

Target Article

Cite this article: Murayama K, Jach HK. (2025) A critique of motivation constructs to explain higher-order behavior: We should unpack the black box. *Behavioral and Brain Sciences* 48, e24: 1–57. doi:10.1017/S0140525X24000025

Target Article Accepted: 5 January 2024

Target Article Manuscript Online: 18 January 2024

Commentaries Accepted: 30 June 2024

Keywords:

causality; computational modeling; essentialism; incentives; intrinsic motivation; measurement model; metatheory; reward

What is Open Peer Commentary? What follows on these pages is known as a Treatment, in which a significant and controversial Target Article is published along with Commentaries (p. 14) and an Authors' Response (p. 50). See bbsonline.org for more information.

Corresponding author:

Kou Murayama;

Email: k.murayama@uni-tuebingen.de

A critique of motivation constructs to explain higher-order behavior: We should unpack the black box

Kou Murayama^{a,b}  and Hayley K. Jach^a 

^aHector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany and

^bResearch Institute, Kochi University of Technology, Kochi, Japan

k.murayama@uni-tuebingen.de; <https://motivationsciencelab.com/>

hayleyjach@gmail.com

Abstract

The constructs of motivation (or needs, motives, etc.) to explain higher-order behavior have burgeoned in psychology. In this article, we critically evaluate such high-level motivation constructs that many researchers define as causal determinants of behavior. We identify a fundamental issue with this predominant view of motivation, which we call the *black-box problem*. Specifically, high-level motivation constructs have been considered as causally instigating a wide range of higher-order behavior, but this does not explain what they actually are or how behavioral tendencies are generated. The black-box problem inevitably makes the construct ill-defined and jeopardizes its theoretical status. To address the problem, we discuss the importance of mental computational processes underlying motivated behavior. Critically, from this perspective, motivation is not a unitary construct that causes a wide range of higher-order behavior – it is an emergent property that people construe through the regularities of subjective experiences and behavior. The proposed perspective opens new avenues for future theoretical development, that is, the examination of how motivated behavior is *realized* through mental computational processes.

1. Introduction

From the inception of psychology and across the ensuing century, constructs of motivation have played a critical role in explaining human behavior (Hull, 1943; McDougall, 1909). Researchers have considered motivation as one of the most essential ingredients in our mind, addressing the fundamental question of why people initiate certain behaviors in the first place (Kanfer & Chen, 2016). Initially the constructs of motivation were mostly proposed for basic behavior such as eating (e.g., hunger drive) or mating (e.g., sex drive). However, later years have seen increasing use of motivation constructs to explain higher-order behavior (though such use was observed in early years; e.g., Murray, 1938). Nowadays, there is a plethora of “high-level motivation constructs” in psychology, including (but not limited to) the need for competence, relatedness, and autonomy (Deci & Ryan, 1985), the need to belong (Baumeister & Leary, 1995), self-affirmation motive (Steele, 1988), desire for status (Anderson, Hildreth, & Howland, 2015), self-enhancement motive (Sedikides & Strube, 1997), achievement motive (McClelland, Atkinson, Clark, & Lowell, 1976), and intrinsic–extrinsic motivation (Deci & Ryan, 1985).

In the current opinion article,¹ we provide a critical analysis of these motivation constructs to explain higher-order human behavior. Specifically, we cast doubt on the theoretical status of high-level motivation in the sense of a construct that directly influences complex behavior. Rather, we contend that such high-level motivation is a subjective construal or emergent property of underlying mental computational processes which determine behavior. From this “psychological construction” perspective, we clarify both strengths and weaknesses of high-level motivation constructs, and offer a new avenue of research that has attracted almost no attention in the past: Theoretical analysis of how motivation is *realized* through mental computational processes.

2. Motivation for higher-order behavior: Definitions and clarifications

2.1. Definition

The definition of motivation varies across different times and fields (e.g., Kleinginna & Kleinginna, 1981; Madsen, 1974), but one common definition is that motivation *energizes* and *directs* behavior (Lewin, 1942; Locke & Latham, 2004; Reeve, 2017; Simpson & Balsam, 2016; VandenBos & American Psychological Association, 2007; Weiner, 1992). Energization

means that motivation instigates or initiates action or behavior (Elliot, 2023), which shall be called the “spring to action” of behavior (James, 1890), a force that impels behavior (Descartes, 1955), or “energy behind our actions” (Wigfield, Muenks, & Eccles, 2021). Direction means that it guides and channels behavior in a certain way. The former aspect of motivation is often referred to as “motives” (Anderson et al., 2015; Atkinson & Raynor, 1978; McClelland et al., 1976) or “needs” (Deci & Ryan, 1985; Dweck, 2017; Maslow, 1943; Stevens & Fiske, 1995). The latter aspect of motivation is referred to as “goals” or “values” (Austin & Vancouver, 1996; Eccles & Wigfield, 2002; Elliot & Fryer, 2008; Fishbach & Ferguson, 2007). Combined together, motivation is conceptualized as a determinant of a certain set of behaviors (Fig. 1A).

Our critical analysis of motivation constructs mainly concerns the former aspect, the function of energization (see also Hinde, 1960, for another critique). Regardless of one’s theory of motivation, motivation is almost always used to explain the initiation (or the intention of the initiation) of behavior; as such, energization is often regarded as the definitive aspect of motivation (e.g., Madsen, 1974). The direction aspect of motivation is also often considered a fundamental aspect of motivation, but we view this aspect as somewhat subsumed within the energization aspect (see also Elliot, 2023). In fact, most of the extant constructs of high-level motives or needs emphasize energization in terms of what they are motivated *for* (i.e., direction). For example, an achievement motive represents the motivation toward a high standard of excellence (McClelland et al., 1976) and a need for autonomy directs people toward fulfilling a sense of agency (Ryan & Deci, 2017). It is difficult to imagine motives or needs that instigate behavior in a completely nonspecific manner, other than a few limited examples (e.g., general Pavlovian instrumental transfer; Corbit & Balleine, 2005). We will briefly revisit the direction aspect of motivation in a later section.

The concept of motivation to explain basic human behavior (e.g., mating, consumption of foods) was initially subject to criticisms concerning its operationalization (e.g., Koch, 1941). Although these points overlap with our criticism to some degree, our main criticism is aimed at the motivation constructs that explain a broad range of higher-order behavior, which we shall call high-level motivation constructs.

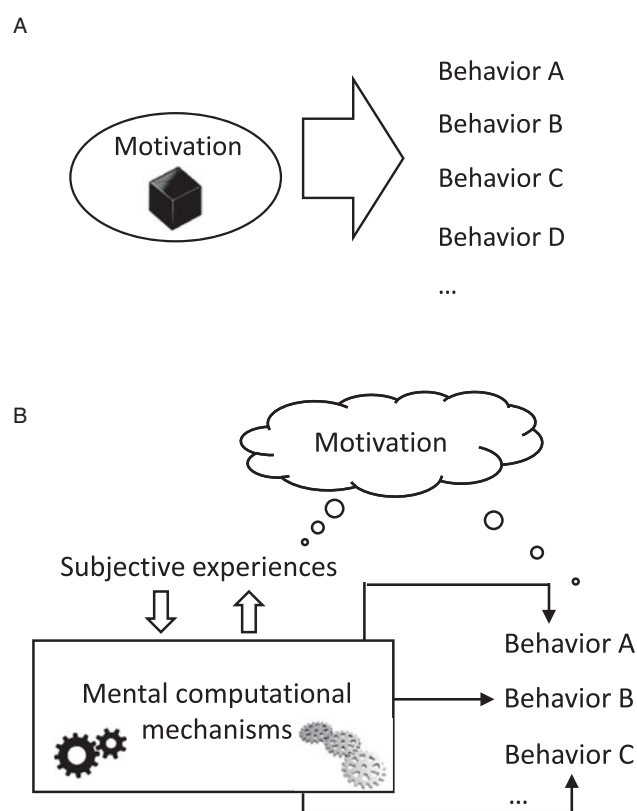


Figure 1. Motivation as the (black box) determinant of behavior (A) and motivation as psychological construction of underlying mental computational processes (B). Graphics are available under creative commons licenses (<https://creativecommons.org/>), downloaded from https://upload.wikimedia.org/wikipedia/commons/d/d9/A_black_box.svg, https://commons.wikimedia.org/wiki/File:Gear_-_Noun_project_7137.svg, and <https://commons.wikimedia.org/wiki/File:Gears-686316.jpg>.

2.2. An example

To render our criticism concrete and easy to understand, we first show one example of how high-level motivation constructs are used to explain behavior: The *need for competence* (Ryan & Deci, 2020). The need for competence is based on the concept of competence proposed by White (1959). According to White (1959), competence reflects the organism’s capacity to effectively interact with its environment. Extending this idea, Deci and Ryan (1985) proposed that humans have a basic psychological need for competence, which is defined as people’s motivation to experience feelings of mastery and success. We use the need for competence as an example simply because it is one of the most accepted high-level motivation constructs in the field (19,800 hits in Google Scholar in September 2023), and recent theoretical progress has made it possible to instantiate our point (Murayama, 2022).

Humans and animals often exhibit behavior that seems to be aimed at mastering the environment. Research showed that monkeys are engaged with solving puzzles (Harlow, 1949). Humans have the capacity to engage in learning for a prolonged period of time (Hidi & Renninger, 2019). Humans and animals also have a tendency to explore the environment without clear rewards (Berlyne, 1966) and a tendency to seek positive feedback (Elliot & Moller, 2003). None of these behaviors can be explained by basic motivation such as a hunger drive, and are considered to be caused by the need for competence (Deci & Ryan, 1985). Using

KOU MURAYAMA is a Humboldt professor in the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen. His research focuses on a number of overlapping questions about how motivation, especially the so-called “intrinsic motivation” or “curiosity,” works in human functioning. With his broad and interdisciplinary background in basic and applied (especially educational) sciences as well as expertise in statistics, his research program features a “multimethod approach,” combining a number of different perspectives and methodologies, to gain a comprehensive understanding of motivation.

HAYLEY K. JACH is a postdoctoral research fellow at the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen. She is an interdisciplinary personality psychologist who employs perspectives from cognitive science, educational science, and motivation science to understand individual differences in emotional and cognitive processes supporting goal commitment, exploration, and information seeking.

this construct, we can explain the behavior in the following way: “We have a tendency to explore because we have a basic motivation to master the environment,” and, “People can sustain their commitment to an activity because people have a fundamental motivation to seek mastery.” The satisfaction of this need for competence is theorized to enhance one’s intrinsic motivation (i.e., motivation to work on a task without relying on explicit incentives), and to lead to many positive long-term outcomes such as higher well-being (Ryan & Deci, 2017, 2020).

3. Fundamental challenges for high-level motivation constructs

3.1. The black-box problem

As the definition of motivation as an energizer and director of behavior attests, researchers typically regard motivation as the initial cause (i.e., origin) of a certain set of behaviors (Fig. 1A). Of course, researchers have assumed that motivation constructs are influenced by many external factors, such as environmental changes, learning, socialization, and development, and that they have genetic origins as well as neural bases (McClelland, 1987). In early research, motivation was conceptualized as an “intervening variable” (Hull, 1943), meaning that it has both external antecedents (e.g., deprivation of food) and outcomes (e.g., increased response). However, many motivation constructs are regarded as the origin of behavior in the sense that they are the internal variable which is supposed to generate the willpower to initiate the action in the first place.

This property of motivation constructs is particularly useful for researchers to understand higher-order behavior. As noted in the example above, we can explain exploratory behavior by proposing that we have a basic motivation to master the environment. Crucially, however, the high-level motivation construct does not truly explain *what* it is or *how* this behavioral tendency is generated (for historical arguments, see Bindra, 1959; Koch, 1956). Instead, motivation is like a black box, where the process that generates the behavior is unknown. By supposing motivation to explain behavior, we implicitly fall prey to the so-called “motivational homunculus” problem (see also Gladwin, Figner, Crone, & Wiers, 2011). That is, to explain how motivated behavior is generated, we posit a construct that works as a “generator” of motivated behavior, but this logic may suffer from the issue of infinite regress (Kenney, 1971). The real danger here is that, because the constructs seemingly explain the set of higher-order behaviors well, we may think that all questions are answered, and soon stop investigating what these constructs are and how they work.

Historically speaking, there was a time when researchers tried to eliminate the black-box property of motivation constructs by proposing physiological or biological causes. For example, Hull’s concept of drive such as hunger drive or thirst drive was directly linked to the physiological deficits of food or water (Hull, 1943). There are also contemporary theories that connect some motivation constructs with simple physiological or biological factors (e.g., testosterone and power motivation; Schultheiss, Campbell, & McClelland, 1999). However, recent studies suggest that such simple one-to-one correspondence between a physiological factor and motivation is not plausible to explain motivation constructs for higher-order complex behavior (e.g., Kim, 2013; Murayama, Izuma, Aoki, & Matsumoto, 2017; Steinman, Duque-Wilckens, & Trainor, 2019). We believe that high-level motivation

constructs partly gained popularity because they paralleled motivations for basic behavior. With the analogy of motivation for food, the statement, “We have a tendency to do X because we have a fundamental motivation for it,” sounds intuitive and convincing. However, this parallelism is misleading, because it provides a false impression that high-level motivation has clear physiological or biological causes.

Note that our argument differs from the issue of circular explanation that is often discussed in the classical literature (e.g., Bindra, 1959; Seward, 1939; i.e., motivation constructs explain a particular type of behavior by arguing that people have a motivation for that behavior). Motivation constructs do have great utility in that they can make generalizable predictions (Berridge, 2004) – for example, by supposing humans have a need for competence, one can predict that humans perform a range of epistemic behaviors, even if these behaviors were not part of the original observation. The constructs also make our explanation parsimonious – now we can conceptualize these behaviors as manifestations of the single motivational construct of the need to belong. By assuming general properties for motivation constructs, we can even make novel predictions for behavior (Baumeister & Leary, 1995). Therefore, circularity is not an issue in our opinion. Our argument is, rather, that motivation is considered a useful explanatory variable (i.e., *explanans*), but it does not have explanatory variables itself (i.e., *explanandum*).

3.2. Lack of consensus on the definition of high-level motivation constructs

The problem of the motivational black box gives rise to another critical issue: Challenges in defining high-level motivation constructs. Because motivation constructs are created to explain a certain set of behaviors without specifying their internal properties, they can be defined only in terms of the behaviors explained by the constructs (see Fig. 1A). As a result, there is always room for ambiguity when one tries to define them based on their internal properties. In other words, high-level motivation constructs have an inherent challenge for precise definition because of their black-box property.

This issue has manifested in various forms in the literature of motivation. In early years, researchers tried to create a comprehensive list of human needs (McDougall, 1909; Murray, 1938), but there was always a question of how we could be certain that two similar needs were distinct and not the same, or whether certain needs were or were not fundamental (Pittman & Zeigler, 2007). In recent years, this issue has often been discussed in the context of jingle-jangle fallacies of motivation constructs (e.g., Bong, 1996; Pekrun, 2023; Pekrun & Marsh, 2022), where the same construct label is used to denote different constructs or different construct labels are used for the same construct (Kelley, 1927). Taking the example of need for competence again, the construct is considered to be a source of intrinsic motivation (Deci & Ryan, 1985; i.e., if this need is satisfied, intrinsic motivation increases). Does this mean they are separate constructs? Or is need for competence a constituent part of intrinsic motivation? There are also other constructs related to need for competence. For example, self-efficacy is a belief in one’s capacity to competently control the environment (Bandura, 1997) and is clearly related to need for competence. Perceived control (Skinner, 1996), self-esteem (Baumeister, 1993), and self-concept (Marsh & Shavelson, 1985) are also similar constructs clearly connected to need for competence. Are they different constructs and, if

not, what are the relationships? Great effort has been devoted to resolving these issues, and the work helped researchers to deepen our thoughts about these motivation constructs. However, given the black-box property, it is virtually impossible to make a conclusive judgment on the difference between constructs, other than arguing “the originator of the theory said so” (see also Murayama, FitzGibbon, & Sakaki, 2019).

This issue of defining high-level motivation constructs is especially problematic when researchers want to make causal inferences. For example, in survey studies, researchers use a variety of designs (e.g., a longitudinal study) to estimate the causal effect of a motivation construct on outcome variables. But does that mean that motivation has a causal effect? To make a causal inference, we estimate the effect when motivation is (hypothetically) “intervened on,” holding other factors constant. However, given the black-box property of high-level motivation constructs, researchers often have difficulty in judging whether or not some potential controlling variables are the inherent property of the motivation constructs (see Eronen, 2020; Hernán & VanderWeele, 2011). For example, think of a situation in which researchers want to learn the causal effect of (satisfaction of) need for competence on math exam performance using longitudinal survey data. One may want to treat self-esteem as a controlling variable, but one could also argue that self-esteem is a constituent part of need for competence. In such cases, controlling for self-esteem does not make sense. But such a decision is often difficult because the constructs are underidentified. Generally speaking, when the target construct is not unambiguously defined, we can never make a solid causal inference from empirical data (Rohrer & Murayama, 2023).

Some may argue that the issue could be addressed empirically. For example, researchers often use the strategy of testing “incremental validity,” in which motivation construct A has predictive power over an outcome variable above and beyond a similar motivation construct B (Smith, Fischer, & Fister, 2003). Some other researchers use factor analysis and show that items representing motivation A and B form distinctive factors (Byrne, 2001; based on certain statistical criteria). Positive results from these analyses often lead researchers to conclude that the two motivation constructs “are overlapping but distinct.” However, the analysis does not directly answer the question of definitions, because the evidence simply shows that the two *measurements* are assessing something different, and is mute to exactly how the two *theoretical constructs* are differently defined. In fact, after such a conclusion, it is often not entirely clear what it really means that two motivational constructs are overlapping but still distinct.

