**On worms, mirror neurons and explaining human behavior**

Markus I. Eronen

Department of Theoretical Philosophy, University of Groningen, The Netherlands

m.i.eronen@rug.nl

Orcid: 0000-0003-2028-3338

Laura F. Bringmann

Department of Psychometrics and Statistics, University of Groningen, The Netherlands

l.f.bringmann@rug.nl

Orcid: 0000-0002-8091-9935

Forthcoming in F. Hutmacher & A. Wendt (eds.) *Theory- and Model Building in Psychology*. Palgrave Macmillan.

**Abstract**

A key challenge for psychological theorizing is: At what level(s) should we look for the best explanations and predictions of human behavior? The dominant view is that potentially the best theories to explain behavior are found in neuroscience and biology. Therefore, we start from the perspective of these lower levels, and consider the prospects of developing theories based on those levels. We focus on *C. elegans*, a model organism whose entire neuronal connectome is mapped, and use it as a case study to highlight the difficulties in developing neuroscientific theories of cognition and behavior. We argue that the neuroscientific levels are not promising avenues for constructing explanatory or predictive psychological theories. This is due to the immense complexity and context-dependency of biological systems, even those as simple as *C. elegans*. In order to tackle this complexity, we need higher-level variables that capture stable features of these systems. For psychological theory development, this implies that we should focus on conceptual work and discovering robust phenomena, and take psychology and behavior as the starting points.

**Introduction**

A central aim in psychology and neuroscience is finding theories that can successfully explain or predict human behavior. The lack of such theories has recently led to calls for more emphasis on theoretical and conceptual research (e.g., Borsboom et al., 2021; Eronen & Bringmann, 2021; Guest & Martin, 2021; Levenstein et al., 2023; van Rooij & Baggio, 2021). However, humans and their behavior are extremely complex, and can be studied at many levels, ranging from molecules to social factors. Thus, a key question for efforts at finding theories to explain human behavior is the following: Is there a privileged level, or levels, where we have the best prospects of finding successful theories to predict and explain behavior?

In different forms, this question has preoccupied philosophers and scientists for decades, if not even centuries, and there is no consensus in sight (see, e.g., Eronen, 2021; Fried, 2022; Miller, 2010). However, in practice the dominant view continues to be that the best explanations can potentially be found at the “lower” levels, including neurobiology, genes and biochemistry. This view, which can be called *explanatory reductionism* (Eronen, 2021), has been explicitly endorsed by many neuroscientists (e.g., Gordon, 2016; Insel & Cuthbert, 2015), and some philosophers (e.g., Barwich, 2021; Bickle, 2006). It is also clearly reflected in how research funding is allocated: For example, the National Institute of Mental Health, the biggest funder of mental health research in the world, has overwhelmingly favored neuroscience and genetic research over more traditional psychological research in the past decades (Frances, 2020).

In this chapter, we argue against this view, and show that there is no reason to privilege lower levels when looking for explanations and theories of human behaviour. To make our case stronger, we start from the perspective of lower levels, and from a best-case scenario for the reductionist. More specifically, we focus on the attempts at explaining the behaviour of *C. elegans*, and show that even in this relatively simple biological case, it is extremely hard to find good theories and explanations based on the lower levels. In contrast, context and behaviour remain the best sources for powerful predictions and explanations. We then discuss the implications this has for finding neuroscientific theories and explanations for human behaviour, using the case of mirror neurons as an example. Finally, we argue that theories that explain or predict human behaviour should build on the (folk) psychological and behavioural levels.

Before we start, it is important to emphasize that this chapter is about theories that can (causally) explain or predict human behaviour. Although explanation and prediction are central goals in science, a theory may be successful even if it does not reach these aims, for example if it provides adequate description or understanding. We briefly return to these other aims at the end of this paper, but our main arguments do not directly concern them. Another important clarification is that ‘theory’ should be understood very broadly here, including both local and case-specific theories, such as mechanistic models (see, e.g., Bechtel, 2009), and broader theories, such as the theory of thermodynamics. Also, the notion of ‘levels’ can be understood in a very loose and heuristic sense, roughly referring to things that are similar to each other in relevant ways (e.g., ‘the level of neurons’), and in most cases could also be replaced with ‘size scales’.

**C. elegans**

*Caenorhabditis elegans* is a nematode (roundworm) of roughly 1 mm in length. It has been used as a model organism since the 1960s, starting in the lab of Sydney Brenner, who later received the Nobel prize for his work on *C. elegans* (Ankeny, 2001). What makes this tiny worm attractive for neuroscience is that it has one of the smallest nervous systems in nature, consisting of only 302 neurons. It is the first organism for which both the genome and the connectome or “wiring diagram” of the nervous system have been completely mapped out (Ankeny, 2001). In addition, the behavior, development, and biochemistry of *C. elegans* has been extensively studied. It is also possible to intervene on specific neurons and track the results of the intervention throughout the nervous system, for example with optogenetics, a technique with which specific genetically engineered neurons can be activated with light (Leifer et al., 2011).

This combination of relative simplicity and large amount of relevant knowledge makes *C. elegans* a “reductionist’s delight” (Schaffner, 2000), and an ideal starting point for our discussion of the limits and possibilities of explanatory reductionism in neuroscience. Early on, the study of *C. elegans* was accompanied with much optimism, and the hope was to determine how exactly genes and elements in the nervous system are linked to behavior (Ankeny, 2001). The idea was that explanations for the behavior of the worm could be found based on the connections in the wiring diagram. Many links between behavior and neurobiology have indeed been discovered. Within the nervous system of the worm, several circuits have been identified and linked to behavior (Flavell & Gordus, 2022). For example, among the motor neurons, groups of neurons have been linked to forward motion and backward motion respectively (Roberts et al., 2016), and models have been proposed for circuits of neurons associated with sleep (Bechtel & Bich, 2023; Maluck et al., 2020).

