychonomic Society.

van Rooij, I. & Blokpoel, M. (2020/forthcoming). Formalising verbal theories: A tutorial by dialogue.

van Rooij, I., Blokpoel, M., Kwisthout, J., Wareham, T. (2019). *Intractability and Cognition: A guide to classical and parameterized complexity analysis*. Cambridge: Cambridge University Press.

van Rooij, I., Wright, C., Kwisthout, J., & Wareham, T. (2018). Rational analysis, intractability, and the prospects of 'as if'-explanations. *Synthese, 195*(2), 491-510

Veale, T., & Keane, M. T. (1997). The competence of sub-optimal theories of structure mapping on hard analogies. In *Proceeding of the 1997 International Joint Conference on Artificial Intelligence*.

Wareham, T. (2017). The roles of internal representation and processing in problem solving involving insight: A computational complexity perspective. *Journal of Problem Solving, 10*(1), 3:1-3:17.

Wareham, T., Evans, P., & van Rooij, I. (2011). What does (and doesn't) make analogical problem solving easy? A complexity-theoretic investigation. *Journal of Problem Solving, 3*(2), 30-71.

Wareham, T. & van Rooij, I (2011). On the computational challenges of analogy-based ceneralization. *Cognitive Systems Research, 12*, 266-280.

Woensdregt, M., Cummins, C., & Smith, K. (2020). A computational model of the cultural co-evolution of language and mindreading. 10.31234/osf.io/3bmsx

Zednik, C., & Jäkel, F. (2014). How does Bayesian reverse-engineering work? In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36).

Zednik, C., & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese, 193*(12), 3951-3985.

Zeppi, A., & Blokpoel, M. (2017). Does mindshaping makes mindreading tractable? Bridging the gap between theory and formal analysis. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society*.