4. Specifying mental computational processes underlying motivation: A potential solution

4.1. High-level motivation as a psychological construction

We offer an alternative perspective on high-level motivation constructs. Specifically, we propose that we should not view high-level motivation constructs as the original causal determinant of behavior. Rather, we argue that such high-level motivation constructs are an emergent property of underlying mental computational processes (Fig. 1B). To make sense of these emergent properties, humans construe the construct of motivation. In other words, high-level motivation constructs are a consequence of psychological construction.

We define mental computational processes as concrete internal mechanisms which produce behavior (see also Marr, 1982). Consider a robot that behaves like humans. The robot employs

particular computational mechanisms to process external input (i.e., sensory input) and decide on actions (i.e., output). Oftentimes some stored information in the robot (memory) plays a critical role for this computation. We call all of these mechanisms mental computational processes. Of course, humans are not robots. Although these mental computational processes determine people’s behavior, humans also have subjective experiences such as positive and negative feelings, and these subjective experiences should be influenced by mental computational processes (LeDoux, 2014).² Figure 1B provides a schematic picture.

Importantly, humans are capable of recognizing regularities in and creating mental categories from their own behaviors and subjective experiences (Reeder, 2009). “Motivation” may be a convenient term to explain these categories. Many studies suggest that we are indeed naturally inclined to infer motivations or intentions from a variety of observations (e.g., Baillargeon et al., 2015). As a result, if we have a tendency to be affiliated with certain social groups, for example, people may be naturally convinced that we have an affiliative or social motivation. However, such inference does not necessarily mean that social motivation is itself represented in our mental computational processes – instead, social motivation could be a consequence of interpreting and categorizing the regularities that exist in behavioral patterns and subjective experiences. As such, motivation is the subjective interpretation of or label for an emergent property arising from underlying mental computational processes.

From this alternative perspective, the key solution to the black-box problem is simple: To unpack the black box. Specifically, researchers on motivation should take further steps to unravel the mental computational processes underlying high-level motivation constructs. In the following section, we demonstrate how this principle can be applied to the motivation construct need for competence.

4.2. A case for reward-learning models of information seeking

We propose that the black box of need for competence can begin to be unpacked with reward-learning models of information-seeking behavior (e.g., FitzGibbon, Lau, & Murayama, 2020; Gottlieb, Oudeyer, Lopes, & Baranes, 2013; Gruber & Ranganath, 2019; Marvin & Shohamy, 2016). In our view, information seeking is an excellent lens from which we can unpack some types of motivation, including need for competence, because the act of seeking information describes effort taken to acquire knowledge, a fundamental process to master the environment. We are not claiming that the reward-learning model is the best model to instantiate need for competence with mental computational processes; there are several alternatives (e.g., Patankar et al., 2023) and there are many different versions of reward-learning models (e.g., Oudeyer & Kaplan, 2007). Our purpose is not to compare these models. We use this model simply for demonstration purposes as it provides a useful example of how our perspective can explain certain types of motivation constructs.

Although there are many different versions of reward-learning models to explain human information-seeking behavior, a common assumption is that information is an intrinsic reward. Murayama (2022; see also Murayama et al., 2019) summarized and expanded on these common aspects of reward-learning models in the “reward-learning framework of knowledge acquisition” (Fig. 2). According to the framework, when an agent identifies some uncertainty in its knowledge (often called “knowledge gap”), the agent computes the expected rewarding value of

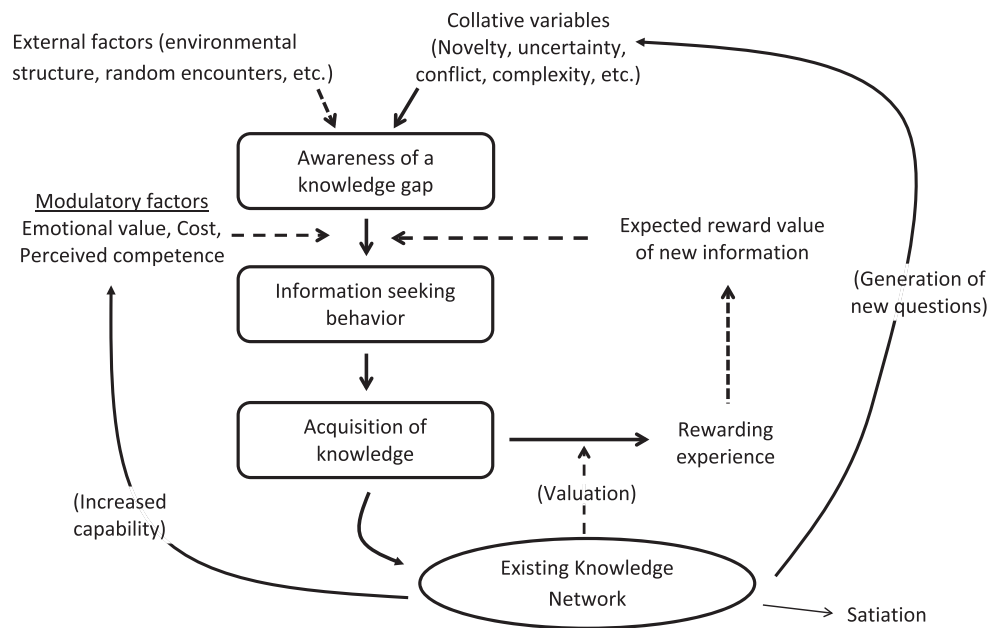


Figure 2. Reward-learning framework of knowledge acquisition (adapted from Murayama, 2022).

upcoming information, and if it is deemed valuable, the agent initiates information-seeking behavior. When the agent successfully acquires the information, the agent experiences a positive rewarding feeling, which in turn strengthens the value of the same sort of information. This is a well-known reinforcement principle (Hull, 1943) but the critical point is that the framework assumes that information itself can have rewarding value. There are different algorithms proposed to accurately quantify the rewarding value of information (e.g., uncertainty; Bennett, Bode, Brydevall, Warren, & Murawski, 2016; van Lieshout et al., 2018). Furthermore, Murayama (2022) argued that acquired information is consolidated into the existing knowledge base, and this expanded knowledge could prompt the agent to become more aware of further knowledge gaps (i.e., “the more we know about a topic, the more likely we realize that there are things that we do not know”). As a result, this system creates a positive feedback loop, sustaining long-lasting information-seeking behavior.

The framework aims to provide a rough summary of current information-seeking models. In these models, we can see several mental computational processes included, such as assessment of knowledge gap, computation of expected reward value of information, selection of information, integration of the information into existing knowledge, and so on. These computations do not normally operate consciously, but rather implicitly (Murayama et al., 2019). Some of these information-seeking models are conceptual (e.g., Gruber & Ranganath, 2019; Jach, DeYoung, & Smillie, 2022). Some other researchers propose computational models to accurately describe people’s information-seeking behavior (for a review, see Baldassarre & Mirolli, 2013) and some other models are even implemented in robots with the aim to replicate people’s (especially infants’ or young children’s) information-seeking behavior (e.g., Baranes & Oudeyer, 2013). This means that, by implementing these mental computational processes, we can create an agent that actively seeks information to expand its knowledge.

Critically, the agent which implements the reward-learning models actively and continuously searches for information that it does not know to expand its knowledge, *as if* it has the

motivations for mastery and competence, despite there being no need for competence featured in the mental computational processes (i.e., there are no boxes or parameters directly representing need for competence in the model). Need for competence thus appears as an emergent property of reward-learning models of information-seeking behavior. Several other motivational concepts can be explained in a similar way. For example, the motivational concept of interest often refers to people’s enduring tendency to engage in particular learning content over time (Hidi & Renninger, 2019; Renninger & Hidi, 2016). The agent can realize this enduring information-seeking behavior by incorporating their knowledge base and resultant positive feedback loop into the system (Murayama, 2022; Murayama et al., 2019). The agent also appears “intrinsically motivated” in the sense that it actively searches its environment without extrinsic incentives.

Of course, there are many motivational constructs that cannot be explained by the presented reward-learning framework of knowledge acquisition (e.g., need to belong). Additionally, although reward-learning models are dominant in the fields of cognitive science and neuroscience, this is not the only way to specify mental computational processing of motivated behavior. The example is simply intended to demonstrate that high-level motivation construct can be a consequence of the subjective construction from behavioral regularities.

We used need for competence and the reward-learning framework of knowledge acquisition as an illustrative example, but there are other potential examples which demonstrate that higher-level motivation constructs can be an emergent property of mental computational processes. For example, Shultz and Lepper (1996) proposed a connectionist model to explain various experimental findings of cognitive dissonance (e.g., Festinger & Carlsmith, 1959). These findings are often explained by positing that humans have a fundamental motive to maintain cognitive consistency or reduce cognitive dissonance (Festinger, 1957; Heider, 1958). However, the model explained the experimental findings without explicitly supposing such a motivation. In addition, O’Reilly (2020) built a multi-layered connectionist

model (inspired by neuroscientific findings) and argued that the model could explain motivated behavior (i.e., the dynamic nature of goal-directed behavior) without explicitly incorporating motivation into the model.

4.3. Strengths of considering mental computational processes

By specifying the mental computational processes underlying higher-order motivated behavior, high-level motivation constructs are no longer black boxes. Instead, they clearly explain *what* the motivation construct is and *how* this behavioral tendency is generated. Importantly, the proposed perspective explicitly refutes the idea that high-level motivation constructs *themselves* cause wide-ranging higher-order human behavior (see Fig. 1A). Rather, our behavior is governed by the mental computational processes, which form a collective dynamic system of interacting elements. Different types of higher-order behavior (e.g., exploring the environment, sustaining epistemic engagement, seeking competence-relevant information) are the consequences of the integration or parts of this collective system, not the consequence of a unitary construct of motivation.

The proposed perspective also neatly sidesteps the fundamental challenge of defining high-level motivation constructs. This is because it is the mental computational processes, not the motivation constructs themselves, that are necessary to understand human behavior. For example, as we presented in the previous section, a description of the mental computational processes that explain how people decide to seek information over time is sufficient to explain the behaviors related to need for competence, intrinsic motivation, and interest. There is no further need to discuss which components of this process represent need for competence, intrinsic motivation, or interest, because we already explained the mechanism and behavior (see also Kidd & Hayden, 2015; Murayama et al., 2019). The proposed perspective indicates that the priority should be given to understanding the underlying computational mechanisms, not discussing the boundaries of (inherently ill-defined) constructs.

An important benefit of the proposed perspective is that it provides a different way of theorizing about and examining motivation: Describing *how* a high-level motivation construct is *realized* by mental computational processes (for a discussion on the merit of this approach in psychology, see van Rooij & Baggio, 2021). Traditionally, researchers start by positing a specific high-level motivation construct and develop a theory by specifying the factors related to the construct (e.g., personality traits, well-being). As a consequence, empirical research has been mainly interested in examining external antecedents (e.g., family environment) and outcomes (e.g., well-being) of motivation (establishing the so-called “nomological network”; Cronbach & Meehl, 1955). As detailed later, we see merit in this approach. However, this is just one way of understanding motivated behavior. An alternative process-focused approach can create new sets of research questions to further advance our understanding of motivation.

For example, we showed that the reward-learning framework of knowledge acquisition (Murayama, 2022) can also provide a theory of how need for competence is realized through reward-learning mechanisms. Once such a theory is developed, researchers can empirically test its validity by examining various parts of the mental processes (e.g., “Does a knowledge gap really facilitate information-seeking behavior?”). The theory can also help us evaluate the antecedents and outcomes identified in the previous empirical literature (e.g., which component of the mental

computational processes can be altered by family environment?). Furthermore, the model prompts researchers to critically analyze the assumptions underlying the model. For example, the reward-learning framework assumes that information works as rewards, but the model does not specify the type of information that is perceived as rewarding. Then researchers can further theorize and empirically test the nature of information that people actively seek, further pinning down the origin of motivated behavior. In fact, this is currently a hot topic of the field (Fitzgibbon & Murayama, 2022; Gottlieb & Oudeyer, 2018; Kidd & Hayden, 2015).

4.4. Reinterpreting past empirical findings

It is important to restate that we do not disregard voluminous empirical survey and experimental studies of motivation in the past century. Much research has found that motivation, as assessed by survey questions or manipulated in experiments, is related to human behavioral and psychological outcomes. These results clearly show the usefulness of motivational constructs in predicting behavior. In fact, “motivation constructs” have been popular in psychology because they can predict various important outcomes, such as well-being, achievement scores, career choice, and so on (e.g., Robbins et al., 2004). These empirical studies can also inform researchers of potential intervention programs (e.g., Lazowski & Hulleman, 2016). However, our argument is that we should not interpret these empirical findings as evidence that high-level motivation constructs directly cause behavior.

For example, in many studies of high-level motivation constructs, researchers assess certain types of motivation using established survey questions. These survey questions often assess certain aspects of mental computational processes (e.g., the positive rewarding feeling following knowledge acquisition) or overall behavior (e.g., the frequency of active information-seeking behavior) related to the motivational construct. When the aggregated scores are related to outcomes and potential confounders are well controlled (Fig. 3, top “Statistical results from data”), we tend to argue that “motivation,” assessed by the survey questions, caused the outcomes (Fig. 3, left, “Motivation as determinant”). This way of thinking is common, and is often strengthened by the recent proliferation of latent variable modeling (Jöreskog, 1969), which visually shows path diagrams where boxes representing motivation predict or are predicted by other variables.

However, the association between aggregated scores and the outcome could happen in a different causal scenario. Specifically, these survey items which reflect parts of mental computational processes (which interact with each other) may directly influence other variables without mediating motivation constructs (Fig. 3, right, “Motivation as psychological construction”). This way of viewing measurement has gained increasing attention in the methodological literature (Donnellan, Usami, & Murayama 2023; e.g., McClure, Jacobucci, & Ammerman, 2021; VanderWeele, 2022). For example, Donnellan et al. (2023) argued that such interpretations are more realistic than latent variable modeling when items represent a wide range of behavior and psychological processes, and proposed a new mixed-effects model to analyze survey data without relying on latent variables (see also Rhemtulla, van Bork, & Borsboom, 2020). From this perspective, aggregated survey scores capture the extent to which the mental computational processes efficiently work as a whole for a particular individual or a situation (see van der Maas et al., 2006, for a similar argument in case of intelligence). Such efficiency scores are parsimonious and useful for prediction,

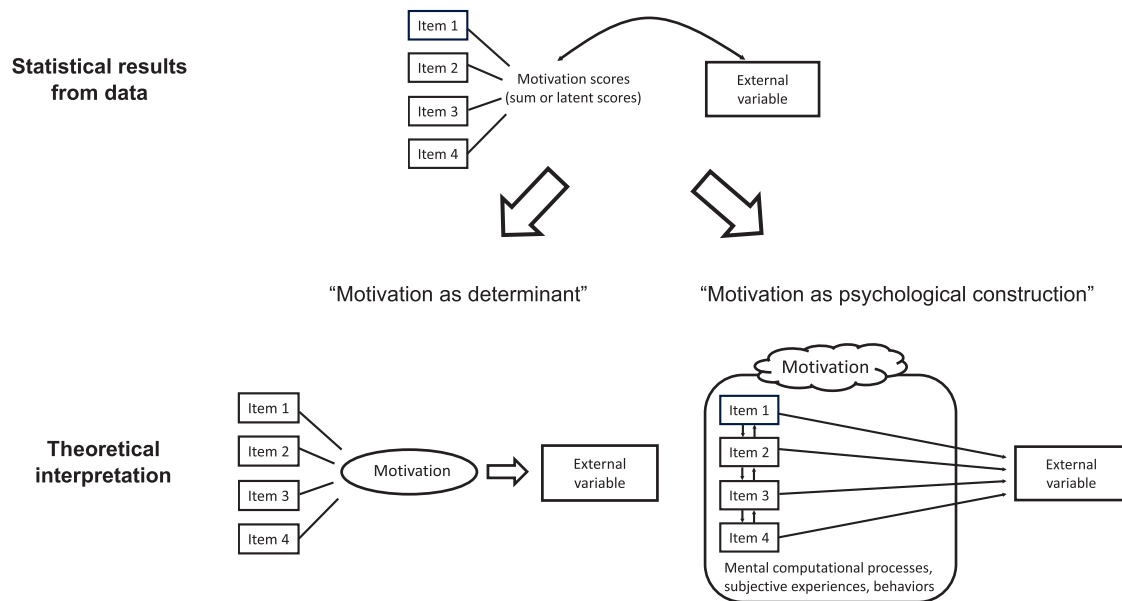


Figure 3. Different interpretations of statistical results according to different perspectives.

but we do not need to assume that these scores capture the motivation construct itself.

We believe this reframing of empirical studies on motivation is particularly important because researchers often face situations in which empirical findings disagree depending on how we measure or manipulate motivation (e.g., Hulleman, Schrager, Bodmann, & Harackiewicz, 2010; Utman, 1997). Recent years have also seen many studies indicating that the same constructs assessed by survey questions and experimental tasks have different correlates (Dang, King, & Inzlicht, 2020; Eisenberg et al., 2019), and this has also been a recurring issue in motivation constructs (Deci, Koestner, & Ryan, 1999; McClelland, Koestner, & Weinberger, 1989; Schultheiss, Yankova, Dirlikov, & Schad, 2009). In such situations, we are often encouraged to accurately define the construct or establish a valid measurement. It is indeed true that we need precise definitions of constructs and valid measurement, but at the same time, the argument rests on an implicit assumption that there is a single construct of motivation that determines behavior and thus the inconsistent results are problematic. From the perspective of psychological construction, on the contrary, such divergent results are natural when different assessments or manipulations tap different aspects of the mental computational processes. Having a precise process model in mind, researchers can then make more fine-grained predictions about how different types of assessments or manipulations result in different outcomes. Such a process-oriented perspective can also help researchers develop targeted interventions that aim to change specific outcomes.