However, in spite of these successes, and the amount of neurobiological knowledge available, understanding of the biological basis of the behavior of *C. elegans* remains limited (see also Krakauer et al., 2017; Stiefel & Brooks, 2019). The dynamics of the nervous system and the “cognition” of *C. elegans* have turned out to be immensely more complicated than originally anticipated.

To start with, the anatomical wiring diagram does not give an adequate picture of how the system functions. The same anatomical connections can play different functional roles depending on the context. Whole-brain imaging and other techniques have revealed that the system is not as stable and invariant as suggested by the wiring diagram, but highly dynamic and evolving, and exhibiting different functions in different sensory and motor contexts (Bargmann, 2012; Flavell & Gordus, 2022). An important factor giving rise to this context-dependence is *neuromodulation*: The activity of a neuron, and the dynamics of a whole circuit, can be modified by molecules released by other neurons (Bargmann, 2012). For example, the switches between dwelling states (slow movement with frequent reversals) and roaming states (faster movement with less reversals) of the worm are partly regulated by a non-synaptic neuromodulatory circuit (Flavell & Gordus, 2022; Roberts et al., 2016). More specifically, due to neuromodulation, the activity of one specific neuron (named AIA) can be correlated with either forward-moving or dwelling behavior, which are basically opposites to each other, depending on whether food odors are present in the environment (Flavell & Gordus, 2022). These are just a few examples: Such changes in function due to neuromodulation are common in the *C. elegans*, and in vertebrate organisms in general (Bargmann, 2012).

The first point that we want draw from this is that even when studying an organism as tiny as a nematode, we need to take into account multiple levels or scales, as well as contextual factors (Stiefel & Brooks, 2019). As Sarma et al. (2018) put it: “Nowhere is the need for models that encompass multiple scales more evident than in the … nematode’s network of 302 neurons, where simple crawling and swimming behaviours remain unexplained … Despite decades of effort, we struggle to describe how individual neurons give rise to such diverse organismal behaviour” (p. 2). The information flow in the circuits of *C. elegans* depends on the context and neuromodulation, and the highly dynamical network of neurons can exhibit far more states and behaviors than the wiring diagram alone would suggest (Bargmann, 2012). Importantly, higher-level or higher-scale factors are not just needed for understanding overall behavior or dynamics of the worm: As the example of neuromodulation shows, we cannot even decipher the function of a *single neuron* without taking its context and the overall dynamics of the system into account.

The second point we would like to emphasize is that we still do not have good explanatory or predictive theories for the behavior of the worm. It is not possible to reliably predict the behavior of the worm based on its nervous system (Hallinen et al., 2021), and the neural mechanisms underlying the decision-making or even the movement patterns of the worm are not well understood (Cook et al., 2019; Gjorgjieva et al., 2014; Hallinen et al., 2021). This is striking, considering the available resources mentioned earlier: The genome and the wiring diagram of the worm is known, it is possible to manipulate specific neurons, and the development and biochemistry of the worm is well studied. Interestingly, more reliable and efficient predictions can be made with models based on behavioral data, for example based on video recordings of the movement patterns of the worms (Costa et al., 2024; Daniels et al., 2019) – a point to which we return later in this chapter.

All in all, this does not bode well for efforts at explaining and predicting far more complex systems, such as the human nervous system. Let us thus move on from worms to our main interest in this chapter, namely human behavior and cognition.

**Mirror neurons**

As neurons are the basic building blocks of the nervous system, an important approach in neuroscience has been building up theories starting from the level of individual neurons. This is evident, for example, in mapping-the-human-connectome projects (Toga et al., 2012), or in mirror neuron research. However, if we take into consideration the problems in explaining the behavior of *C*. *elegans* discussed above, it is very unlikely that this would be a feasible way of developing predictive or explanatory theories of human behavior. To illustrate this, let us look more closely at the case of mirror neuron research, which explicitly starts from the level of neurons.

Mirror neuron research was highly influential and visible in neuroscience in the first decades of this century, and is still actively pursued (Bonini et al., 2022). Mirror neurons were discovered by Giacomo Rizzolatti and colleagues at the University of Parma in the early 1990s, when investigating the activity of individual brain cells of macaque monkeys (Rizzolatti et al., 1996). These researchers noticed a peculiar phenomenon: some neurons in the premotor cortex of the monkeys were firing not only when an action was performed (e.g., a monkey grabbed a banana), but also when the same action was perceived (e.g., the monkey saw another monkey grab a banana).

Rizzolatti and his colleagues soon saw immense potential in their discovery. They proposed that mirror neurons could explain how animals understand the actions of others in a direct way, “from within,” by simulating these actions in the motor area of their own brain (Rizzolatti et al., 1996). In further research, mirror neurons were found in multiple brain regions, and in different animal species (Bonini et al., 2022). In humans, it is difficult to measure the activity of individual neurons (you have to open the skull to do so), but one can measure the activity of the “mirror neuron system” indirectly, for example using functional magnetic resonance imagining (fMRI; Bonini et al., 2022). Over the years, mirror neurons have been proposed to (partially) explain a broad range of phenomena, such as language acquisition, autism, imitation, or empathy (Cook et al., 2014).

But what does the firing of a neuron in a given context really mean? Does it really give away its role in the mechanism or the brain? Importantly, mirror neurons are not morphologically or developmentally distinguishable from other neurons, so the only thing that makes them “mirror” neurons is their pattern of firing (Hickok, 2014). And the problem, considering the *C. elegans* case, is that it is highly implausible that we could derive the role of a neuron just from its firing activity in a specific context. As we saw, the firing of a neuron in *C. elegans* can signify different things, depending on the context and the overall state of the network.