5. Answering questions

To further elaborate and avoid misunderstanding, here we answer some follow-up questions that readers might have about our perspective.

5.1. What is the distinction between basic motivation constructs and high-level motivation constructs?

The high-level motivation constructs evaluated in this article explain a wide-range of higher-order behavior, whereas basic

motivation constructs have a relatively narrow set of behaviors as explanandum (e.g., the motivation for sex to explain people's desire to copulate). However, although we made a distinction in this article for the purpose of simplicity, the distinction is continuous rather than dichotomous. We do not argue that motivation for basic behavior does not suffer from the issue of black box at all – this is a matter of degree. Our point is that the problem is exacerbated for high-level motivation constructs.

5.2. Are mental computational processes free from motivation constructs?

No, they are not. As mental computational processes produce behaviors, we must logically assume something which initiates behavior either explicitly or implicitly. In the reward-learning framework of knowledge acquisition, there is an implicit assumption that people choose to seek information that has a high reward value. Therefore, one could argue that the framework posits a “motivation to seek rewarding information,” perhaps just before the box of information-seeking behavior in Figure 2. Unless we understand the basic biological/physical mechanisms to produce behavior, we can never eliminate motivation-like constructs.

But our argument is not to eliminate motivation constructs from explanation. Our proposal is that theorizing mental computational processes (i.e., analyzing motivation from a different level) can provide new insights into what researchers have called “motivation” in the literature, enabling a deeper understanding of our motivated behavior. For example, both exploration behavior and long-term intellectual engagement are thought to be caused by need for competence, but there is no explanation for how these outcomes are related. By specifying the mental computational processes, we can propose that exploration behavior is caused by the rewarding value of information, whereas long-term engagement is a consequence of a positive feedback loop created by expanding existing knowledge from new information (Fig. 2). Both behaviors are produced by the same mental computational system (so it makes sense that they are explained by the same construct) but the way they are produced within the system is different.

We can go even further down the line to unpack motivation-like constructs in the specified mental computational processes. As noted above, the reward-learning framework implicitly assumes a motivation to seek rewarding information. But what are the basic building blocks of rewarding value for information? Is that novelty (Poli, Meyer, Mars, & Hunnius, 2022), entropy (Loewenstein, 1994), or Kullback–Leibler divergence (Ningombam, Yoo, Kim, Song, & Yi, 2022)? As we do not have direct access to such information in reality, how can such a metric be computed in the real world? (Gottlieb & Oudeyer, 2018). Note that as we go down the level, motivation-like constructs become narrower and narrower (or more and more specific), and at some point we may no longer feel comfortable to call it motivation. When we orient attention to an object that suddenly comes into our visual field, for example, is this a manifestation of motivation? We can in theory suppose such a motivation that causes the orientation behavior, but many researchers would say it is not necessary or useful (Murayama, 2023b).

In a sense our proposal is consistent with the idea that the functioning of human behavior can be understood at different levels of granularity with the lowest being the biological or physical level (e.g., Dennett, 1987; Marr, 1982; Newell, 1994; in case of motivation, see Nagengast & Trautwein, 2023). Our proposed perspective indicates that high-level motivation constructs reflect higher-level explanations whereas mental computational processes represent lower-level explanations. No level of understanding should be dismissed as “wrong” (i.e., one level of explanation should not be replaced with a lower-level explanation), because they explain the behavior for different purposes, but the problem of motivation literature is that most researchers are satisfied with higher-level explanations (i.e., supposing high-level motivation constructs to explain behavior) and little effort has been made to pursue lower-level explanations.

5.3. Has your perspective truly not been discussed in the prior motivation literature?

The idea of psychological construction is not novel at a general level – it has been discussed in various forms in psychological and cognitive science as well as philosophy of science (Brick, Hood, Ekroll, & de-Wit, 2022; Churchland, 1979; Dalege et al., 2016; Danziger, 1990; MacCorquodale & Meehl, 1948; Pessoa, Medina, & Desfilis, 2022; Stich & Ravenscroft, 1994). Other psychological constructs such as emotions, personality, intelligence, and consciousness have also received similar scrutiny (Boag, 2018; Fiske, 2020; Lau, 2009; Russell, 2003; van der Maas et al., 2006). Recent theoretical developments in psychological network analysis (Borsboom & Cramer, 2013) have a similar root (for application to motivational constructs, see Sachisthal et al., 2019; Tamura et al., 2022). However, the construct of motivation can be uniquely positioned in this context because motivation is characterized as the causal determinant of behavior. As we showed earlier, this unique property poses fundamental challenges when theorizing or interpreting motivated behavior, and the psychological construction perspective has clear utility to circumvent the issue.

Nevertheless, there is some (albeit limited) work related to our psychological construction perspective in the literature of motivation. For example, Bem (1967) suggested that dissonance reduction motive (Festinger, 1957) can be interpreted as the consequence of self-perception. The idea is that people are not motivated by the dissonance reduction motive, but simply act

and infer attitudes by observing/interpreting what they did (i.e., attribution process). This idea is seemingly similar to the psychological construction perspective in that it does not assume a motivation to explain higher-order behavior. However, the difference is that Bem’s (1967) theory uses the attribution process to explain people’s actual behavior or attitude (e.g., “people changed their behavior because of attribution”). On the contrary, our proposed perspective would suggest an interplay between different mental computational processes that together produced something we would categorize as a “dissonance reduction motive.” The attribution process could be part of it, but not the whole (in fact, it is not plausible that higher-order behavior is largely explained by attribution processes). Therefore, although we acknowledge that there were similar ideas in the literature of motivation, our proposed psychological construction perspective is critically distinct in that we stress the importance of specifying mental computational processes.

That said, this line of work suggests another interesting line of work for future research. Specifically, we could examine the processes of how people construct motivation from the regularities of observable behavior and subjective feelings. Although early research on attribution provided many insights into this psychological construction process (e.g., Bem, 1967; Nisbett & Wilson, 1977), the proliferation of high-level motivation constructs in recent years seems to have suppressed this tradition of work. But by combining the work on mental computational processes and psychological construction processes, we may be able to achieve deeper insights into motivated behavior. For example, there is a possibility that such attributional process of motivation works as a reinforcer of motivated behavior itself, influencing mental computational processes (e.g., “I like to study because I think I have high achievement motivation”; see Palminteri & Lebreton, 2022). Future studies could examine such interactive processes.

5.4. Does the criticism only apply to the energization aspect of motivation and not the directional aspect?

Our discussion focused mainly on the energization aspect of motivation, but motivation is said to additionally have a directional aspect. The directional aspect of motivation channels people’s behavior in a certain way, and researchers argue that values (Eccles & Wigfield, 2002) or goals (Elliot & Fryer, 2008) play a critical role here. Does our proposed perspective have implications for the directional aspect?

As noted earlier, the distinction between energization and direction aspects in motivation is somewhat ambiguous. To clarify the distinction, Niv, Joel, and Dayan (2006) introduced a normative account of motivation, by defining the directional aspect of motivation as a modulation of the mapping between reward outcome and utility (Fig. 4). When one is in a state of hunger, the utility of food reward would be increased. When one is in a state of thirst, the utility of food may not be as high but the utility of water would be increased. This means that hunger motivation altered the utility of food. This is one of the commonly accepted definitions of motivation in the classic literature (Berridge, 2004; Dickinson & Balleine, 2002; Hull, 1943). By viewing motivation as the modulator of reward utility, we can independently define motivation as separate from the energization aspect (Niv et al., 2006) – it simply changes the mapping of utility without directly causing behavior. In this respect, therefore, our criticism does not apply to the directional aspect of motivation as exemplified by hunger or thirst.

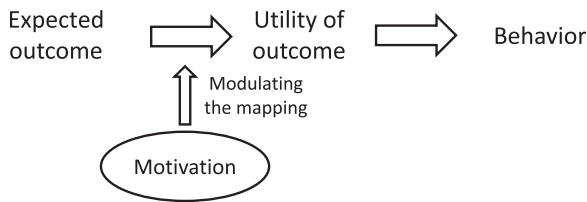


Figure 4. Direction aspect of motivation: Motivation functions as utility mapping.

Goals and values could be considered in the same manner. Like hunger or thirst, goals and values function as the modulator of utility; when one values acquiring mathematics competence, for example, they would assign high utility or experience stronger positive feelings when succeeding in a math exam than those who do not value math. When one pursues performance goals, outperforming others would give stronger positive feelings than those pursuing mastery goals. Goals and values are also not viewed as initial determinants of behavior; rather, we usually think that goals and values are set in people's mind through certain mechanisms. They are often externally provided (e.g., external goal setting; Harackiewicz & Sansone, 1991; Locke & Latham, 1990; value intervention; Rozek, Svoboda, Harackiewicz, Hulleman, & Hyde, 2017) or internally generated (e.g., self-set goals; Barron & Harackiewicz, 2001; Locke, 2001; internalization of values; Renninger & Hidi, 2016). Therefore, our criticism does not directly apply to these concepts.

However, unlike hunger or thirst, theories of such higher-order constructs such as goals and values are still unclear about the mental computational mechanism of how they are internally generated (for some exceptions, see Ballard, Palada, Griffin, & Neal, 2021; Vancouver, Wang, & Li, 2020). For example, prominent theories, such as expectancy-value theory (Eccles & Wigfield, 2020) and the hierarchical model of achievement motivation (Elliot, 1997), identified a number of factors that influence goals or values (e.g., implicit beliefs, personal characteristics, subjective perceptions, and affective states). However, these theories do not answer *how* these factors lead to the adoption of goals and values. In fact, the hierarchical model does not identify what kind of decision-making process is involved when one adopts mastery goals over performance goals. Expectancy-value theory does not specify how value is incorporated and represented into the existing knowledge structure. Like the case of needs and motives, we believe that specifying the computational mechanisms of goal/value adoption and transformation would provide a new landscape of understanding these concepts even better. In sum, although directional aspects of motivation are indeed less immune to our criticism, higher-order motivational concepts such as goals and values still have large room to benefit from the proposed perspective.³

5.5. What about the evolutionary account of motivation?

Researchers often use evolution to justify motivation constructs, for example, people are motivated to understand the environment because this is crucial for survival. It is true that sets of behavior explained by need for competence can have survival value. But this perspective only addresses the *why* question ("why do people have motivation X?"), not the *what* ("what is this motivation X?") or *how* ("how motivation X is realized") questions. In addition, an evolutionary perspective does not mean that we should keep motivation constructs a black box. In fact, the evolutionary perspective is open to the possibility that evolution shaped our

mental computational processes in a particular manner. In other words, thinking about adaptive value and ecological constraints would help researchers specify more realistic mental computational processes (Anderson & Milson, 1989; Lieder & Griffiths, 2020). Therefore, evolution is not incompatible with our proposal that we should unpack the black-box property of high-level motivation constructs – evolution is, rather, a useful tool to unpack the black-box.

6. Conclusion: Toward the future of motivation science

In this article, we provided a critical evaluation of motivation constructs that explain a wide range of higher-order behavior. Although such high-level motivation constructs seem to explain higher-order behavior quite well, they do not specify what they are or how they work, which we called the *black-box problem*. To address the black-box problem, we highlighted the utility of specifying the mental computational processes underlying motivated behavior. Importantly, according to this perspective, high-level motivation constructs can be understood as people's subjective construction of these mental processes, and should not be considered the determinant of behavior. The idea of psychological construction is rather metatheoretical; as such, it does not contradict or refute the vast number of empirical findings in the field. Nevertheless, the proposed perspective points to important avenues for future theoretical development – theories addressing how motivation is *realized* in our mental computational processes.

We do not intend to refute the utility of existing theories of motivation. Existing theories of motivation constructs profoundly shaped our academic field, orienting researchers toward important phenomena that would otherwise have been overlooked. For example, to explain human behaviors that are not driven by clear extrinsic incentives (e.g., money, food, etc.), Deci and Ryan (1985) proposed that humans have intrinsic motivation. Since the introduction of the concept of intrinsic motivation, numerous studies have examined the nature of human behavior that is not driven by extrinsic incentives, which significantly enhanced our understanding of such behavior.

At the same time, the black-box issue in the existing theories deeply constrains our theories of motivated behavior. For example, it is generally challenging to understand the relationship between different high-level motivation constructs. This is because both constructs do not specify what the construct is composed of, providing room for different ways of theorizing their relationship. Consequently, as many researchers have repeatedly indicated (Anderman, 2020; Baumeister, 2016; Murphy & Alexander, 2000), there are currently too many motivation theories and constructs, with little integration between them. Identification of mental computational processes may provide a way to understand motivated behavior more parsimoniously, because the same process can give rise to different types of motivated behavior. In fact, the reward-learning framework we presented (Fig. 2) can explain the manifestation of several motivation constructs in a single framework (e.g., need for competence, intrinsic motivation, curiosity, interest; for a more comprehensive treatment, see Murayama, 2022).

We have consistently used the need for competence as an example so that readers can follow our argument easily, and because there has been great progress in specifying computational processes underlying exploration, a behavior closely linked to need for competence (Baldassarre & Mirolli, 2013; Cogliati Dezza, Schulz, & Wu, 2022; Oudeyer & Kaplan, 2007). However, the idea applies to many

higher-level motivation constructs. Need to belong, for example, is defined as a drive to form and maintain at least a minimum quantity of lasting, positive, and significant interpersonal relationships (Baumeister & Leary, 1995). The construct is used to explain a wide range of social behaviors, such as people's bond-forming behavior, the tendency to preserve social relations, the effective/prioritized processing of social information, social satiation (i.e., reduced inclination to form new social relationships if one already has sufficient social bonds), negative emotional experiences after rejection, and so on. But the construct does not explain the processes through which these types of social behaviors unfold. For instance, the tendency to preserve social relationships may be explained by the positive feedback loop of social rewards (a positive rewarding experience because of instant social interaction) – once a person is in a group, the person may gain constant positive social feedback from peers such that the person is not willing to leave the group. Here the term “social rewards” should still be unpacked, but it is much more narrowly defined than the need to belong. Regarding the effective processing for social stimuli, recent work using deep neural networks showed that such effective/prioritized processing (e.g., face recognition) as well as the brain functional localization of social information may emerge from our category of discrimination process (Dobs, Martinez, Kell, & Kanwisher, 2022; Kanwisher, Khosla, & Dobs, 2023). This is just a simple thinking exercise (not even a hypothesis) but it opens a new avenue for theorizing or examining the need to belong.

Of course, such theorizing is not easy. This is because mental computational processes are not observable and people are typically unaware of such processes (Kihlstrom, 1987; Nisbett & Wilson, 1977). However, the field of cognitive science and other related areas have provided different ways to theorize about our mental computational processes (including reward-learning models), and researchers can take advantage of these precedents as the starting point (e.g., Dayan & Abbott, 2005; Friston & Kiebel, 2009; Lieder & Griffiths, 2020). Biological constraints often inform mental computational processes of motivation (see, e.g., Rolls, 2016). It is also important to add that the mental computational processes do not have to be described as quantitative formulas (i.e., computational modeling). Such a formulation is useful but neither necessary nor sufficient. The reward-learning framework we discussed above (Murayama, 2022), for example, did not specify the computational function of how expected reward value is calculated, but it still explains how the need for competence can emerge from such a system. There are some other attempts to conceptually describe the mechanisms underlying motivated behavior (e.g., Brehm, 1999; Richter, Gendolla, & Wright, 2016). In fact, the use of mathematical formulation can be challenging to describe the complexity of motivated behavior, especially in real-world contexts (DeYoung & Krueger, 2020). Conversely, we can also easily trick the mathematical expression to explain motivated behavior (e.g., adding a “constant” or “bonus” in the utility value for the motivated behavior one wants to explain). In such cases, the resultant model does not really explain the origin of motivated behavior, even if the model is mathematically expressed.⁴ Our point is that many roads can lead to Rome: Our goal should be to describe the processes underlying motivated behavior, regardless of how this goal is achieved.

From the proposed perspective, one important avenue for future theoretical work is to understand how intraindividual states are related to stable interindividual differences (i.e., states and traits; see also Baumeister, 2016; this is also an issue of timescale). When motivation is defined as the determinant of behavior,

motivational traits can be simply quantified and operationalized as “general strength” of motivation. Motivational states, on the contrary, can be quantified and operationalized as short within-person fluctuation of the strengths. However, mental computational processes occur at the within-person level by definition. They do not have an enduring “general strength,” and it is not obvious how stable motivational traits can be explained. To understand motivational states and traits, we need to develop a theory of mental computational processes that explicitly addresses how intraindividual processes translate into long-term development (see Dalege et al., 2016; Murayama, 2022; see also Atkinson, Bongort, & Price, 1977).

One big implication of the proposed perspective is that we should no longer see motivation as an inherent category. In fact, based on the reward-learning framework of knowledge acquisition (Fig. 2) one can argue that intrinsically motivated behavior is controlled by cognitive computational processes (e.g., calculation of expected reward value of new information) as well as affective experiences (i.e., rewarding feelings). Learning is also at the heart of this motivational process. There has been a long tradition in psychology to distinguish several inherent categories of mental functioning such as motivation, emotion, and cognition (Danziger, 1990), and these categories formed distinct research fields with distinct theories. Using these categories, there have also been attempts to discuss how they are related. For example, some emotion theories argue that emotions have a motivating functioning, subsuming motivation under the category of emotion (e.g., Arnold, 1969; Brehm, 1999; Frijda, 1986). These categories are certainly useful for academic communications, and understanding the relationship between these categories should advance our understanding of motivated behavior. However, we also suggest that these categories could constrain our thinking, potentially hindering the development of comprehensive theories to explain behavior and decision making (Murayama, 2023a). In fact, as long as we can correctly specify the mental computational processes to explain behavior, it is not that important to discuss which part of the processes is categorized as motivation, emotion, or cognition.

Halliday (1983, p. 105) states that when a psychological variable such as motivation is invented, it simply means “some phenomenon that requires explanation has been identified” (see also MacCorquodale & Meehl, 1948). We need to take this statement seriously. Motivation is not an explanation itself. It is the starting point for explanation. We hope that the proposed “unpacking the black-box” perspective motivates researchers (via, of course, mental computational processes) to explore new forms of research on human motivated behavior.

Acknowledgments. We thank Bernard Weiner and Andrew Elliot for providing comments on an earlier draft of the manuscript.

Financial support. This work was supported by the Alexander von Humboldt Foundation (Kou Murayama, the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research).

Competing interest. None.

Notes

1 The preliminary idea of the manuscript was already discussed in a short essay (Murayama, 2023b).

2 As indicated by the figure, subjective experiences can exert impact on mental computational processes. There has been a long discussion on whether this

is true or not (e.g., Sheldon, 2022; Wegner, 2004) but our argument holds regardless of the standpoint on the matter. The key point is that these effects are, if any, mediated by mental computational processes.

3 One might argue that high-level motivation constructs which we discussed, such as need for competence or need to belong, can also be conceptualized as having this directional function (e.g., people add high utility for intimate social group formation). However, for the same reason described here, this does not address the essential problem of these high-level motivation constructs.

4 Berridge (2023) discussed this point by comparing two different computational models of “wanting” (Smith & Read, 2022; Zhang, Berridge, Tindell, Smith, & Aldridge, 2009).

References

- Anderman, E. M. (2020). Achievement motivation theory: Balancing precision and utility. *Contemporary Educational Psychology*, 61, 101864. <https://doi.org/10.1016/j.cedpsych.2020.101864>
- Anderson, C., Hildreth, J. A. D., & Howland, L. (2015). Is the desire for status a fundamental human motive? A review of the empirical literature. *Psychological Bulletin*, 141, 574–601. <https://doi.org/10.1037/a0038781>
- Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96, 703–719.
- Arnold, M. B. (1969). Emotion, motivation, and the limbic system. *Annals of the New York Academy of Sciences*, 159(3), 1041–1058. <https://doi.org/10.1111/j.1749-6632.1969.tb12996.x>
- Atkinson, J. W., Bongort, K., & Price, L. H. (1977). Explorations using computer simulation to comprehend thematic apperceptive measurement of motivation. *Motivation and Emotion*, 1, 1–27. <https://doi.org/10.1007/BF00997578>
- Atkinson, J. W., & Raynor, J. O. (1978). *Personality, motivation, and achievement*. Hemisphere.
- Austin, J. T., & Vancouver, J. B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin*, 120, 338–375.
- Baillargeon, R., Scott, R. M., He, Z., Sloane, S., Setoh, P., Jin, K.-s., ... Bian, L. (2015). Psychological and sociomoral reasoning in infancy. In M. Mikulincer, P. R. Shaver, E. Borgida, & J. A. Bargh (Eds.), *APA handbook of personality and social psychology*, Vol. 1. *Attitudes and social cognition* (pp. 79–150). American Psychological Association. <https://doi.org/10.1037/14341-003>
- Baldassarre, G., & Mirolli, M. (2013). *Intrinsically motivated learning in natural and artificial systems*. Springer.
- Ballard, T., Palada, H., Griffin, M., & Neal, A. (2021). An integrated approach to testing dynamic, multilevel theory: Using computational models to connect theory, model, and data. *Organizational Research Methods*, 24(2), 251–284. <https://doi.org/10.1177/1094428119881209>
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. Freeman.
- Baranes, A., & Oudeyer, P.-Y. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1), 49–73. <https://doi.org/10.1016/j.robot.2012.05.008>
- Barron, K. E., & Harackiewicz, J. M. (2001). Achievement goals and optimal motivation: Testing multiple goal models. *Journal of Personality and Social Psychology*, 80, 706–722.
- Baumeister, R. F. (1993). *Self-esteem: The puzzle of low self-regard*. Plenum Press.
- Baumeister, R. F. (2016). Toward a general theory of motivation: Problems, challenges, opportunities, and the big picture. *Motivation and Emotion*, 40(1), 1–10. <https://doi.org/10.1007/s11031-015-9521-y>
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117, 497–529.
- Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, 74(3), 183–200. <https://doi.org/10.1037/h0024835>
- Bennett, D., Bode, S., Brydevall, M., Warren, H., & Murawski, C. (2016). Intrinsic valuation of information in decision making under uncertainty. *PLoS Computational Biology*, 12(7), e1005020. <https://doi.org/10.1371/journal.pcbi.1005020>
- Berlyne, D. E. (1966). Curiosity and exploration. *Science*, 153(3731), 25–33. <https://doi.org/10.1126/science.153.3731.25>
- Berridge, K. C. (2004). Motivation concepts in behavioral neuroscience. *Physiology & Behavior*, 81(2), 179–209. <https://doi.org/10.1016/j.physbeh.2004.02.004>
- Berridge, K. C. (2023). Separating desire from prediction of outcome value. *Trends in Cognitive Sciences*, 27(10), 932–946. <https://doi.org/10.1016/j.tics.2023.07.007>
- Bindra, D. (1959). *Motivation: A systematic reinterpretation* (1st ed./1st printing). Ronald Press.
- Boag, S. (2018). Personality dynamics, motivation, and the logic of explanation. *Review of General Psychology*, 22(4), 427–436. <https://doi.org/10.1037/gpr0000150>
- Bong, M. (1996). Problems in academic motivation research and advantages and disadvantages of their solutions. *Contemporary Educational Psychology*, 21, 149–165. <https://doi.org/10.1006/ceps.1996.0013>
- Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, 9(1), 91–121. <https://doi.org/10.1146/annurev-clinpsy-050212-185608>
- Brehm, J. W. (1999). The intensity of emotion. *Personality and Social Psychology Review*, 3(1), 2–22. https://doi.org/10.1207/s15327957pspr0301_1
- Brick, C., Hood, B., Ekroll, V., & de-Wit, L. (2022). Illusory essences: A bias holding back theorizing in psychological science. *Perspectives on Psychological Science*, 17(2), 491–506. <https://doi.org/10.1177/1745691621991838>
- Byrne, B. M. (2001). Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validity of a measuring instrument. *International Journal of Testing*, 1(1), 55–86. https://doi.org/10.1207/S15327574IJT0101_4
- Churchland, P. M. (1979). *Scientific realism and the plasticity of mind*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511625435>
- Cogliati Dezza, I., Schulz, E., & Wu, C. M. (Eds.). (2022). *The drive for knowledge: The science of human information seeking*. Cambridge University Press. <https://doi.org/10.1017/9781009026949>
- Corbit, L. H., & Balleine, B. W. (2005). Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of Pavlovian-instrumental transfer. *Journal of Neuroscience*, 25(4), 962–970. <https://doi.org/10.1523/JNEUROSCI.4507-04.2005>
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Dale, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. J. (2016). Toward a formalized account of attitudes: The causal attitude network (CAN) model. *Psychological Review*, 123(1), 2–22. <https://doi.org/10.1037/a0039802>
- Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences*, 24(4), 267–269. <https://doi.org/10.1016/j.tics.2020.01.007>
- Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*. Cambridge University Press.
- Dayan, P., & Abbott, L. F. (2005). *Theoretical neuroscience: Computational and mathematical modeling of neural systems* (Revised ed.). MIT Press.
- Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6), 627.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Plenum.
- Dennett, D. C. (1987). *The intentional stance* (pp. xi, 388). MIT Press.
- Descartes, R. (1955). *The philosophical works of Descartes* (2 vols.) (pp. ix, 832). Dover.
- DeYoung, C. G., & Krueger, R. F. (2020). To wish impossible things: On the ontological status of latent variables and the prospects for theory in psychology. *Psychological Inquiry*, 31(4), 289–296. <https://doi.org/10.1080/1047840X.2020.1853462>
- Dickinson, A., & Balleine, B. (2002). The role of learning in the operation of motivational systems. In H. Pashler & R. Gallistel (Eds.), *Steven's handbook of experimental psychology: Learning, motivation, and emotion* (3rd ed., pp. 497–533). John Wiley & Sons, Inc. <https://doi.org/10.1002/0471214426.pas0312>
- Dobs, K., Martinez, J., Kell, A. J., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. *Science Advances*, 8(11), eabl8913.
- Donnellan, E., Usami, S., & Murayama, K. (2023). Random item slope regression: An alternative measurement model that accounts for both similarities and differences in association with individual items. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000587>
- Dweck, C. S. (2017). From needs to goals and representations: Foundations for a unified theory of motivation, personality, and development. *Psychological Review*, 124(6), 689–719. <https://doi.org/10.1037/rev0000082>
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53(1), 109–132. <https://doi.org/10.1146/annurev.psych.53.100901.135153>
- Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, 61, 101859. <https://doi.org/10.1016/j.cedpsych.2020.101859>
- Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications*, 10(1), Article 1. <https://doi.org/10.1038/s41467-019-10301-1>
- Elliot, A. J. (1997). Integrating “classic” and “contemporary” approaches to achievement motivation: A hierarchical model of approach and avoidance achievement motivation. In P. R. Pintrich & M. L. Maehr (Eds.), *Advances in motivation and achievement* (Vol. 10, pp. 143–179). JAI Press.
- Elliot, A. J. (2023). Energization and direction are both essential parts of motivation. In M. Bong, J. Reeve, & S. Kim (Eds.), *Motivation science: Controversies and insights* (pp. 10–14). Oxford University Press. <https://doi.org/10.1093/oso/9780197662359.003.0002>

- Elliot, A. J., & Fryer, J. W. (2008). The goal construct in psychology. In J. Y. Shah & W. L. Gardner (Eds.), *Handbook of motivation science* (pp. 235–250). Guilford Press.
- Elliot, A. J., & Moller, A. C. (2003). Performance-approach goals: Good or bad forms of regulation? *International Journal of Educational Research*, 39, 339–356.
- Eronen, M. I. (2020). Causal discovery and the problem of psychological interventions. *New Ideas in Psychology*, 59, 100785. <https://doi.org/10.1016/j.newideapsych.2020.100785>
- Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.
- Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology*, 58(2), 203–210. <https://doi.org/10.1037/h0041593>
- Fishbach, A., & Ferguson, M. J. (2007). The goal construct in social psychology. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (2nd ed., pp. 490–515). Guilford Press.
- Fiske, A. P. (2020). The lexical fallacy in emotion research: Mistaking vernacular words for psychological entities. *Psychological Review*, 127(1), 95–113. <https://doi.org/10.1037/rev0000174>
- FitzGibbon, L., Lau, J. K. L., & Murayama, K. (2020). The seductive lure of curiosity: Information as a motivationally salient reward. *Current Opinion in Behavioral Sciences*, 35, 21–27. <https://doi.org/10.1016/j.cobeha.2020.05.014>
- Fitzgibbon, L., & Murayama, K. (2022). Counterfactual curiosity: Motivated thinking about what might have been. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1866), 20210340. <https://doi.org/10.1098/rstb.2021.0340>
- Frijda, N. H. (1986). *The emotions* (pp. xii, 544). Editions de la Maison des Sciences de l'Homme.
- Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521), 1211–1221. <https://doi.org/10.1098/rstb.2008.0300>
- Gladwin, T. E., Figner, B., Crone, E. A., & Wiers, R. W. (2011). Addiction, adolescence, and the integration of control and motivation. *Developmental Cognitive Neuroscience*, 1(4), 364–376. <https://doi.org/10.1016/j.dcn.2011.06.008>
- Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, 19(12), Article 12. <https://doi.org/10.1038/s41583-018-0078-0>
- Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information seeking, curiosity and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences*, 17(11), 585–593. <https://doi.org/10.1016/j.tics.2013.09.001>
- Gruber, M. J., & Ranganath, C. (2019). How curiosity enhances hippocampus-dependent memory: The prediction, appraisal, curiosity, and exploration (PACE) framework. *Trends in Cognitive Sciences*, 23(12), 1014–1025. <https://doi.org/10.1016/j.tics.2019.10.003>
- Halliday, T. H. (1983). Motivation. In T. H. Halliday & P. J. B. Slater (Eds.), *Animal behaviour: Causes and effects* (Vol. 1, pp. 100–133). Blackwell Scientific.
- Harackiewicz, J. M., & Sansone, C. (1991). Goals and intrinsic motivation: You can get there from here. In M. L. Maehr & P. R. Pintrich (Eds.), *Advances in motivation and achievement* (Vol. 7, pp. 21–49). JAI Press.
- Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, 56(1), 51–65. <https://doi.org/10.1037/h0062474>
- Heider, F. (1958). *The psychology of interpersonal relations* (pp. ix, 326). John Wiley. <https://doi.org/10.1037/10628-000>
- Hernán, M. A., & VanderWeele, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology (Cambridge, MA)*, 22(3), 368–377. <https://doi.org/10.1097/EDE.0b013e3182109296>
- Hidi, S. E., & Renninger, K. A. (2019). Interest development and its relation to curiosity: Needed neuroscientific research. *Educational Psychology Review*, 31(4), 833–852. <https://doi.org/10.1007/s10648-019-09491-3>
- Hinde, R. A. (1960). Energy models of motivation. *Symposia of the Society for Experimental Biology*, 14, 199–213.
- Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory*. Appleton-Century.
- Hulleman, C., Schrager, S., Bodmann, S., & Harackiewicz, J. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin*, 136, 422–449. <https://doi.org/10.1037/a0018947>
- Jach, H. K., DeYoung, C. G., & Smillie, L. D. (2022). Why do people seek information? The role of personality traits and situation perception. *Journal of Experimental Psychology: General*, 151, 934–959. <https://doi.org/10.1037/xge0001109>
- James, W. (1890). *The principles of psychology* (Vol. 1). Henry Holt and Co. <https://doi.org/10.1037/10538-000>
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2), 183–202. <https://doi.org/10.1007/BF02289343>
- Kanfer, R., & Chen, G. (2016). Motivation in organizational behavior: History, advances and prospects. *Organizational Behavior and Human Decision Processes*, 136, 6–19. <https://doi.org/10.1016/j.obhdp.2016.06.002>
- Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask “why” questions of minds and brains. *Trends in Neurosciences*, 46(3), 240–254. <https://doi.org/10.1016/j.tins.2022.12.008>
- Kelley, T. L. (1927). *Interpretation of educational measurements* (p. 353). World Book.
- Kenny, A. (1971). The homunculus fallacy. In M. G. Grene & I. Prigogine (Eds.), *Interpretations of life and mind* (pp. 155–165). Humanities Press.
- Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, 88(3), 449–460. <https://doi.org/10.1016/j.neuron.2015.09.010>
- Kihlstrom, J. F. (1987). The cognitive unconscious. *Science*, 237(4821), 1445–1452. <https://doi.org/10.1126/science.3629249>
- Kim, S.-I. (2013). Neuroscientific model of motivational process. *Frontiers in Psychology*, 4, 98–98. <https://doi.org/10.3389/fpsyg.2013.00098>
- Kleinginna, P. R., & Kleinginna, A. M. (1981). A categorized list of motivation definitions, with a suggestion for a consensual definition. *Motivation and Emotion*, 5(3), 263–291. <https://doi.org/10.1007/BF00993889>
- Koch, S. (1941). The logical character of the motivation concept. I. *Psychological Review*, 48, 15–38. <https://doi.org/10.1037/h0062042>
- Koch, S. (1956). Behavior as “intrinsically” regulated: Work notes towards a pre-theory of phenomena called “motivational”. In M. R. Jones (Ed.), *Nebraska symposium on motivation* (Vol. 4, pp. 42–87). Nebraska University Press.
- Lau, H. C. (2009). Volition and the function of consciousness. In N. Murphy, G. F. R. Ellis, & T. O'Connor (Eds.), *Downward causation and the neurobiology of free will* (pp. 153–169). Springer. https://doi.org/10.1007/978-3-642-03205-9_9
- Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education: A meta-analytic review. *Review of Educational Research*, 86(2), 602–640. <https://doi.org/10.3102/0034654315617832>
- LeDoux, J. E. (2014). Coming to terms with fear. *Proceedings of the National Academy of Sciences of the United States of America*, 111(8), 2871–2878. <https://doi.org/10.1073/pnas.1400335111>
- Lewin, K. (1942). Field theory and learning. *Teachers College Record*, 43(10), 215–242. <https://doi.org/10.1177/016146814204301006>
- Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43, e1. <https://doi.org/10.1017/S0140525X1900061X>
- Locke, E. A. (2001). Self-set goals and self-efficacy as mediators of incentives and personality. In M. Erez, U. Kleinbeck, & H. Thierry (Eds.), *Work motivation in the context of a globalizing economy* (pp. 13–26). Lawrence Erlbaum Associates Publishers.
- Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance*. Prentice Hall.
- Locke, E. A., & Latham, G. P. (2004). What should we do about motivation theory? Six recommendations for the twenty-first century. *Academy of Management Review*, 29(3), 388–403. <https://doi.org/10.5465/amr.2004.13670974>
- Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116(1), 75–98. <http://doi.org/10.1037/0033-2909.116.1.75>
- MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, 55(2), 95–107. <https://doi.org/10.1037/h0056029>
- Madsen, K. B. (1974). *Modern theories of motivation: A comparative metascientific study*. Wiley.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Freeman.
- Marsh, H. W., & Shavelson, R. J. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist*, 20, 107–123.
- Marvin, C. B., & Shohamy, D. (2016). Curiosity and reward: Valence predicts choice and information prediction errors enhance learning. *Journal of Experimental Psychology: General*, 145(3), 266–272. <https://doi.org/10.1037/xge0000140>
- Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, 50(4), 370–396. <https://doi.org/10.1037/h0054346>
- McClelland, D. C. (1987). *Human motivation*. Cambridge University Press.
- McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1976). *The achievement motive*. Irvington.
- McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review*, 96, 690–702.
- McClure, K., Jacobucci, R., & Ammerman, B. A. (2021). Are items more than indicators? An examination of psychometric homogeneity, item-specific effects, and consequences for structural equation models. *PsyArXiv*. <https://doi.org/10.31234/osf.io/n4mxv>
- McDougall, W. (1909). The principal instincts and the primary emotions of man. In W. McDougall (Ed.), *An introduction to social psychology* (2nd ed., pp. 45–89). John W Luce & Company. <https://doi.org/10.1037/13634-003>
- Murayama, K. (2022). A reward-learning framework of knowledge acquisition: An integrated account of curiosity, interest, and intrinsic-extrinsic rewards. *Psychological Review*, 129(1), 175–198. <https://doi.org/10.1037/rev0000349>
- Murayama, K. (2023a). Are cognition, motivation, and emotion the same or different?: Let's abandon that thinking. In M. Bong, J. Reeve, & S. Kim (Eds.), *Motivation science: Controversies and insights* (pp. 243–245). Oxford University Press. <https://doi.org/10.1093/oso/9780197662359.003.0011>
- Murayama, K. (2023b). Motivation resides only in our language, not in our mental processes. In M. Bong, J. Reeve, & S. Kim (Eds.), *Motivation science: Controversies and*