This context-dependence of neuronal behavior is not just some peculiar feature of *C.* *elegans*: instead, neuromodulation is a widespread mechanism in organisms, including mammals (Bargmann, 2012). And due to the fact that neurons in the macaque or human cortex are on average connected to thousands of other neurons (Yuste, 2015), their activity and function in the overall system is likely to be much *more* context-dependent than that of *C. elegans* neurons. Interestingly, also some proponents of mirror neuron research acknowledge this complication: “it is clear that a neuron endowed with a given *functional property* (e.g., discharging during grasping) does not necessarily play a relevant *functional role* in that specific function (e.g., controlling grasping execution). Indeed, the anatomical location and connections of neurons with similar functional properties enable them to play (or not to play) a role in a given function” (Bonini, 2017, p. 59). In other words, the fact that the firing of a neuron is correlated with a specific function or behavior does not imply that it is a neuron *for* that specific function or behavior.

There are also further complications: Mirror neurons do not exclusively fire in the mirroring context, but also often fire to some extent in response to other stimuli (Cook et al., 2014). There is also clear evidence that mirror neurons can be “rewired”: For example, at the end of one study, mirror neurons were firing not just for grasping hand movements, but also when the monkey saw a plier or a stick, which had been used in the experiment and therefore repeatedly observed by the monkey (Heyes & Catmur, 2022; Hickok, 2014). All of this suggests that we probably do not know much at all about the role of mirror neurons in the macaque or human nervous system, and it is not even clear that they form a distinct class of neurons (as the title ‘mirror neuron’ suggests).

More generally, it is highly doubtful whether the level of neurons is a plausible starting point when looking for explanations for higher-level phenomena like imitation, action understanding or autism (Churchland, 2011; Hickok, 2014; Krakauer et al., 2017). Even in the case of *C. elegans*, explanations for complex behaviors require higher-level dynamics and variables (Sarma et al., 2018; Stiefel & Brooks, 2019). In order to get closer to explaining phenomena like action understanding or autism, we first need a good understanding of what these phenomena even are (see also Bringmann et al., 2022). Then, we need to learn more about the cognitive mechanisms underlying or producing these phenomena (Bechtel, 2009; Craver, 2007). And (mirror) neurons that are firing do not yet make up a cognitive mechanism. At best, they are a small part of such a mechanism, and to know what role (if any) these small parts are playing, we need to know more about the overall mechanism, at various levels or scales. When a mirror neuron is firing, what are the neurons it is sending information to? From what kinds of neurons is the mirror neuron receiving input? What is the mechanism that this circuit is part of, and how are things like movements and perceptual contents exactly represented in this mechanism? How does this mechanism, or the interaction of several mechanisms, give rise to the higher-level phenomenon of interest (e.g., action understanding)? Considering that a neuron in the macaque or human cortex can receive inputs from, and send out information to, thousands of other neurons, answers to these questions at the level of neurons are likely to be intractable in practice (see also Rich et al., 2021). Thus, taking higher-level factors into account is necessary, and looking for mechanistic explanations for phenomena such as action understanding or autism at the level of neurons is rather hopeless.

Reductionists and mirror neuron researchers would probably largely agree with this, and point out that (1) the mirror neuron theory is not just about single neurons, but also about the “mirror neuron system”, and (2) cognitive neuroscientists are usually looking for multilevel mechanisms, and not just for explanations starting from the level of neurons. We will consider the second point in the next section, and focus here on the first one. According to mirror neuron proponents, “mirroring” can also be detected with brain imaging at the level of brain areas, by showing that some brain areas are active both when observing and executing an action. However, it is doubtful whether such mirroring patterns have been actually found in fMRI data (Hickok, 2014). And more importantly, what is being measured with fMRI is increased activity in very large populations of neurons: One voxel, which is the smallest unit fMRI can measure, usually contains at least 100 000 neurons (Miller et al., 2014). This is over 300 times the number of neurons in the nervous system of *C. elegans*, which already is complex enough to prevent simple explanations for the roles of neurons. In addition, considering that even in the mirroring areas of monkeys, only around one third of the neurons are actual mirror neurons (Bonini et al., 2022), the majority of neurons in any given voxel or brain area will not be mirror neurons. All of this suggests that the existence of a mirror neuron system in humans is a wild hypothesis at best (see Hickok, 2014; Heyes & Catmur, 2022 for further criticism).

To summarize, the functional role of mirror neurons is unknown. It cannot be derived from the firing pattern only, and we also do not have plausible evidence for the existence of a mirror neuron system or mechanism.[[1]](#footnote-1) In general, looking at the neuronal level is not a plausible starting point for explaining higher-level phenomena such as action understanding, autism, or empathy. And if we already have higher-level explanations, it is not at all clear what neuron-level findings can add to these explanations – a topic to which we return in the next two sections.

**Cognitive neuroscience**

In the previous section, we have argued that starting from the level of neurons is not a feasible approach when explaining human behavior. But what are the prospects for cognitive neuroscience in general, if we consider also higher-level and multiscale approaches?

When discussing the *C. elegans*, we pointed out that we still do not have neuroscientific explanations or theories to causally explain or predict the behavior of this worm. And it is crucial to keep in mind just how small the *C. elegans* brain really is, with only 302 neurons. Now imagine a system that is a hundred times larger, with around 30 000 neurons. This system would already pose far greater challenges for modeling, as many aspects of complexity increase exponentially, not linearly. For example, whereas a network of 300 nodes has around 45 000 possible (undirected) connections between the nodes, a network of 30 000 nodes has around 450 000 000 possible (undirected) connections between nodes.[[2]](#footnote-2) The same pattern can be observed in actual biological cases: whereas *C. elegans* has around 5 000 synaptic connections between neurons (Cook et al., 2019), a fruit fly larva with around 3 000 neurons already has more than 500 000 synapses (Winding et al., 2023).