- insights (pp. 65–69). Oxford University Press. <https://doi.org/10.1093/oso/9780197662359.003.0011>
- Murayama, K., FitzGibbon, L., & Sakaki, M. (2019). Process account of curiosity and interest: A reward-learning perspective. *Educational Psychology Review*, 31(4), 875–895. <https://doi.org/10.1007/s10648-019-09499-9>
- Murayama, K., Izuma, K., Aoki, R., & Matsumoto, K. (2017). “Your choice” motivates you in the brain: The emergence of autonomy neuroscience. In S. Kim, J. Reeve, & M. Bong (Eds.), *Advances in motivation and achievement* (Vol. 19, *Recent developments in neuroscience research on human motivation*) (pp. 95–125). Emerald.
- Murphy, P. K., & Alexander, P. A. (2000). A motivated exploration of motivation terminology. *Contemporary Educational Psychology*, 25(1), 3–53. <https://doi.org/10.1006/ceps.1999.1019>
- Murray, H. A. (1938). *Explorations in personality*. Oxford University Press.
- Nagengast, B., & Trautwein, U. (2023). Theoretical and methodological disintegration is the most fundamental limitation in contemporary motivation research. In M. Bong, J. Reeve, & S. Kim (Eds.), *Motivation science: Controversies and insights* (pp. 419–424). Oxford University Press. <https://doi.org/10.1093/oso/9780197662359.003.0068>
- Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.
- Ningombam, D. D., Yoo, B., Kim, H. W., Song, H. J., & Yi, S. (2022). CuMARL: Curiosity-based learning in multiagent reinforcement learning. *IEEE Access*, 10, 87254–87265. <https://doi.org/10.1109/ACCESS.2022.3198981>
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259. <https://doi.org/10.1037/0033-295X.84.3.231>
- Niv, Y., Joel, D., & Dayan, P. (2006). A normative perspective on motivation. *Trends in Cognitive Sciences*, 10, 375–381.
- O'Reilly, R. C. (2020). Unraveling the mysteries of motivation. *Trends in Cognitive Sciences*, 24(6), 425–434. <https://doi.org/10.1016/j.tics.2020.03.001>
- Oudeyer, P.-Y., & Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurobotics*, 1, 6. <https://doi.org/10.3389/neuro.12.006.2007>
- Palminteri, S., & Lebreton, M. (2022). The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences*, 26(7), 607–621. <https://doi.org/10.1016/j.tics.2022.04.005>
- Patanekar, S. P., Zhou, D., Lynn, C. W., Kim, J. Z., Ouellet, M., Ju, H., ... Bassett, D. S. (2023). Curiosity as filling, compressing, and reconfiguring knowledge networks. *Cognitive Intelligence*, 2(4), 26339137231207633. <https://doi.org/10.1177/26339137231207633>
- Pekrun, R. (2023). Jingle-jangle fallacies in motivation science: Toward a definition of core motivation. In M. Bong, J. Reeve, & S. Kim (Eds.), *Motivation science: Controversies and insights* (pp. 52–58). Oxford University Press. <https://doi.org/10.1093/oso/9780197662359.003.0009>
- Pekrun, R., & Marsh, H. W. (2022). Research on situated motivation and emotion: Progress and open problems. *Learning and Instruction*, 81, 101664. <https://doi.org/10.1016/j.learninstruc.2022.101664>
- Pessoa, L., Medina, L., & Desfilis, E. (2022). Refocusing neuroscience: Moving away from mental categories and towards complex behaviours. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1844), 20200534. <https://doi.org/10.1098/rstb.2020.0534>
- Pittman, T. S., & Zeigler, K. R. (2007). Basic human needs. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (2nd ed., pp. 473–489). Guilford Press.
- Polí, F., Meyer, M., Mars, R. B., & Hunnius, S. (2022). Contributions of expected learning progress and perceptual novelty to curiosity-driven exploration. *Cognition*, 225, 105119. <https://doi.org/10.1016/j.cognition.2022.105119>
- Reeder, G. (2009). Mindreading: Judgments about intentionality and motives in dispositional inference. *Psychological Inquiry*, 20(1), 1–18. <https://doi.org/10.1080/10478400802615744>
- Reeve, J. (2017). *Understanding motivation and emotion* (7th ed.). Wiley.
- Renninger, K. A., & Hidi, S. (2016). *The power of interest for motivation and engagement*. Routledge.
- Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, 25(1), 30–45. <https://doi.org/10.1037/met0000220>
- Richter, M., Gendolla, G. H. E., & Wright, R. A. (2016). Three decades of research on motivational intensity theory: What we have learned about effort and what we still don't know. *Advances in Motivation Science*, 3, 149–186.
- Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill, factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, 130(2), 261–288. <https://doi.org/10.1037/0033-2909.130.2.261>
- Rohrer, J. M., & Murayama, K. (2023). These are not the effects you are looking for: Causality and the within-/between-persons distinction in longitudinal data analysis. *Advances in Methods and Practices in Psychological Science*, 6(1), 25152459221140842. <https://doi.org/10.1177/25152459221140842>
- Rolls, E. T. (2016). Motivation explained: Ultimate and proximate accounts of hunger and appetite. In A. J. Elliot (Ed.), *Advances in motivation science* (Vol. 3, pp. 187–249). Elsevier. <https://doi.org/10.1016/bs.adms.2015.12.004>
- Rozeek, C. S., Svoboda, R. C., Harackiewicz, J. M., Hulleman, C. S., & Hyde, J. S. (2017). Utility-value intervention with parents increases students' STEM preparation and career pursuit. *Proceedings of the National Academy of Sciences of the United States of America*, 114(5), 909–914. <https://doi.org/10.1073/pnas.1607386114>
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145–172. <https://doi.org/10.1037/0033-295X.110.1.145>
- Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Press.
- Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, 61, 101860. <https://doi.org/10.1016/j.cedpsych.2020.101860>
- Sachisthal, M. S. M., Jansen, B. R. J., Peetsma, T. T. D., Dalege, J., van der Maas, H. L. J., & Raijmakers, M. E. J. (2019). Introducing a science interest network model to reveal country differences. *Journal of Educational Psychology*, 111(6), 1063–1080. <https://doi.org/10.1037/edu0000327>
- Schultheiss, O. C., Campbell, K. L., & McClelland, D. C. (1999). Implicit power motivation moderates men's testosterone responses to imagined and real dominance success. *Hormones and Behavior*, 36(3), 234–241. <https://doi.org/10.1006/hbeh.1999.1542>
- Schultheiss, O. C., Yankova, D., Dirlikov, B., & Schad, D. J. (2009). Are implicit and explicit motive measures statistically independent? A fair and balanced test using the picture story exercise and a cue- and response-matched questionnaire measure. *Journal of Personality Assessment*, 91(1), 72–81. <https://doi.org/10.1080/00223890802484456>
- Sedikides, C., & Strube, M. J. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be better. *Advances in Experimental Social Psychology*, 29, 209–269.
- Seward, G. H. (1939). Dialectic in the psychology of motivation. *Psychological Review*, 46, 46–61. <https://doi.org/10.1037/h0057732>
- Sheldon, K. M. (2022). *Freely determined: What the new psychology of the self teaches us about how to live*. Basic Books.
- Shultz, T. R., & Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review*, 103(2), 219–240. <https://doi.org/10.1037/0033-295X.103.2.219>
- Simpson, E. H., & Balsam, P. D. (2016). The behavioral neuroscience of motivation: An overview of concepts, measures, and translational applications. *Current Topics in Behavioral Neurosciences*, 27, 1–12. https://doi.org/10.1007/7854_2015_402
- Skinner, E. A. (1996). A guide to constructs of control. *Journal of Personality and Social Psychology*, 71, 549–570.
- Smith, B. J., & Read, S. J. (2022). Modeling incentive salience in Pavlovian learning more parsimoniously using a multiple attribute model. *Cognitive, Affective, & Behavioral Neuroscience*, 22(2), 244–257. <https://doi.org/10.3758/s13415-021-00953-2>
- Smith, G. T., Fischer, S., & Fister, S. M. (2003). Incremental validity principles in test construction. *Psychological Assessment*, 15(4), 467–477.
- Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 261–302). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60229-4](https://doi.org/10.1016/S0065-2601(08)60229-4)
- Steinman, M. Q., Duque-Wilckens, N., & Trainor, B. C. (2019). Complementary neural circuits for divergent effects of oxytocin: Social approach versus social anxiety. *Biological Psychiatry*, 85(10), 792–801. <https://doi.org/10.1016/j.biopsych.2018.10.008>
- Stevens, L. E., & Fiske, S. T. (1995). Motivation and cognition in social life: A social survival perspective. *Social Cognition*, 13(3), 189–214. <https://doi.org/10.1521/soco.1995.13.3.189>
- Stich, S. P., & Ravenscroft, R. (1994). What is folk psychology? *Cognition*, 50, 447–468.
- Tamura, A., Ishii, R., Yagi, A., Fukuzumi, N., Hatano, A., Sakaki, M., ... Murayama, K. (2022). Exploring the within-person contemporaneous network of motivational engagement. *Learning and Instruction*, 81, 101649. <https://doi.org/10.1016/j.learninstruc.2022.101649>
- Utman, C. H. (1997). Performance effects of motivational state: A meta-analysis. *Personality and Social Psychology Review*, 1(2), 170–182. https://doi.org/10.1207/s15327957pspr0102_4
- Vancouver, J. B., Wang, M., & Li, X. (2020). Translating informal theories into formal theories: The case of the dynamic computational model of the integrated model of work motivation. *Organizational Research Methods*, 23(2), 238–274. <https://doi.org/10.1177/1094428118780308>
- VandenBos, G. R., & American Psychological Association (Ed.) (2007). *APA dictionary of psychology* (1st ed.). American Psychological Association.
- van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, 113(4), 842–861. <https://doi.org/10.1037/0033-295X.113.4.842>
- VanderWeele, T. J. (2022). Constructed measures and causal inference: Towards a new model of measurement for psychosocial constructs. *Epidemiology (Cambridge, MA)*, 33(1), 141. <https://doi.org/10.1097/EDE.0000000000001434>
- van Lieshout, L. L. F., Vandenbroucke, A. R. E., Müller, N. C. J., Cools, R., & de Lange, F. P. (2018). Induction and relief of curiosity elicit parietal and frontal activity. *Journal of Neuroscience*, 38(10), 2579–2588. <https://doi.org/10.1523/JNEUROSCI.2816-17.2018>
- van Rooij, L., & 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>

- Wegner, D. M. (2004). Précis of The illusion of conscious will. *Behavioral and Brain Sciences*, 27(5), 649–659. <https://doi.org/10.1017/S0140525X04000159>
- Weiner, B. (1992). *Human motivation: Metaphors, theories, and research*. Sage.
- White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review*, 66(5), 297–333.
- Wigfield, A., Muenks, K., & Eccles, J. S. (2021). Achievement motivation: What we know and where we are going. *Annual Review of Developmental Psychology*, 3(1), 87–111. <https://doi.org/10.1146/annurev-devpsych-050720-103500>
- Zhang, J., Berridge, K. C., Tindell, A. J., Smith, K. S., & Aldridge, J. W. (2009). A neural computational model of incentive salience. *PLoS Computational Biology*, 5(7), e1000437. <https://doi.org/10.1371/journal.pcbi.1000437>

Open Peer Commentary

Endogenous reward is a bridge between social/cognitive and behavioral models of choice

George Ainslie* 

Department of Veterans Affairs Medical Center, Coatesville, PA, USA
info@picoeconomics.org
www.picoeconomics.org

*Corresponding author.

doi:10.1017/S0140525X24000463, e25

Abstract

Endogenous reward (intrinsic reward at will) is a *fiat currency* that is *occasioned* by steps toward any goals which are challenging and/or uncommon enough to prevent its debasement by inflation. A “theory of mental computational processes” should propose what properties let goals grow from appetites for endogenous rewards. Endogenous reward may be the universal selective factor in all modifiable mental processes.

The authors propose a promising link between social/cognitive and behavioral approaches to choice by arguing that the course of purely mental processes is determined by the same kind of reward that governs external goal pursuit (citing Murayama, 2022). This conclusion is acceptable to any behaviorist who is not still bound by the old Skinnerian dictum against dealing with mental processes. It perhaps marks the falling away of the last theoretical barrier between the two approaches. The target article does not pursue a remaining holdover, its odd distinction between motivation and reward, but rather calls for moving on to find “the type of information that is perceived as rewarding” (target article, sect. 4.3).

Certainly Figure 2 depicts a perfect reward-learning model. The authors argue that the various high-level goals described in social/cognitive psychology are not natural types but rather clusters of related outcomes. They are not elementary variables but “black boxes” (e.g., target article, sect. 3.1), the building blocks of which are computed values. Thus, “priority should be given to understanding the underlying computational mechanisms” (target article, sect. 4.3). Behavioral reward theory has already provided candidates for such building blocks, by analyzing the most elementary computations as sequences of binary choices.

At least with external rewards, brain imaging shows rehearsal of sequences leading up to choices – *vicarious trial and error* – that looks like human deliberation (Redish, 2016). As with a computer, such binary choices should be able to form high-order processes of great complexity.

As for what makes information rewarding, the authors settle on curiosity that is induced by awareness of “information gaps” (target article, sect. 4.2) which lead to the experiences of “novelty, uncertainty, conflict, complexity, etc.” (Fig. 2); but even these will require some fleshing out to become “basic building blocks of rewarding value” (target article, sect. 5.2). Here is where a behavioral approach can make further contributions. Whatever the reward is for narrowing the knowledge gap or for satisfying curiosity, it should (a) perform like rewards that have been studied in other contexts; (b) have a variable effect over a time course; and (c) depend on some kind of appetite.

- (a) As an example of intrinsic reward performing like other kinds: Opportunities to add up small numbers can be shown to reward human subjects’ attention-paying responses so these follow Herrnstein’s matching law of reward, despite no physical behavior and no feedback about getting the answers right (Heyman & Moncaleano, 2021).
- (b) High-order goals are clearly subject to nonexponential discounting – probably hyperbolic – which entails the overeffectiveness of near-term rewards. The pursuit of almost any of the authors’ examples requires solving intertemporal conflicts: Long-term competence is threatened by present-paying laziness, self-esteem by impulsiveness, and so on. The formation of high-order goals may depend on identifying larger but more delayed rewards that can overcome faster-rewarding alternatives only by making common cause with similar delayed rewards, that is, by being perceived as serving a shared aspiration that is at stake in each relevant choice (Ainslie, 1992, pp. 144–162; 2005, pp. 5–9; Read, Lowenstein, & Rabin, 1999; but see also Rachlin, 1995).
- (c) Appetite will be the key variable in rewards that do not depend on physical states, and in the goals built from them. Given the variety of goals that people have made central to their lives, including powers, loves, collections, knowledge, faiths, theories, and delusions, sources of intrinsic reward are probably not limited to inborn turn-key patterns – not the “inherently interesting or enjoyable” (Deci & Ryan, 1985) – but open to arbitrary choice. This author has elsewhere proposed that much reward is thus not just *intrinsic* but *endogenous*, available at will (Ainslie, 1992, pp. 243–263; 2013, pp. 8–13; 2017, pp. 178–184). Some attention-directing skill is apparently learned to promote a goal above moment-to-moment satisfactions by holding off reward until appetite gets strong, and only then harvesting it. But this skill will itself be subject to hyperbolic impatience, so it must find criteria outside of its control for harvesting its investment. For a basic example without extrinsic incentives: A solitary player is deterred from claiming a win until the cards *occasion* it by remembering that cheating has always made the investment worthless. The art of exploiting endogenous reward is to find *occasions* that are *singular* – distinct and infrequent – enough to prevent a reward’s overuse and hence inflation.