However, we are still nowhere close to the complexity of the human brain. For that, you have to imagine a system that is around 280 000 000 times larger than the nervous system *of C. elegans*. According to current estimates, the human brain consists of around 85 000 000 000 neurons (von Bartheld et al., 2016), most of these neurons synaptically connecting to thousands or tens of thousands other neurons (Yuste, 2015). If a system that is a “reductionist’s delight” and nearly 300 000 000 times smaller the human brain is already sufficiently complex to make the development of explanatory theories extremely difficult, what are the prospects of developing neuroscientific theories of human behavior? Prima facie, rather hopeless – see also Rich et al. (2021), who argue that inferring the structure of our cognitive systems based on empirical data is a computationally intractable problem.

However, one source of hope is theoretical breakthroughs in neuroscience, which might allow us deal with complexity in a more efficient way. Something like this has happened in physics. For example, the behavior of gas molecules is extremely complex and often computationally intractable. One liter of oxygen contains around 2x1022 gas molecules, which is more than a million times the number of neurons in the human brain. Fortunately, to efficiently explain and predict the behavior of gasses, we do not have to model each individual gas molecule, but we can use macrovariables such as pressure, volume and temperature. As James Woodward (2021) points out, in physics we have found variables and relationships which allow us to effectively ignore or screen off the details at the lower levels. Woodward calls this “conditional irrelevance”: Conditional on the values of the higher-level variables (e.g., pressure and volume), further variations in the lower-level variables (e.g., positions of individual gas molecules) are not relevant for the phenomenon of interest (Woodward, 2021). Once we have found causally explanatory higher-level variables for which conditional irrelevance holds with respect to lower-level variables, we can just use the higher-level variables with little or no loss of explanatory power (see also Gebharter & Eronen, 2023).

If something like this could be found in neuroscience, that would be a significant breakthrough, and a way out of the conundrum described in this section (see also Stiefel & Brooks, 2019). However, there are at least three problems or challenges with this approach: (1) current methods dominating cognitive neuroscience are not very conducive to reaching this aim, (2) cognitive neuroscience is very different from physics, so there might not even be useful higher-level variables to discover, and (3) if we discover them, they may turn out to be psychological or behavioral variables instead of neuroscientific ones. The third point is a topic for the next section, so let us briefly consider the first two.

Regarding the first point: In physics, figuring out useful higher-level variables has required careful theoretical and conceptual work (Bringmann & Eronen, 2016; Chang, 2004; Eronen & Romeijn, 2020). These variables did not just fall out of the data, but were the result of centuries of iterative cycles, where measurement, conceptualization, and mathematical theory interacted (Bringmann et al., 2022; Chang, 2004). In contrast, in cognitive neuroscience there is relatively little focus on theoretical and conceptual work, and often, conceptual unclarity abounds: “cognitive neuroscience is awash in a sea of conflicting terms and concepts” (Poldrack & Yarkoni, 2016, p. 588).

In practice, cognitive neuroscience is largely technologically and empirically driven, which is not a fruitful approach to figuring out the causal structure of the system, or finding the kinds of higher-level variables discussed above. This is nicely illustrated by Jonas and Kording (2017), who applied state-of-the-art neuroscience methods to study a microprocessor. The idea was that if we apply these methods to a system whose structure and organization is known, we can assess how much of the structure can be revealed with those methods. The conclusion was that although the methods produced results that looked plausible and interesting, they provided little insights into the actual functioning and causal structure of the system (Jonas & Kording, 2017). Indeed, as far as we know, cognitive neuroscientists have not yet found any causally explanatory higher-level variables in the nervous system.

Then to the second point. Even if cognitive neuroscience focuses efforts on finding higher-level variables and regularities, it is likely that things will not work out as neatly as in physics. In the end, we have not yet found good causal theories even for the behavior of *C. elegans*, with a nervous system about 300 000 000 times smaller than the human one. One aspect that makes neuroscience different from physics is that biology is very messy: Whereas all gas molecules in a gas are the same, and it is straightforward to take a statistical average of their behavior, each neuron is a distinct biological individual. Each neuron has its own developmental history, extremely complex organization that can differ in significant ways from neurons of the same type, and highly context-dependent function (Gjorgjieva et al., 2016; Huang, 2009, see also previous section). This suggests that the behavior of populations of neurons might not result in higher-level variables and relationships for which conditional irrelevance with regard to lower levels holds. In other words, it may be that nature has no useful higher-level variables to offer in this case, as the behavior of the system is sensitive to the specific features of individual neurons.

This raises a rather unsettling possibility: On the one hand, we cannot predict and explain human behavior at the level of neurons, due to the overwhelming complexity at that level. But on the other hand, we also cannot predict and explain human behavior with higher-level neuroscientific variables, because there are not enough useful higher-level neuroscientific variables to be found. This could be called “ill-behaved complexity”: Complexity that is (nearly) intractable if we start at lower levels, and also does not result in sufficiently well-behaved mechanisms or causal regularities at higher levels.[[3]](#footnote-3) If this, or even something close to it, is the case, the prospects of developing predictive or explanatory theories of human behavior with cognitive neuroscience are very dim. Therefore, it may be better to direct efforts to the psychological and behavioral levels, to which we now turn to.

**Mindreading and mindshaping**

Based on the pessimism of the previous section, one might wonder if predicting human behavior is possible at all. But in practice, we seem to be quite successful in dealing with each other in everyday life, and anticipating each other’s behavior. For example, when my partner shouts “bye!” and turns around, I can predict that they will very likely walk out of the door and close the door behind them. So how do we manage to predict and explain each other, in the face of the intractable complexity of the human brain?

The first part of the answer is that we manage this by tracking real patterns of behavior (Bermúdez, 2003; Dennett, 1987). Even in the case of *C. elegans*, currently the most efficient and reliable way of predicting its behavior is by starting from patterns of behavior, and not from neurobiology (Costa et al., 2024). If there is a lot of food available, the worms will probably dwell; if you remove the food, they will probably switch to food searching behavior (Flavell et al., 2020). Similar robust patterns are evident in human behavior: when the light turns green at a crossroads, I can predict that the pedestrians who have been waiting will start crossing the street, whereas the people in cars will stop driving.

As these examples suggest, *context* is crucial for successful prediction. If you put a traffic light into a lecture hall, people will not start walking when it turns green. When someone is screaming “Help!” in a theater play, it is no cause for concern, but if someone is screaming “Help!” next to a burning building, we should act quickly. And so on. Thus, the combination of contextual information and knowledge of robust patterns of behavior gets us very far in predicting human behavior (see also Bermúdez, 2003; Hutto, 2007).

However, this is not yet enough. Sometimes people think before they act, make intricate plans, collaborate in complex projects, and so. Thus, often we have to figure out what other people think, and to engage in “mindreading”. By mindreading, we refer to the practice of attributing mental states to agents in order to make sense of and predict their behavior (Spaulding, 2020). Consider the following example:

By attributing to Steve the desire for a banana and the belief that there are no more bananas at home but there are some left at the local grocery store, you can: (i) *explain* why Steve has just left home; (ii) *predict* where Steve is heading; and (iii) *coordinate* your behavior with his (meet him at the store, or prepare a surprise party while he is gone). Without mindreading, (i)–(iii) do not come easily—if they come at all. (Barlassina, 2017)

This example illustrates the power of mindreading based on “folk psychological” mental states such as beliefs, desires, and emotions. Although there are wide-ranging debates on the nature of folk psychology and mindreading, it is widely agreed that humans commonly engage in mindreading, that it is crucial for our social lives, and that it often works (Barlassina, 2017; Dennett, 1987; Fodor, 1987; Zawidzki, 2013). If we treat other people as agents with beliefs, desires, and other mental states, we can predict and explain their behavior remarkably well (see also Dennett, 1987, 1991).

This may still seem puzzling considering the near-intractable complexity of the brain discussed in the previous section, and the problems in understanding the cognition of an organism as simple as *C. elegans*. Indeed, it has been pointed out in the literature that figuring out what mental states another person has, and predicting their next actions based on those mental states, should be a computationally intractable problem (see, e.g., Rich et al., 2021; Zawidzki, 2013). There are just too many possible beliefs, thoughts, or other mental states that one could in principle have in any given situation.

To reconcile this difficulty with our evident success in mindreading, several authors have proposed in recent years that we humans are “mindshaping” ourselves and others in order to make life more tractable (Heyes, 2018; McGeer, 2007, 2021; Peters, 2019; Zawidzki, 2013). The hypothesis is that the framework of mindreading and folk psychology has developed through (cultural) evolution as a way to make behavior predictable, and to coordinate behavior so that we can function together efficiently (Heyes, 2018; Slors, 2021; Sterelny, 2012). Humans are a thoroughly cooperative species, and our capability to co-operate efficiently (e.g., when hunting large mammals) is a key factor in our evolutionary success (Sterelny, 2012). Through norms, agreements, rules, and shared vocabulary we *make* our behavior predictable and understandable so that we can better get along, and coordination and cooperation become easier (Peters 2019). As children grow up, they learn about beliefs, thoughts, and intentions, how to appropriately use such concepts, how they ought to think and act in certain situations, and so on (Heyes, 2018; McGeer, 2007). If they fail to follow these norms and conventions, they are corrected or even disciplined.

We can predict the behavior of people based on these norms, rules, and regularities, because people have learned to follow them. As a species, we are very proficient at shaping our minds, shaping other minds, and tracking these shapes or patterns, and all of this results in the predictive success of mindreading. Hence, whereas the human nervous system is (nearly) intractably complex, human behavior is not. It follows real patterns that are (at least partly) the result of mindshaping by pedagogy, rules, conventions, etc. These patterns are to a large extent normative: People follow them because they *ought to* follow them. However, this normative dimension, and the predictive and explanatory power that comes with it, is completely lost if we depart from folk psychology and search for theories and explanations in the brain (see also Borsboom et al., 2019). Norms are not visible in brain states. Therefore, without the normative dimension, we are faced with the ocean of complexity described in the previous section.

To summarize, humans have developed mindshaping and mindreading, which are based on norms and conventions, as a solution for dealing with the potentially intractable complexity of human behavior. It is unclear how we could explain and predict human behavior without relying on context, behavior and the normative patterns resulting from mindshaping.

**Conclusions**

We can conclude that any realistic attempt to construct a theory to predict and explain human behavior has to rely on behavioral patterns and (folk) psychology.[[4]](#footnote-4) But importantly, we should not just accept folk psychological concepts as they are. Instead, they should be amenable to improvement and adjustment based on empirical findings and conceptual work (Bringmann et al., 2022; Eronen, 2020b). In practice, psychological science has been very much empirically and statistically driven (similarly to cognitive neuroscience), which partly explains the lack of successful theories in psychology (Eronen & Bringmann, 2021). In order to make progress, it is fundamentally important to focus on the basic concepts and how to iteratively improve them (Bringmann et al., 2022). However, psychological theory constructors should also keep in mind that finding good psychological theories is extremely hard (see also Eronen & Bringmann, 2021), so progress is likely to be very slow.