Repeated successes make some kinds of gambles lose value through habituation, but let others stand out by revealing related

gambles that build appetite. Complex patterns of occasioning should sometimes proliferate into major preoccupations or lifestyles (outlined in Ainslie, 2013), and thus form high-level goals as in the authors' examples (target article, sect. 4.2). Endogenous reward is a *fiat currency* in which agents can indulge freely, limited only by depletion of their appetite for it. However, such reward is subject to competition by not only extrinsic incentives, but also by different patterns of endogenous reward that build their appetite and harvest their reward on alternative time-tables. Painful thoughts, for instance, would offer a combination of rapid reward for attention but with an inhibition of other sources of reward.

The authors have only begun to tap the radical potential of endogenous reward, which is, in effect, a behavior (Ainslie, 2023, pp. 19–22). If reward governs cognitive functions in general, it may be the universal selective factor in all modifiable mental processes. Of course tendencies toward many responses are inborn – for example, in emotions such as anger after frustration or fear when facing danger, or in the authors' example of orienting attention to an object that suddenly appears (target article, sect. 5.2). Obedience to pre-existing tendencies is conventionally ascribed to unmotivated black-box factors such as incentive salience, simple pairing (“conditioning”), or the actor's having formed a habit. But many examples show pre-existing tendencies to be modifiable in the marketplace of motivation: They can be overcome by competing rewards with the right magnitude and timing, as in learned emotional control (Ainslie & Monterosso, 2005) or strong attentional focus (Beecher, 1948). After all, incentive salience is still incentive, emotions all have valences (Miller, 1969) and conditioning doesn't occur to neutral unconditioned stimuli (Goldwater, 1972, pp. 350–351). As for habits, even rats switch flexibly into and out of them (Keramati, Smittenaar, Dolan, & Dayan, 2016). Mental processes in general may be *pulled* by reward much more than they are *pushed* by prior stimuli.

Financial support. This material was supported by the Department of Veterans Affairs Medical Center, Coatesville, PA, USA. The opinions expressed are not those of the Department of Veterans Affairs or the US Government.

Competing interest. None.

References

- Ainslie, G. (1992). *Picoeconomics: The strategic interaction of successive motivational states within the person*. Cambridge U.
- Ainslie, G. (2005). Précis of Breakdown of Will. *Behavioral and Brain Sciences*, 28(5), 635–673.
- Ainslie, G. (2013). Grasping the impalpable: The role of endogenous reward in choices, including process addictions. *Inquiry*, 56, 446–469.
- Ainslie, G. (2017). De gustibus disputare: Hyperbolic delay discounting integrates five approaches to choice. *Journal of Economic Methodology*, 24(2), 166–189.
- Ainslie, G. (2023). Behavioral construction of the future. *Psychology of Addictive Behaviors*, 37(1), 13–24.
- Ainslie, G., & Monterosso, J. (2005). Why not emotions as motivated behaviors? *Behavior and Brain Sciences*, 28, 194–195.
- Beecher, H. K. (1948). Pain in men wounded in battle. *Annals of Surgery*, 123(1), 96–105.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Plenum.
- Goldwater, B. C. (1972). Psychological significance of pupillary movements. *Psychological Bulletin*, 77(5), 340.
- Heyman, G. M., & Moncaleano, S. (2021). Behavioral psychology's matching law describes the allocation of covert attention: A choice rule for the mind. *Journal of Experimental Psychology: General*, 150(2), 195.
- Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. *Proceedings of the National Academy of Sciences*, 113(45), 12868–12873.
- Miller, N. (1969). Learning of visceral and glandular responses. *Science*, 163, 434–445.
- Murayama, K. (2022). Motivation resides only in our language, not in our mental processes. In M. Bong, S. Kim, & J. Reeve (Eds.), *Motivation science: Controversies and insights*. Oxford University Press.
- Rachlin, H. (1995). Behavioral economics without anomalies. *Journal of the Experimental Analysis of Behavior*, 64, 396–404.
- Read, D., Lowenstein, G., & Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty*, 19(1), 171–197.
- Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience*, 17(3), 147–159.

Resurrecting the “black-box” conundrum

Patricia A. Alexander* 

Department of Human Development and Quantitative Methodology, University of Maryland, College Park, College Park, MD, USA
palexand@umd.edu

*Corresponding author.

doi:10.1017/S0140525X24000451, e26

Abstract

In their article, Murayama and Jach contend that a mental computational model demonstrates that high-level motivations are emergent properties from underlying cognitive processes rather than instigators of behaviors. Despite points of agreement with the authors' critiques of the motivation literature, I argue that their claim of dismantling the black box of the human mind has been constructed on shaking grounds.

Resurrecting the “black-box” conundrum

In their provocative treatise entitled, “A critique of motivation constructs to explain higher-order behavior: We should unpack the black box,” Murayama and Jach (M&J) offer a detailed analysis of the motivation research and the causal claims populating that literature regarding the initiating of behaviors. The “black box” to which the authors refer is the longstanding contention that aspects of human mental functioning are not accessible for reflection or analysis (Skinner, 1989), even by individuals executing those functions. As a counterpoint to the black-box argument, the authors offer an alternative framework for investigating the complex “motivation-behavior” enigma based on mental computational modeling. Their mental computational model promises nothing short of a solution to the black-box problem. As the authors boldly stated:

By specifying the mental computational processes underlying higher-order motivated behavior, high-level motivation constructs are no longer black boxes.

As I will discuss, I agree with several insights the authors draw from their critical analysis of motivation research. That agreement notwithstanding, my principal contention is that the authors' bold claim of unpacking or dismantling the black box of the human mind has been constructed on theoretically shaky grounds. The justifications for this counter-position are the authors' oversimplification of the complex and dynamic nature of mental functioning

and the questionable conceptualizations guiding their mental computational model.

Points of agreement

As noted, several of M&J's critiques of the motivation literature and its explanation of human behavior are well-founded. For one, motivation is *not* a unitary construct. As well documented in the philosophical and psychological literature (Skinner, 2023), motivation is a meta-term encompassing innumerable constructs and sub-constructs. Those constructs and sub-constructs can be domain-general or domain- and task-specific, trait-like or state-like, and tacit or explicit. This plethora of terms means that there is an inherent vagueness when we speak about human motivation that has been exacerbated by the multitude of labels generated to identify this ever-growing litany of forms (Alexander, 2024; Alexander, Grossnickle, & List, 2014). Broadening this conceptual morass, researchers frequently fail to define what aspect of motivation they are addressing, coin a new term when relevant labels exist, or use an existing word to mean something else (Dinsmore, Alexander, & Loughlin, 2008; Murphy & Alexander, 2000). Some refer to this phenomenon, which my colleagues and I have repeatedly documented, as the "jingle-jangle fallacy" (Bong, 1996; Pekrun, 2023). I prefer to identify this simply as "poor science." Moreover, such conceptual ambiguity carries over into research measures and procedures, as M&J assert.

One of the persistent criticisms of motivation is that researchers often rely on participants' self-reports as the primary or sole evidence (King & Fryer, 2024). In their defense, motivation theorists and researchers counter this criticism by arguing that motivation remains largely in the realm of beliefs or perceptions, which are potent forces in how humans think and act (Greene, 2015; Van Meter, 2020). I do not deny that perceptions can, at times, be more powerful than reality in explaining human thoughts and actions (Alexander & Baggetta, 2014; Hurley, 2001). Nonetheless, I agree with M&J that corroborating those self-reports with biophysiological or neurocognitive data would strengthen what can be inferred or predicted about *why* humans think and act as they do.

M&J also asserted that "when the target construct is not unambiguously defined, we can never make a solid causal inference from empirical data"; another point of consensus with the authors (Alexander, 2013, 2024). However, meeting standards that allow for causal claims is challenging even when researchers explicitly define their constructs (Steiner, Shadish, & Sullivan, 2023).

Counterpoints

Agreements aside, I ultimately take issue with the authors' overall contention that their proposed mental computational model will effectively dismantle the black box of the human mind and provide solutions to "why" questions about motivation and behavior. First and foremost, the functioning of the mind cannot be reduced to simple linear or reciprocal models like those M&J promote. In effect, even seemingly uncomplicated behaviors can arise from a confluence of internal and external forces that operate dynamically and that can remain below human awareness.

Regrettably, the authors' efforts to narrow the scope of their modeling to what they regard as high-level motivations and higher-order behaviors cannot constrain mental functioning to the degree required to resolve the black-box problem. For one, the defining attributes of high-level versus low-level motivation

constructs are notoriously nebulous, as the authors rightfully acknowledged. Further, even if a more defensible distinction for high-level motivation constructs were possible, the most basic or primary drives underlying human functioning could align with intentional and complex goals and, thus, with subsequent behaviors (Butler & Rice, 1963). Additionally, there can be competing goals or mixed motives that accompany non-automatic or reflexive thoughts and actions (Linnenbrink & Pintrich, 2001). Thus, singular paths or directional hypotheses about the "motivation-behavior" enigma are not theoretically defensible even if they can be empirically demonstrated.

Finally, there are too many intervening and unacknowledged internal and external factors that can morph or redirect individuals' motivations *and* their concomitant actions. Consequently, the claim that motivation is "an emergent property that people construe" (M&J) is as indefensible as claims that motivations initiate behaviors. Even the most sophisticated of mental computational models cannot establish such directionality when it comes to the complex and dynamic of human motivations or human behaviors, however defined.

Financial support. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sector.

Competing interest. The author declares that she has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Alexander, P. A. (2013). In praise of (reasoned and reasonable) speculation: A response to Robinson et al.'s moratorium on recommendations for practice. *Educational Psychology Review*, 25(2), 303–308. <https://doi.org/10.1007/s10648-013-9234-2>
- Alexander, P. A. (2024). Hybridizing psychological theories: Weighing the ends against the means. *Educational Psychology Review*, 36(1), 23. <https://doi.org/10.1007/s10648-024-09856-3>
- Alexander, P. A., & Baggetta, P. (2014). Percept-concept coupling and human error. In D. N. Rapp & J. L. G. Baasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 297–327). MIT Press.
- Alexander, P. A., Grossnickle, E. M., & List, A. (2014). Navigating the labyrinth of teacher motivations and emotions. In P. Richardson, S. Karabenick & H. Watt (Eds.), *Teacher motivation: Theory and practice* (pp. 150–163). Routledge.
- Bong, M. (1996). Problems in academic motivation research and advantages and disadvantages of their solutions. *Contemporary Educational Psychology*, 21, 149–165. <https://doi.org/10.1006/ceps.1996.0013>
- Butler, J. M., & Rice, L. N. (1963). Audience, self-actualization, and drive theory. In J. M. Wepman & R. W. Heine (Eds.), *Concepts of personality* (pp. 79–110). Aldine Publishing Co. <https://doi.org/10.1037/11175-004>
- Dinsmore, D. L., Alexander, P. A., & Loughlin, S. M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review*, 20, 391–409. <https://doi.org/10.1007/s10648-008-9083-6>
- Greene, B. A. (2015). Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research. *Educational Psychologist*, 50(1), 14–30. <https://doi.org/10.1080/00461520.2014.989230>
- Hurley, S. (2001). Perception and action: Alternative views. *Synthese*, 129, 3–40. <https://doi.org/10.1023/A:1012643006930>
- King, R. B., & Fryer, L. K. (2024). Hybridizing motivational strains: How integrative models are crucial for advancing motivation science. *Educational Psychology Review*, 36, 38. <https://doi.org/10.1007/s10648-024-09850-9>
- Linnenbrink, E. A., & Pintrich, P. R. (2001). Multiple goals, multiple contexts: The dynamic interplay between personal goals and contextual goal stresses. In S. Volet & S. Järvelä (Eds.), *Motivation in learning contexts: Theoretical advances and methodological implications* (pp. 251–269). Pergamon Press.
- Murphy, P. K., & Alexander, P. A. (2000). A motivated exploration at motivation terminology. [Special Issue]. *Contemporary Educational Psychology*, 25(1), 3–53. <https://doi.org/10.1006/ceps.1999.1019>
- Pekrun, R. (2023). Jingle-jangle fallacies in motivation science: Toward a definition of core motivation. In M. Bong, J. Reeve & S. Kim (Eds.), *Motivation science: Controversies and insights* (pp. 52–58). Oxford University Press. <https://doi.org/10.1093/oso/9780197662359.003.0009>

- Skinner, B. F. (1989). The origins of cognitive thought. *American Psychologist*, 44(1), 13. <https://doi.org/10.1037/0003-066X.44.1.13>
- Skinner, E. A. (2023). Four guideposts toward an integrated model of academic motivation: Motivational resilience, academic identity, complex social ecologies, and development. *Educational Psychology Review*, 35(3), 80. <https://doi.org/10.1007/s10648-023-09790-w>
- Steiner, P. M., Shadish, W. R., & Sullivan, K. J. (2023). Frameworks for causal inference in psychological science. In H. Cooper, M. N. Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Foundations, planning, measures, and psychometrics* (2nd ed., pp. 23–56). American Psychological Association. <https://doi.org/10.1037/0000318-002>
- Van Meter, P. N. (2020). Commentary: Measurement and the study of motivation and strategy use: Determining if and when self-report measures are appropriate. *Frontline Learning Research*, 8(3), 174–184. <https://doi.org/10.14786/flr.v8i3.631>

Exploring novelty to unpack the black-box of motivation

Nico Bunzeck^{a*}  and Sebastian Haesler^b 

^aDepartment of Psychology, and Center of Brain, Behavior and Metabolism, University of Lübeck, Lübeck, Germany and ^bNeuroelectronics Research Flanders (NERF), and Department of Neuroscience, KU Leuven, Leuven, Belgium
nico.bunzeck@uni-luebeck.de
sebastian.haesler@nerf.be
<https://www.ipsy1.uni-luebeck.de/>
<https://haeslerlab.com>

*Corresponding author.

doi:10.1017/S0140525X24000505, e27

Abstract

Murayama and Jach point out that we do not sufficiently understand the constructs and mental computations underlying higher-order motivated behaviors. Although this may be generally true, we would like to add and contribute to the discussion by outlining how interdisciplinary research on *novelty-evoked exploration* has advanced the study of learning and curiosity.

When confronted with a novel or unexpected stimulus, such as a sudden innocuous noise in an otherwise silent environment, humans and other animals orient toward the source of the stimulus. This “orienting reflex” is an expression of increased attention and has long been linked to specific components in the electroencephalography signal (Sokolov, 1963). Novelty also drives a wide range of higher-order behaviors through diverse sets and intertwined neural mechanisms. For instance, novelty can promote spatial memory via changes in dopamine dependent long-term potentiation in rats (Wang, Redondo, & Morris, 2010), cue-evoked dopamine promotes associative learning in mice (Morrens, Aydin, Janse van Rensburg, Esquivelzeta Rabell, & Haesler, 2020), novelty shapes salience memories via activity in the prefrontal cortex and basal ganglia in monkeys (Ghazizadeh & Hikosaka, 2022), and, also in monkeys, basal forebrain neurons distinguish between novelty and familiarity in recognition memory tasks (Wilson & Rolls, 1990). This is compatible with human imaging studies showing that novelty processing improves subsequent recognition memory via neural beta oscillations (13–25 Hz) (Steiger, Sobczak, Reineke, & Bunzeck, 2022), and the fact that novelty responses in the human substantia nigra (SN)/ventral tegmental area (VTA) are modulated by dopaminergic and cholinergic

stimulation (Bunzeck, Guitart-Masip, Dolan, & Düzel, 2014). Finally, environmental novelty enhances memory via co-release of dopamine from locus coeruleus axon terminals in the hippocampus (Takeuchi et al., 2016).

Although these findings help to clarify how novelty affects exploration and learning, the underlying motives remain unclear. From an evolutionary perspective, exploring the unknown can be advantageous for several reasons. First, it can help to reduce uncertainty in light of possible rewards and punishments (e.g., nutritious vs. poisonous foods, or harmful vs. harmless snakes, etc.). Second and along these lines, novelty-evoked exploration can promote cognitive and behavioral flexibility since it requires adaptation and possibly innovative solutions (e.g., learning how to open a coconut without losing the valuable milk). Third, together with potentially increasing genetic diversity and escaping current adverse living conditions, the exploration of novel environments can, therefore, increase the chances for survival both at an individual and group level. As a mechanism to realize motivated behaviors, computational modelling has offered a simple solution. By treating novelty as reward, the dopaminergic reward system creates an “exploration bonus,” which incentivizes exploration similar to how it motivates the search for rewards (Kakade & Dayan, 2002). In line with this notion, contextual novelty changes reward representations in the human striatum (Guitart-Masip, Bunzeck, Stephan, Dolan, & Düzel, 2010; Wittmann, Daw, Seymour, & Dolan, 2008) and the neural dynamics of reward anticipation (Bunzeck, Guitart-Masip, Dolan, & Düzel, 2011). Together with the observation that monetary rewards accelerate the onset of neural novelty signals in humans (Bunzeck, Doeller, Fuentemilla, Dolan, & Düzel, 2009), this further demonstrates the close empirical link between novelty processing and reward motivation.

From a psychological perspective, the motive to explore novelty relates to personality traits and corresponding current states. Specifically, the personality trait novelty-seeking positively correlated with hemodynamic activity in the SN/VTA elicited by novel cues (Krebs, Schott, & Düzel, 2009). Moreover, the temporary and situational intrinsic motivation to acquire new knowledge, also referred to as *state* epistemic curiosity, drives long-term memory via the dopaminergic mesolimbic system (Gruber, Gelman, & Ranganath, 2014; Kang et al., 2009). *Trait* epistemic curiosity, that is, the rather stable and consistent desire to acquire new knowledge, contributes to various aspects of motivation and performance in academic and professional contexts, including goal setting and learning (Litman & Mussel, 2013). The underlying principles of such higher-order curious behaviors can be described by reinforcement learning (RL)-frameworks in which novelty generates an intrinsic reward signal in addition to an extrinsic (primary) reward signal (Modirshanechi, Kondrakiewicz, Gerstner, & Haesler, 2023). By separating intrinsic (novelty-driven) from extrinsic (reward-driven) motivational processes, these RL-frameworks can also explain non-optimal behavior such as continuing to seek novelty even when novelty-seeking leads to distraction by reward-independent stochasticity (Modirshanechi, Xu, Lin, Herzog, & Gerstner, 2022) also referred to as “noisy-TV” problem (Aubret, Matignon, & Hassas, 2023). Importantly, neuroimaging supports the presence of separate intrinsic and extrinsic signals at the neural level (Filimon, Nelson, Sejnowski, Sereno, & Cottrell, 2020).