A closely related point is that there are other goals in psychology and neuroscience besides prediction and causal explanation of behavior, and these other goals may in fact be more realistic and fruitful. For example, psychology and neuroscience often aim at detecting phenomena or developing purely descriptive models. Also, non-mainstream approaches such as phenomenology (Wertz, 2024) or psychological humanities (Teo, 2017) do not have prediction and causal explanation as goals, but rather, for example, understanding. Our arguments do not apply to these other goals and approaches. In fact, the *C. elegans* case nicely illustrates the importance and usefulness of purely descriptive research. Arguably the greatest success in *C. elegans* research has been developing a very complete descriptive model of the biology of the worm (Ankeny, 2000, 2001), which also provides a springboard for eventual causal theories of the behavior of the worm. Sufficient knowledge of relevant robust phenomena and the organization of the system can even be seen as a prerequisite for developing good explanatory theories of the system (Eronen & Bringmann, 2021). Also, discoveries of robust phenomena can be important as such: Consider examples such as cognitive biases or inattentional blindness, which are very important and societally relevant discoveries, even though we do not know the underlying mechanisms (Eronen, 2020a; Haig, 2013). In addition, discovering new patterns in the interaction of psychological states, behavior and context can be very fruitful and important (e.g., studying in which contexts we are most likely to experience inattentional blindness). In this sense, our arguments can be seen as supporting the importance of non-causal descriptive research and phenomena detection, both in neuroscience and in psychology.

We started this paper with the question: Is there a privileged level, or levels, where we have the best prospects of finding successful theories to predict and explain human behavior? Based on our analysis, we offer a preliminary answer: The best prospects are not at biological or neuroscientific levels, but at the level of psychology and behavior.

**References**

Ankeny, R. A. (2000). Fashioning descriptive models in biology: Of worms and wiring diagrams. *Philosophy of Science*, *67*, S260–S272. https://doi.org//10.1086/392824

Ankeny, R. A. (2001). The natural history of *Caenorhabditis elegans* research. *Nature Reviews Genetics*, *2*(6), 474–479. https://doi.org/10.1038/35076538

Bargmann, C. I. (2012). Beyond the connectome: How neuromodulators shape neural circuits. *BioEssays*, *34*(6), 458–465. https://doi.org/10.1002/bies.201100185

Barlassina, L. (2017). Folk psychology as mental simulation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* *(Summer 2017 Edition)*. https://plato.stanford.edu/archives/sum2017/entries/folkpsych-simulation/

Barwich, A.-S. (2021). Imaging the living brain: An argument for ruthless reductionism from olfactory neurobiology. *Journal of Theoretical Biology*, *512*, 110560. https://doi.org/10.1016/j.jtbi.2020.110560

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology*, *22*(5), 543–564. https://doi.org/10.1080/09515080903238948

Bechtel, W., & Bich, L. (2023). Using neurons to maintain autonomy: Learning from *C. elegans*. *Biosystems*, *232*, 105017. https://doi.org/10.1016/j.biosystems.2023.105017

Bermúdez, J. L. (2003). The domain of folk psychology. *Royal Institute of Philosophy Supplements*, *53*, 25–48. https://doi.org/10.1017/S1358246100008250

Bickle, J. (2006). Reducing mind to molecular pathways: Explicating the reductionism implicit in current cellular and molecular neuroscience. *Synthese*, *151*(3), 411–434. https://doi.org/10.1007/s11229-006-9015-2

Bonini, L. (2017). The extended mirror neuron network: Anatomy, origin, and functions. *The Neuroscientist*, *23*(1), 56–67. https://doi.org/10.1177/1073858415626400

Bonini, L., Rotunno, C., Arcuri, E., & Gallese, V. (2022). Mirror neurons 30 years later: Implications and applications. *Trends in Cognitive Sciences*, *26*(9), 767–781. https://doi.org/10.1016/j.tics.2022.06.003

Borsboom, D., Cramer, A. O. J., & Kalis, A. (2019). Brain disorders? Not really: Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, *42*. https://doi.org/10.1017/S0140525X17002266

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to basics: The importance of conceptual clarification in psychological science. *Current Directions in Psychological Science*, *31*(4), 340–346. https://doi.org/10.1177/09637214221096485

Bringmann, L. F., & Eronen, M. I. (2016). Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory and Psychology*, *26*(1), 27–43. https://doi.org/10.1177/0959354315617253

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.

Churchland, P. S. (2011). *Braintrust: What neuroscience tells us about morality*. Princeton University Press.

Cook, R., Bird, G., Catmur, C., Press, C., & Heyes, C. (2014). Mirror neurons: From origin to function. *Behavioral and Brain Sciences*, *37*(2), 177–192. https://doi.org/10.1017/S0140525X13000903

Cook, S. J., Jarrell, T. A., Brittin, C. A., Wang, Y., Bloniarz, A. E., Yakovlev, M. A., Nguyen, K. C. Q., Tang, L. T.-H., Bayer, E. A., Duerr, J. S., Bulow, H. E., Hobert, O., Hall, D. H., & Emmons, S. W. (2019). Whole-animal connectomes of both *Caenorhabditis elegans* sexes. *Nature*, *571*(7763), 63–71. https://doi.org/10.1038/s41586-019-1352-7

Costa, A. C., Ahamed, T., Jordan, D., & Stephens, G. J. (2024). A Markovian dynamics for *Caenorhabditis elegans* behavior across scales. *Proceedings of the National Academy of Sciences*, *121*(32), e2318805121. https://doi.org/10.1073/pnas.2318805121

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.

Daniels, B. C., Ryu, W. S., & Nemenman, I. (2019). Automated, predictive, and interpretable inference of *Caenorhabditis elegans* escape dynamics. *Proceedings of the National Academy of Sciences*, *116*(15), 7226–7231. https://doi.org/10.1073/pnas.1816531116

Dennett, D. C. (1987). *The intentional stance*. MIT Press.

Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy*, *88*(1), 27–51. https://doi.org/10.2307/2027085

Eronen, M. I. (2020a). Causal discovery and the problem of psychological interventions. *New Ideas in Psychology*, *59*. 100785. https://doi.org/10.1016/j.newideapsych.2020.100785

Eronen, M. I. (2020b). Interventionism for the intentional stance: True believers and their brains. *Topoi*, *39*(1), 45–55. https://doi.org/10.1007/s11245-017-9513-5

Eronen, M. I. (2021). The levels problem in psychopathology. *Psychological Medicine*, *51*(6), 927–933. https://doi.org/10.1017/S0033291719002514

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Eronen, M. I., & Romeijn, J.-W. (2020). Philosophy of science and the formalization of psychological theory. *Theory & Psychology*, *30*, 786–799. https://doi.org/10.1177/0959354320969876

Flavell, S. W., & Gordus, A. (2022). Dynamic functional connectivity in the static connectome of *Caenorhabditis elegans*. *Current Opinion in Neurobiology*, *73*, 102515. https://doi.org/10.1016/j.conb.2021.12.002

Flavell, S. W., Raizen, D. M., & You, Y.-J. (2020). Behavioral states. *Genetics*, *216*(2), 315–332. https://doi.org/10.1534/genetics.120.303539

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT Press.

Frances, A. (2020). The lure of ‘cool’ brain research is stifling psychotherapy. Aeon. https://aeon.co/ideas/the-lure-of-cool-brain-research-is-stifling-psychotherapy

Fried, E. I. (2022). Studying mental health problems as systems, not syndromes. *Current Directions in Psychological Science*, *31*(6), 500–508.

Gebharter, A., & Eronen, M. I. (2023). Quantifying proportionality and the limits of higher-level causation and explanation. *The British Journal for the Philosophy of Science*, *74*(3), 573–601.

Gjorgjieva, J., Biron, D., & Haspel, G. (2014). Neurobiology of *Caenorhabditis elegans* locomotion: Where do we stand? *Bioscience*, *64*(6), 476–486. https://doi.org/10.1093/biosci/biu058

Gjorgjieva, J., Drion, G., & Marder, E. (2016). Computational implications of biophysical diversity and multiple timescales in neurons and synapses for circuit performance. *Current Opinion in Neurobiology*, *37*, 44–52. https://doi.org/10.1016/j.conb.2015.12.008

Gordon, J. A. (2016). On being a circuit psychiatrist. *Nature Neuroscience*, *19*, 1385–1386.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Haig, B. D. (2013). Detecting psychological phenomena: Taking bottom-up research seriously. *The American Journal of Psychology*, *126*(2), 135–153. https://doi.org/10.5406/amerjpsyc.126.2.0135

Hallinen, K. M., Dempsey, R., Scholz, M., Yu, X., Linder, A., Randi, F., Sharma, A. K., Shaevitz, J. W., & Leifer, A. M. (2021). Decoding locomotion from population neural activity in moving *C. elegans*. *eLife*, *10*, e66135. https://doi.org/10.7554/eLife.66135

Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press.

Heyes, C., & Catmur, C. (2022). What happened to mirror neurons? *Perspectives on Psychological Science*, *17*(1), 153–168. https://doi.org/10.1177/1745691621990638

Hickok, G. (2014). *The myth of mirror neurons: The real neuroscience of communication and cognition*. W. W. Norton & Co.

Huang, S. (2009). Non-genetic heterogeneity of cells in development: More than just noise. *Development*, *136*(23), 3853–3862. https://doi.org/10.1242/dev.035139

Hutto, D. D. (2007). The narrative practice hypothesis: Origins and applications of folk psychology. *Royal Institute of Philosophy Supplements*, *60*, 43–68.

Insel, T. R., & Cuthbert, B. N. (2015). Brain disorders? Precisely. *Science*, *348*(6234), 499–500. https://doi.org/10.1126/science.aab2358

Jonas, E., & Kording, K. P. (2017). Could a neuroscientist understand a microprocessor? *PLOS Computational Biology*, *13*(1), e1005268. https://doi.org/10.1371/journal.pcbi.1005268

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. https://doi.org/10.1016/j.neuron.2016.12.041

Leifer, A. M., Fang-Yen, C., Gershow, M., Alkema, M. J., & Samuel, A. D. T. (2011). Optogenetic manipulation of neural activity in freely moving *Caenorhabditis elegans*. *Nature Methods*, *8*(2), 147–152. https://doi.org/10.1038/nmeth.1554

Levenstein, D., Alvarez, V. A., Amarasingham, A., Azab, H., Chen, Z. S., Gerkin, R. C., Hasenstaub, A., Iyer, R., Jolivet, R. B., Marzen, S., Monaco, J. D., Prinz, A. A., Quraishi, S., Santamaria, F., Shivkumar, S., Singh, M. F., Traub, R., Nadim, F., Rotstein, H. G., & Redish, A. D. (2023). On the role of theory and modeling in neuroscience. *Journal of Neuroscience*, *43*(7), 1074–1088. https://doi.org/10.1523/JNEUROSCI.1179-22.2022

Maluck, E., Busack, I., Besseling, J., Masurat, F., Turek, M., Busch, K. E., & Bringmann, H. (2020). A wake-active locomotion circuit depolarizes a sleep-active neuron to switch on sleep. *PLOS Biology*, *18*(2), e3000361. https://doi.org/10.1371/journal.pbio.3000361

McGeer, V. (2007). The regulative dimension of folk psychology. In D. D. Hutto & M. Ratcliffe (Eds.), *Folk psychology re-assessed* (pp. 137–156). Springer. https://doi.org/10.1007/978-1-4020-5558-4\_8

McGeer, V. (2021). Enculturating folk psychologists. *Synthese*, *199*(1–2), 1039–1063. https://doi.org/10.1007/s11229-020-02760-7

Miller, D. J., Balaram, P., Young, N. A., & Kaas, J. H. (2014). Three counting methods agree on cell and neuron number in chimpanzee primary visual cortex. *Frontiers in Neuroanatomy*, *8*, 36. https://doi.org/10.3389/fnana.2014.00036

Miller, G. A. (2010). Mistreating psychology in the decades of the brain. *Perspectives on Psychological Science*, *5*(6), 716–743. https://doi.org/10.1177/1745691610388774

Peters, U. (2019). The complementarity of mindshaping and mindreading. *Phenomenology and the Cognitive Sciences*, *18*(3), 533–549. https://doi.org/10.1007/s11097-018-9584-9

Poldrack, R. A., & Yarkoni, T. (2016). From brain maps to cognitive ontologies: Informatics and the search for mental structure. *Annual Review of Psychology*, *67*, 587–612. https://doi.org/10.1146/annurev-psych-122414-033729

Rich, P., De Haan, R., Wareham, T., & Van Rooij, I. (2021). How hard is cognitive science? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*. https://doi.org/10.31234/osf.io/k79nv

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*(2), 131–141. https://doi.org/10.1016/0926-6410(95)00038-0

Roberts, W. M., Augustine, S. B., Lawton, K. J., Lindsay, T. H., Thiele, T. R., Izquierdo, E. J., Faumont, S., Lindsay, R. A., Britton, M. C., Pokala, N., Bargmann, C. I., & Lockery, S. R. (2016). A stochastic neuronal model predicts random search behaviors at multiple spatial scales in *C. elegans*. *eLife*, *5*, e12572. https://doi.org/10.7554/eLife.12572

Sarma, G. P., Lee, C. W., Portegys, T., Ghayoomie, V., Jacobs, T., Alicea, B., Cantarelli, M., Currie, M., Gerkin, R. C., Gingell, S., Gleeson, P., Gordon, R., Hasani, R. M., Idili, G., Khayrulin, S., Lung, D., Palyanov, A., Watts, M., & Larson, S. D. (2018). OpenWorm: Overview and recent advances in integrative biological simulation of *Caenorhabditis elegans*. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1758), 20170382. https://doi.org/10.1098/rstb.2017.0382

Schaffner, K. F. (2000). Behavior at the organismal and molecular levels: The case of *C. elegans*. *Philosophy of Science*, *67*(S3), S273–S288. https://doi.org/10.1086/392825

Slors, M. (2021). A cognitive explanation of the perceived normativity of cultural conventions. *Mind & Language*, *36*(1), 62–80. https://doi.org/10.1111/mila.12265

Spaulding, S. (2020). What is mindreading? *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(3), e1523. https://doi.org/10.1002/wcs.1523

Sterelny, K. (2012). *The evolved apprentice: How evolution made humans unique*. MIT Press.

Stiefel, K. M., & Brooks, D. S. (2019). Why is there no successful whole brain simulation (yet)? *Biological Theory*, *14*(2), 122–130. https://doi.org/10.1007/s13752-019-00319-5

Teo, T. (2017). From psychological science to the psychological humanities: Building a general theory of subjectivity. *Review of General Psychology*, *21*(4), 281–291. https://doi.org/10.1037/gpr0000132

Toga, A. W., Clark, K. A., Thompson, P. M., Shattuck, D. W., & Van Horn, J. D. (2012). Mapping the human connectome. *Neurosurgery*, *71*(1), 1–5. https://doi.org/10.1227/NEU.0b013e318258e9ff

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682–697. https://doi.org/10.1177/1745691620970604

von Bartheld, C. S., Bahney, J., & Herculano-Houzel, S. (2016). The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. *Journal of Comparative Neurology*, *524*(18), 3865–3895. https://doi.org/10.1002/cne.24040

Wertz, F. J. (2024). Phenomenology and natural science: What kind of a discipline is psychology? *Theory & Psychology*, *34*(3), 405–424. https://doi.org/10.1177/09593543241248128

Winding, M., Pedigo, B. D., Barnes, C. L., Patsolic, H. G., Park, Y., Kazimiers, T., Fushiki, A., Andrade, I. V., Khandelwal, A., Valdes-Aleman, J., Li, F., Randel, N., Barsotti, E., Correia, A., Fetter, R. D., Hartenstein, V., Priebe, C. E., Vogelstein, J. T., Cardona, A., & Zlatic, M. (2023). The connectome of an insect brain. *Science*, *379*(6636), eadd9330. https://doi.org/10.1126/science.add9330

Woodward, J. (2021). Explanatory autonomy: The role of proportionality, stability, and conditional irrelevance. *Synthese*, *198*(1), 237–265. https://doi.org/10.1007/s11229-018-01998-6

Yuste, R. (2015). From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, *16*(8), 487–497. https://doi.org/10.1038/nrn3962

Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. MIT Press.

1. The mirror neuron theory has been closely associated with the simulation theory of mindreading and with embodied and enactive approaches to cognition. However, these do not depend on the mirror neuron theory, and our criticism is not aimed at them. [↑](#footnote-ref-1)
2. Calculated with the equation (*n*(*n*-1))/2, where *n* is the number of nodes, see for example https://en.wikipedia.org/wiki/Complete\_graph [↑](#footnote-ref-2)
3. Woodward (2021) also mentions this possibility, and refers to it as “model chaos”. [↑](#footnote-ref-3)
4. In philosophical terms, we are defending a strong form of epistemic or explanatory antireductionism in psychology. However, this does not imply or require any ontological antireductionism, or even realism regarding (folk) psychology. The predictive power of (folk) psychological states is compatible with a range of ontological positions, ranging from Dennettian instrumentalism to realism (see Eronen, 2020b for more). [↑](#footnote-ref-4)