Finally, disease-related and age-related brain changes, especially within the mesolimbic system and interconnected brain regions, impact on novelty exploration, motivation, and learning. For instance, age-related degeneration of the dopaminergic mid-brain affects neural novelty signals and long-term memory

performance (Düzel, Bunzeck, Guitart-Masip, & Düzel, 2010), whereas iron levels and myelin content in the ventral striatum predict memory performance in older adults (Steiger, Weiskopf, & Bunzeck, 2016). More recent studies have pointed out that locus coeruleus integrity, associated with dopaminergic and noradrenergic neuromodulation, also plays a key role in learning novel information in healthy and pathological aging (Dahl et al., 2019, 2023). Along these lines, behavioral, neurobiological, and computational changes in motivation-based exploration and learning have been identified in several neuropsychiatric conditions, such as attention deficit hyperactivity disorder (Véronneau-Veilleux, Robaey, Ursino, & Nekka, 2022) and psychotic disorders (Kesby, Eyles, McGrath, & Scott, 2018). In fact, in adolescents with high novelty-seeking scores, multimodal biomarkers predicted longitudinal risk behavior (including alcohol drinking) (Qi et al., 2021).

Taken together, *novelty* motivates a wide range of higher-order behaviors, especially learning and associated memory formation, and recent studies gave detailed insights into the underlying psychological, neurobiological, and computational processes across the lifespan. As such, we agree with Murayama and Jach that there are myriad unanswered questions, but we would like to stress that translational and interdisciplinary research has tremendously helped – and will help in the future – to further conceptualize the construct of motivation.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interest. None.

References

- Aubret, A., Matignon, L., & Hassas, S. (2023). An information-theoretic perspective on intrinsic motivation in reinforcement learning: A survey. *Entropy*, 25(2), 327. <https://doi.org/10.3390/e25020327>
- Bunzeck, N., Doeller, C. F., Fuentemilla, L., Dolan, R. J., & Düzel, E. (2009). Reward motivation accelerates the onset of neural novelty signals in humans to 85 milliseconds. *Current Biology*, 19(15), 1294–1300.
- Bunzeck, N., Guitart-Masip, M., Dolan, R. J., & Düzel, E. (2011). Contextual novelty modulates the neural dynamics of reward anticipation. *The Journal of Neuroscience*, 31(36), 12816–12822. <https://doi.org/10.1523/JNEUROSCI.0461-11.2011>
- Bunzeck, N., Guitart-Masip, M., Dolan, R. J., & Düzel, E. (2014). Pharmacological dissociation of novelty responses in the human brain. *Cerebral Cortex*, 24(5), 1351–1360. <https://doi.org/10.1093/cercor/bhs420>
- Dahl, M. J., Mather, M., Düzel, S., Bodammer, N. C., Lindenberger, U., Kühn, S., & Werkle-Bergner, M. (2019). Rostral locus coeruleus integrity is associated with better memory performance in older adults. *Nature Human Behaviour*, 3(11), 1203–1214. <https://doi.org/10.1038/s41562-019-0715-2>
- Dahl, M. J., Kulesza, A., Werkle-Bergner, M., & Mather, M. (2023). Declining locus coeruleus–dopaminergic and noradrenergic modulation of long-term memory in aging and Alzheimer's disease. *Neuroscience & Biobehavioral Reviews*, 153, 105358. <https://doi.org/10.1016/j.neubiorev.2023.105358>
- Düzel, E., Bunzeck, N., Guitart-Masip, M., & Düzel, S. (2010). NOvelty-related Motivation of Anticipation and exploration by Dopamine (NOMAD): Implications for healthy aging. *Neuroscience & Biobehavioral Reviews*, 34(5), 660–669. <https://doi.org/10.1016/j.neubiorev.2009.08.006>
- Filimon, F., Nelson, J. D., Sejnowski, T. J., Sereno, M. I., & Cottrell, G. W. (2020). The ventral striatum dissociates information expectation, reward anticipation, and reward receipt. *Proceedings of the National Academy of Sciences*, 117(26), 15200–15208. <https://doi.org/10.1073/pnas.1911778117>
- Ghazizadeh, A., & Hikosaka, A. (2022). Salience memories formed by value, novelty and aversiveness jointly shape object responses in the prefrontal cortex and basal ganglia. *Nature Communications*, 13(1), 6338. <https://doi.org/10.1038/s41467-022-33514-3>
- Gruber, M. J., Gelman, B. D., & Ranganath, C. (2014). States of curiosity modulate hippocampus-dependent learning via the dopaminergic circuit. *Neuron*, 84, 486–496. <https://doi.org/10.1016/j.neuron.2014.08.060>
- Guitart-Masip, M., Bunzeck, N., Stephan, K. E., Dolan, R. J., & Düzel, E. (2010). Contextual novelty changes reward representations in the Striatum. *The Journal of Neuroscience*, 30(5), 1721–1726. <https://doi.org/10.1523/JNEUROSCI.5331-09.2010>
- Kakade, S., & Dayan, P. (2002). Dopamine: Generalization and bonuses. *Neural Networks*, 15(4–6), 549–559. [https://doi.org/10.1016/S0893-6080\(02\)00048-5](https://doi.org/10.1016/S0893-6080(02)00048-5)
- Kang, M. J., Hsu, M., Krajchich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T., & Camerer, C. F. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science*, 20(8), 963–973. <http://doi.org/10.1111/j.1467-9280.2009.02402.x>
- Kesby, J. P., Eyles, D. W., McGrath, J. J., & Scott, J. G. (2018). Dopamine, psychosis and schizophrenia: The widening gap between basic and clinical neuroscience. *Translational Psychiatry*, 8(1), 1–12. <https://doi.org/10.1038/s41398-017-0071-9>
- Krebs, R. M., Schott, B. H., & Düzel, E. (2009). Personality traits are differentially associated with patterns of reward and novelty processing in the human substantia nigra/ventral tegmental area. *Biological Psychiatry*, 65(2), 103–110. <https://doi.org/10.1016/j.biopsych.2008.08.019>
- Litman, J. A., & Mussel, P. (2013). Validity of the interest-and deprivation-type epistemic curiosity model in Germany. *Journal of Individual Differences*, 34(2), 59–68. <https://doi.org/10.1027/1614-0001/a000100>
- Modirshanechi, A., Xu, H. A., Lin, W.-H., Herzog, M. H., & Gerstner, W. (2022). *The curse of optimism: A persistent distraction by novelty* (p. 2022.07.05.498835). bioRxiv. <https://doi.org/10.1101/2022.07.05.498835>
- Modirshanechi, A., Kondrakiewicz, K., Gerstner, W., & Haesler, S. (2023). Curiosity-driven exploration: Foundations in neuroscience and computational modeling. *Trends in Neurosciences*, 46, 1054–1066. <https://doi.org/10.1016/j.tins.2023.10.002>
- Morrens, J., Aydin, Ç., Janse van Rensburg, A., Esquivelzeta Rabell, J., & Haesler, S. (2020). Cue-evoked dopamine promotes conditioned responding during learning. *Neuron*, 106, 142–153.e7. <https://doi.org/10.1016/j.neuron.2020.01.012>
- Qi, S., Schumann, G., Bustillo, J., Turner, J. A., Jiang, R., Zhi, D., ... IMAGEN Consortium. (2021). Reward processing in novelty seekers: A transdiagnostic psychiatric imaging biomarker. *Biological Psychiatry*, 90(8), 529–539. <https://doi.org/10.1016/j.biopsych.2021.01.011>
- Sokolov, E. N. (1963). Higher nervous functions: The orienting reflex. *Annual Review of Physiology*, 25(1), 545–580. <https://doi.org/10.1146/annurev.ph.25.030163.002553>
- Steiger, T. K., Weiskopf, N., & Bunzeck, N. (2016). Iron level and myelin content in the ventral striatum predict memory performance in the aging brain. *The Journal of Neuroscience*, 36(12), 3552–3558. <https://doi.org/10.1523/JNEUROSCI.3617-15.2016>
- Steiger, T. K., Sobczak, A., Reineke, R., & Bunzeck, N. (2022). Novelty processing associated with neural beta oscillations improves recognition memory in young and older adults. *Annals of the New York Academy of Sciences*, 1511(1), 228–243. <https://doi.org/10.1111/nyas.14750>
- Takeuchi, T., Duszakiewicz, A. J., Sonneborn, A., Spooner, P. A., Yamasaki, M., Watanabe, M., ... Morris, R. G. M. (2016). Locus coeruleus and dopaminergic consolidation of everyday memory. *Nature*, 537(7620), 357–362. <https://doi.org/10.1038/nature19325>
- Véronneau-Veilleux, F., Robaey, P., Ursino, M., & Nekka, F. (2022). A mechanistic model of ADHD as resulting from dopamine phasic/tonic imbalance during reinforcement learning. *Frontiers in Computational Neuroscience*, 16, 849323. <https://doi.org/10.3389/fncom.2022.849323>
- Wang, S.-H., Redondo, R. L., & Morris, R. G. M. (2010). Relevance of synaptic tagging and capture to the persistence of long-term potentiation and everyday spatial memory. *Proceedings of the National Academy of Sciences*, 107(45), 19537–19542. <https://doi.org/10.1073/pnas.1008638107>
- Wilson, F. A., & Rolls, E. T. (1990). Learning and memory is reflected in the responses of reinforcement-related neurons in the primate basal forebrain. *Journal of Neuroscience*, 10(4), 1254–1267. <https://doi.org/10.1523/JNEUROSCI.10-04-01254.1990>
- Wittmann, B. C., Daw, N. D., Seymour, B., & Dolan, R. J. (2008). Striatal activity underlies novelty-based choice in humans. *Neuron*, 58(6), 967–973. <https://doi.org/10.1016/j.neuron.2008.04.027>

Motivation needs cognition but is not just about cognition

Nathalie André^{a*} and Roy F. Baumeister^b

^aSport Sciences Faculty, University of Poitiers, Poitiers, France and

^bDepartment of Psychology, Harvard University, Cambridge, MA, USA
nathalie.andre@univ-poitiers.fr
r.baumeister@uq.edu.au

*Corresponding author.

doi:10.1017/S0140525X24000529, e28

Abstract

Murayama and Jach offer valuable suggestions for how to integrate computational processes into motivation theory, but these processes cannot do away with motivation altogether. Rewards are only rewarding because people want and like them – that is, because of motivation. Sexual desire is not primarily a quest for rewarding information. Elucidating the interface between motivation and cognition seems a promising way forward.

Motivation theory has long been a battleground, as evidenced in part by its long history of competing lists of basic drives and motives (or wants and needs), with no strong method for evaluating such lists. One perennially attempted solution has been to reduce motivation to cognition. Murayama and Jach (henceforth M&J) provide one of the more intelligent and reasonable efforts of this sort. Perhaps people do not have wants or needs at all – instead they have computational processes.

We are pleased to see that after discussing how to get rid of concepts of motivation, M&J conclude that they are not eliminating motivation after all. They hope to add a new level of analysis that can improve our understanding. This is promising. Computational processes might not replace motivation after all but elaborate how motivations work.

At points they do say that motivation processes can be re-framed as computational processes, such as by seeking rewarding information. But, crucially – what makes a reward rewarding? The answer is that the person wants or needs it. (This applies to information also.) Motivation is something inside the person (or other animal) that makes the reward appealing. Identical stimuli can be highly rewarding to some people but completely indifferent to others. The rewardingness is not in the stimulus but in the person, or at best in its relationship to the person. Computers can perform calculations faster and better than humans, but information is not rewarding to the computer – precisely because the computer lacks motivation. By and large, cognition serves motivation, but both are intertwined and cannot replace each other. Motivation without cognition would be endless frustration. Cognition without motivation would not know what to do.

We think motivation is fundamental to psychology, because it is most closely linked to sustaining life. Organisms evolved to want food, safety, sleep, sex, and the like, and these motivations helped them to survive and reproduce. Computers, contrast, are indifferent to whether they survive or reproduce. Cognition can serve motivation by helping the animal understand how to obtain the resources it needs to sustain life. Human cognition can even inhibit behaviors leading to immediate but problematic rewards in order to obtain delayed high-quality rewards.

Much of M&J's analysis relies on information seeking. This seems an atypical example that is exceptionally conducive to attempts to reduce motivation to cognition. Desires for sex, power, social status, money, or even fame would seemingly pose a more formidable challenge. It is not clear to us how to frame such desires as computational processes. Is the desire to copulate at bottom a search for rewarding information? And if so, what makes copulation rewarding? Sexual intercourse is not the filling of a couple's knowledge gap, except perhaps for their first time. Saving money for the proverbial rainy day is likewise not about filling gaps in knowledge.

Moreover, the existence of an information gap is not enough to produce motivation. There are many things that we do not know, and we know that we do not know them, but we are not motivated to learn them, such as the phone numbers of far-off strangers, or

the middle names of thousands of deceased people. To their credit, M&J acknowledge that the reward-learning framework fails to state what kind of information is rewarding (or why). This is a crucial point that is not a detail but indicates a central shortcoming of the entire approach.

Ultimately, something crucial is missing. The desire for information is not the most basic human drive, from which all other seemingly motivated patterns can be deduced. In our view, motivation evolved as a subjective craving for things that contributed to biological success at survival and reproduction. Yes, having good information helps those things. But more fundamentally, sex, social status, power, food, safety, and similar things are keys to survival and reproduction. Some of these might not be included under their umbrella concept of “higher-order” motivations. But we wish they would define what differentiates higher-order motivations from others and perhaps provide a short list.

Child development research may also give pause to those who seek to explain motivation with cognition. Redding, Morgan, and Harmon (1988) found that task persistence measures correlated more weakly with cognitive measures among older than younger children, and they concluded that motivation and cognition may become increasingly separate as children grow older. Likewise, babies show exploratory behavior and selective interest – so motivation operates long before there is much in the way of prefrontal computational process.

Furthermore, the hallmarks of depression (i.e., lack of interest, slowness in decision-making, difficulty paying attention, poor concentration, and passive inactivity) seem to be explained by the interaction between motivational and cognitive processes (Grahek, Shenhav, Musslick, Krebs, & Koster, 2019). Put another way, the relationship between reward processing and cognitive control across different clinical population suggests that these two systems remain independent even if they are strongly intertwined. These current perspectives are opposed to that proposed by M&J.

M&J are correct to note the challenges and frustrating problems in motivation theory. Elucidating computational processes may be valuable for filling gaps. But trying to replace motivation theory with cognitive theory has never worked.

Financial support. Not applicable.

Competing interest. N. A. and R. F. B. declare that they have no conflict of interest.

References

- Grahek, I., Shenhav, A., Musslick, S., Krebs, R. M., & Koster, E. H. W. (2019). Motivation and cognitive control in depression. *Neuroscience and Biobehavioral Reviews*, 102, 371–381. <https://doi.org/10.1016/j.neubiorev.2019.04.011>
- Redding, R. E., Morgan, G., & Harmon, R. J. (1988). Mastery motivation in infants and toddlers: Is it greatest when tasks are moderately challenging? *Infant Behavior and Development*, 11, 419–430.

The unboxing has already begun: One motivation construct at a time

Ruud Custers^{a*} , Baruch Eitam^b  and E. Tory Higgins^c

^aPsychology Department, Utrecht University, Utrecht, The Netherlands;

^bPsychology Department, University of Haifa, Haifa, Israel and ^cPsychology Department, Columbia University, New York, NY, USA

r.custers@uu.nl; www.goallab.nl

beitam@psy.haifa.ac.il
tory@psych.columbia.edu

*Corresponding author.

doi:10.1017/S0140525X24000554, e29

Abstract

Murayama and Jach argue that it is not clearly specified how motivation constructs produce behavior and that this black box should be unpacked. We argue that the authors overlook important classic theory and highlight recent research programs that already started unboxing. We feel that without relying on the mechanisms that such programs uncover, the proposed computational approach will be fruitless.

The authors critique the vague use of “high-level” motivation constructs as explanatory variables in various sections of the psychological literature and propose that the “black box” of mechanistic explanation should be opened. We fully concur that in many theories on motivation, processes are underspecified or ignored. At the same time, though, research on the underlying processes in motivation has been steadily going on for more than half a century in various corners of the field. While we acknowledge that there are literatures that utilize motivation constructs for various purposes other than explanatory (e.g., to predict behavior), we highlight here three (of many) very different lines of research (including our own) that have unpacked the black box to identify various degrees of “mechanism.”

First, in focusing heavily on drive theory, the author completely overlooked incentive theory (Bindra, 1974; Bolles, 1972; Toates, 1986), which replaced drive theory as the dominant motivation theory in the 1970s/1980s, as well as the work that spawned from it. According to this theory, needs or motives do not directly affect behavior, but rather change the incentive value of behavioral opportunities and stimuli in the environment. Thus, needs or motives – in combination with deprivation – modulate the value of behavioral goals in the situation at hand, while these in turn energize and give direction to behavior (Custers & Aarts, 2010). Put differently, Custers and Aarts (2005) have argued that the causal starting point of behavior can best be understood to be the environmental cues that activate mental representations of goals, with their effects on the direction of behavior and the effort invested in it being moderated by their subjective value at the time of activation. This subjective value is determined by abstract motivational constructs such as needs and situational variables such as deprivation or discrepancy relative to the goal state. Importantly, goal representations are not magical, and are subject to mechanisms applied to all mental representations such as accessibility and its dynamics (Eitam & Higgins, 2014). This literature therefore provides specific mechanistic paths by which high-level motivation constructs do not cause, but rather moderate the effects of the environment on behavior (see Berridge & Robinson, 1998, for a popular neural implementation of the incentive value theory).

Second, self-regulatory systems theory is the product of a three decades-long research program that details both “chronic” and transient (situational) changes in the balance of motivational orientations (Higgins & Cornwell, 2016). This research program has shown a myriad of effects from a shifting focus from a “prevention” state – minimizing a negative (“−1”) discrepancy between

one’s current state and a baseline state (“0”) – versus a “promotion” state – maximizing a positive discrepancy (“+1”) between one’s current state and a baseline state (“0”). While differing in the level of explanation from our previous example, this research program has been anything but a black box by showing how such motivational orientations are affected by parenting (Manian, Papadakis, Strauman, & Essex, 2006) and how they influence basic processes such as judgment and decision making (Förster, Higgins, & Bianco, 2003; Higgins & Cornwell, 2016).

Our third and final example is the work on reinforcement from sensorimotor predictability (Eitam, Kennedy, & Higgins, 2013), which can be easily cast as an attempt to open the “black box” of the abstract need for autonomy and control. What this body of work shows is that sensorimotor prediction, which is considered part of the brain’s mechanism to execute planned or volitional movement, also serves as a reinforcement signal for “effective” motor plans. More specifically, motor programs that are associated with more successful predictions are reinforced above and beyond their utility or association with tangible rewards. This has shown to occur in healthy adults (Hemed, Bakbani-Elkayam, Teodorescu, Yona, & Eitam, 2020) as well as in clinically depressed individuals (Bakbani-Elkayam, Dolev-Amit, Hemed, Zilcha-Mano, & Eitam, 2024), and more recently in a mouse model. This process is another example that hardly fits the author’s depiction of motivational concepts as “initiating behavior,” as it reflects a subtle interplay between environmental input and a computational process, together, creating a direction of behavior.

Thus, such efforts to explain how high-level motivation constructs affect behavior have been going on for quite a while, admittedly with varying degrees of success. Given the above, we suggest that any effort to advance a general framework of motivation would benefit tremendously from the more local work that has already been done to unpack the relevant black boxes. Such work may provide the necessary glue to connect abstract motivation constructs to the lower computation level, at which behavior is controlled by rewards.

Financial support. None.

Competing interest. None.

References

- Bakbani-Elkayam, S., Dolev-Amit, T., Hemed, E., Zilcha-Mano, S., & Eitam, B. (2024). Intact modulation of response vigor in major depressive disorder. *Motivation and Emotion*, 48(2), 209–221. <https://doi.org/10.1007/s11031-024-10059-0>
- Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, 28(3), 309–369. [https://doi.org/10.1016/S0165-0173\(98\)00019-8](https://doi.org/10.1016/S0165-0173(98)00019-8)
- Bindra, D. (1974). A motivational view of learning, performance, and behavior modification. *Psychological Review*, 81(3), 199–213. <https://doi.org/10.1037/h0036330>
- Bolles, R. C. (1972). Reinforcement, expectancy, and learning. *Psychological Review*, 79(5), 394–409. <https://doi.org/10.1037/h0033120>
- Custers, R., & Aarts, H. (2005). Positive affect as implicit motivator: On the nonconscious operation of behavioral goals. *Journal of Personality and Social Psychology*, 89(2), 129–142. <https://doi.org/10.1037/0022-3514.89.2.129>
- Custers, R., & Aarts, H. (2010). The unconscious will: How the pursuit of goals operates outside of conscious awareness. *Science*, 329(5987), 47–50. <https://doi.org/10.1126/science.1188595>
- Eitam, B., & Higgins, E. T. (2014). What’s in a goal? The role of motivational relevance in cognition and action. *Behavioral and Brain Sciences*, 37(2), 141–142. <https://doi.org/10.1017/S0140525X13002008>
- Eitam, B., Kennedy, P. M., & Higgins, E. T. (2013). Motivation from control. *Experimental Brain Research*, 229, 475–484. <https://doi.org/10.1007/s00221-012-3370-7>
- Förster, J., Higgins, E. T., & Bianco, A. T. (2003). Speed/accuracy decisions in task performance: Built-in trade-off or separate strategic concerns? *Organizational Behavior*

- and *Human Decision Processes*, 90(1), 148–164. [https://doi.org/10.1016/S0749-5978\(02\)00509-5](https://doi.org/10.1016/S0749-5978(02)00509-5)
- Hemed, E., Bakbani-Elkayam, S., Teodorescu, A. R., Yona, L., & Eitam, B. (2020). Evaluation of an action's effectiveness by the motor system in a dynamic environment. *Journal of Experimental Psychology: General*, 149(5), 935–948. <https://doi.org/10.1037/xge0000692>
- Higgins, E. T., & Cornwell, J. F. M. (2016). Securing foundations and advancing frontiers: Prevention and promotion effects on judgment & decision making. *Organizational Behavior and Human Decision Processes*, 136, 56–67. <https://doi.org/10.1016/j.obhdp.2016.04.005>
- Manian, N., Papadakis, A. A., Strauman, T. J., & Essex, M. J. (2006). The development of children's ideal and ought self-guides: Parenting, temperament, and individual differences in guide strength. *Journal of Personality*, 74(6), 1619–1646. <https://doi.org/10.1111/j.1467-6494.2006.00422.x>
- Toates, F. (1986). *Motivational systems*. Cambridge University Press.

It's bigger on the inside: mapping the black box of motivation

Marco Del Giudice* 

Department of Life Sciences, University of Trieste, Trieste, Italy
marco.delgiudice@units.it
<https://marcodg.net>

*Corresponding author.

doi:10.1017/S0140525X24000402, e30

Abstract

Many motivational constructs are opaque “black boxes,” and should be replaced by an explicit account of the underlying psychological mechanisms. The theory of motivational systems has begun to provide such an account. I recently contributed to this tradition with a general architecture of motivation, which connects “energization” and “direction” through the goal-setting activity of emotions, and serves as an evolutionary grounded map of motivational processes.

In the target article, Murayama and Jach aim at the opacity of “higher-order” motivational constructs such as needs for competence, relatedness, and autonomy, and contend that “such high-level motivation is a subjective construal or emergent property of underlying mental computational processes which determine behavior.” Their point is well taken, and a useful reminder that (just like psychometrically derived personality traits) constructs such as needs and motives should be temporary placeholders for the operations of yet-to-be-identified psychological mechanisms.

My goal in this commentary is to bring some good news: there is an entire tradition of research on *motivational systems*, rooted in ethology and flourished within psychology thanks to the work of John Bowlby (1982) and others, that has been peering into the black box for decades with very interesting results (e.g., Kenrick, Griskevicius, Neuberg, and Schaller, 2010). I recently contributed to this tradition with an architectural account of the mechanisms involved in motivation and their interplay (Fig. 1; Del Giudice, 2023a, 2023b, 2024). The *General Architecture of Motivation* (GAM) clarifies the respective roles of different kinds of mechanisms, and seamlessly connects the two main functions of motivation – the “energization” and

“direction” of behavior – through the goal-setting activity of emotions (see below).

Motivational systems are not conceptualized as amorphous “internal variables,” but as specialized control systems that regulate behavior in fitness-critical domains such as mating, attachment, affiliation, caregiving, social status, as well as physical safety and exploration (see Del Giudice, 2024; Kenrick et al., 2010, 2022). They are cognitively impenetrable but experience- and context-sensitive, are typically regulated by feedback processes (though feedforward, anticipatory regulation is likely to be important for at least some of them), and orchestrate the onset of specific emotions when they are activated (or deactivated) by cues that signal domain-specific threats and opportunities. Motivational systems are amenable to computational modeling, as demonstrated by the various simulations of the attachment system proposed over the years (see Cittern and Edalat, 2014; Petters and Beaudoin, 2017; Schneider, 2001). They energize and orient the individual in the pursuit of evolved goals, from more “basic” to more complex, such as obtaining protection and security, learning about the environment, defending and enhancing one's status, or caring for one's offspring and kin (see Del Giudice, 2024). Their neurobiological substrates include functionally specialized “hubs” that collect and integrate cues relevant to a particular domain to orchestrate behavior and physiology on a broad scale (for a striking example, see the work on parenting circuits by Kohl and colleagues [2018; Kohl, 2020; Kohl and Dulac, 2018]).

What this approach does not explain is how individuals pursue *instrumental goals* – the explicitly represented, hierarchically organized goals that guide moment-to-moment actions throughout daily life (typically associated with the direction of behavior), and that are linked only indirectly to the unrepresented, innate goals embodied by motivational systems (associated with the energization of behavior). Historically, these two kinds of goals have been addressed by different, largely non-overlapping research communities. The GAM integrates them with the inclusion of a “programmable,” general purpose control system tasked with managing hierarchies of goals (each with its own importance/urgency, abstraction, and location in time); pursuing currently active goals by generating appropriate sub-goals and monitoring their success or failure; and sending concrete *actionable goals* to downstream systems for action selection and motor control. This *Instrumental Goal Pursuit System* (IGPS) is the natural computational substrate for “higher-order” motivations related to competence and mastery, which are not well accounted for by classic models of motivational systems.

But how can motivational systems regulate behavior, if moment-to-moment instrumental goals are under the control of the IGPS? One of the key insights of the GAM is that motivational systems control behavior indirectly by activating *emotions*, which in turn provide the IGPS with urgent, abstract goals and/or “stop signals” that instruct the IGPS to suspend or terminate currently active goals. The idea that emotions generate abstract, high-priority goals for the individual (e.g., “avoiding danger” in the case of fear; “cleansing oneself” in the case of disgust; “reaching proximity to the caregiver” in the case of attachment anxiety) is consistent with motivational theories of emotions, which speak of *relational goals* (Scarantino, 2014) and *emotivational goals* (Roseman, 2011). The IGPS evaluates these goals according to their importance/urgency (likely based on the intensity of the corresponding emotions) and their compatibility with the existing goal structure; as a result, the goal hierarchy may be rearranged to

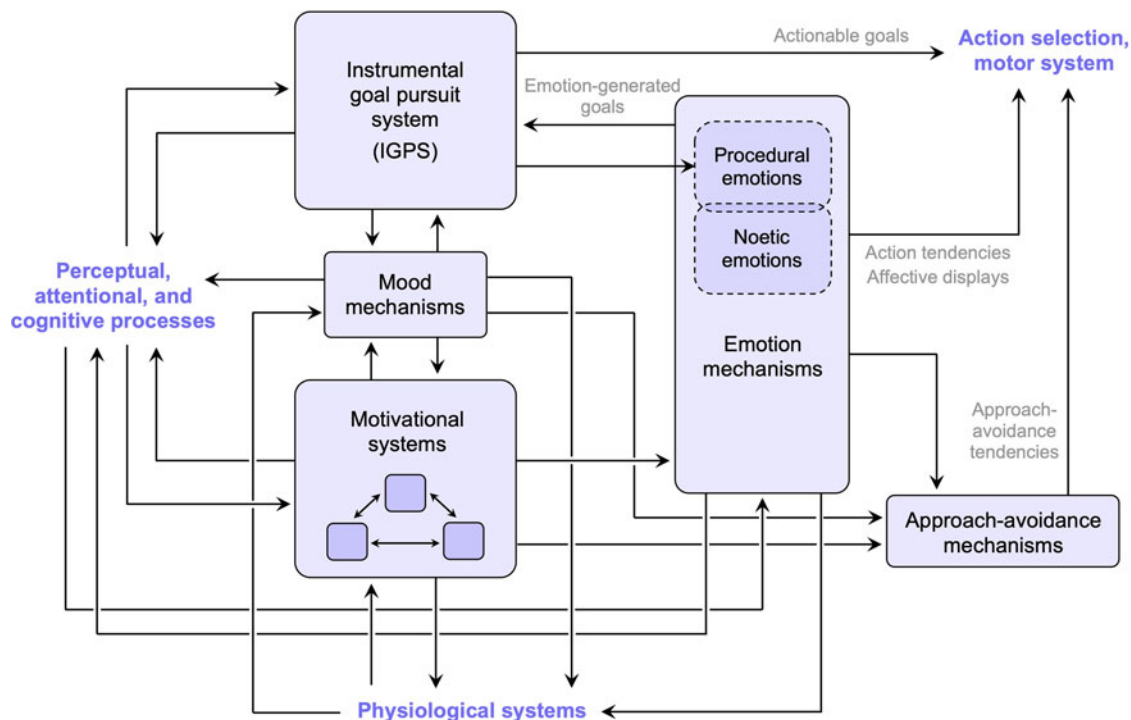


Figure 1 (Del Giudice). Schematic diagram of the general architecture of motivation (GAM). Reproduced with permission from Del Giudice (2023a).

include the new emotion-generated goals, derive concrete sub-goals, etc. In this way, emotions bridge the gap between qualitatively different kinds of goals, and serve as the “glue” that binds together multiple control mechanisms into a coordinated whole.

At the same time, the pursuit of instrumental goals by the IGPS is regulated by a set of *procedural emotions* that signal success and failure across domains and help regulate the allocation of the individual’s effort – emotions like frustration, satisfaction, disappointment, and anxious indecision in response to unresolved conflicts between goals. This means that the control of goal-directed behavior can be represented by two nested loops, an outer loop managed by motivational systems and an inner loop managed by the IGPS. Crucially, both loops involve emotions, further underscoring the fact that energization and direction are not separate but intermixed functions. (A related implication is that goal pursuit *always* has an affective component, even in the case of instrumental goals, although there are differences in the specific kinds of emotions involved.)

There are other aspects of the GAM than I cannot discuss in this brief commentary, including a theory of moods as higher-order coordination mechanisms and the conceptual tools to describe individual differences in motivation (and personality) as differences in the operating parameters of motivational systems (e.g., activation and deactivation thresholds) and the IGPS (e.g., depth of the goal hierarchy, rigidity of goal priorities, persistence of striving in the face of failure, stringency of criteria for determining success). The latter are especially relevant in light of Murayama and Jach’s call for “a theory of mental computational processes that explicitly addresses how intra-individual processes translates into long-term development” of stable individual differences (see Del Giudice, 2023a, 2023b). In short, I believe that this framework dovetails perfectly with the renovation project advocated in the target article, and offers researchers an evolutionarily

grounded map of what used to be the inscrutable black box of motivation.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. The author declares none.

References

- Bowlby, J. (1982). *Attachment and Loss: Vol. I: Attachment* (revised ed.). Basic Books.
- Cittern, D., & Edalat, A. (2014). An arousal-based neural model of infant attachment. In *2014 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)* (pp. 57–64). IEEE.
- Del Giudice, M. (2023a). A general motivational architecture for human and animal personality. *Neuroscience and Biobehavioral Reviews* 144, 104967.
- Del Giudice, M. (2023b). Motivation, emotion, and personality: Steps to an evolutionary synthesis. *PsyArXiv*, 1–24. doi: 10.31234/osf.io/3ry7b
- Del Giudice, M. (2024). The motivational architecture of emotions. In L. Al-Shawaf & T. K. Shackelford (Eds.), *The Oxford handbook of evolution and the emotions* (pp. 99–132). Oxford University Press.
- Kenrick, D. T., Griskevicius, V., Neuberg, S. L., & Schaller, M. (2010). Renovating the pyramid of needs: Contemporary extensions built upon ancient foundations. *Perspectives on Psychological Science* 5, 292–314.
- Kenrick, D. T., & Lundberg-Kenrick, D. E. (2022). *Solving modern problems with a stone-age brain: Human evolution and the seven fundamental motives*. American Psychological Association.
- Kohl, J. (2020). Parenting – a paradigm for investigating the neural circuit basis of behavior. *Current Opinion in Neurobiology* 60, 84–91.
- Kohl, J., Babayan, B. M., Rubinstein, N. D., Autry, A. E., Marin-Rodriguez, B., Kapoor, V., ... Dulac, C. (2018). Functional circuit architecture underlying parental behaviour. *Nature* 556, 326–331.
- Kohl, J., & Dulac, C. (2018). Neural control of parental behaviors. *Current Opinion in Neurobiology* 49, 116–122.
- Petters, D., & Beaudoin, L. (2017). Attachment modelling: From observations to scenarios to designs. In P. Erdi, S. Bhattacharya, & A. L. Cochran (Eds.), *Computational neurology and psychiatry* (pp. 227–271). Springer.
- Roseman, I. (2011). Emotional behaviors, emotivational goals, emotion strategies: Multiple levels of organization integrate variable and consistent responses. *Emotion Review* 3, 434–444.

Scarantino, A. (2014). The motivational theory of emotions. In J. D'Arms & D. Jacobson (Eds.), *Moral psychology and human agency: Philosophical essays on the science of ethics* (pp. 156–185). Oxford University Press.

Schneider, M. E. (2001). System theory of motivational development. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (pp. 10120–10125). Elsevier.

Human motivation is organized hierarchically, from proximal (means) to ultimate (ends)

Edgar Dubourg*, Valérien Chambon and Nicolas Baumard

Département d'études cognitives, Institut Jean Nicod, Ecole normale supérieure, Université PSL, EHESS, CNRS, Paris, France

edgar.dubourg@gmail.com

valerian.chambon@ens.psl.eu

nicolas.baumard@gmail.com

<https://www.edgardubourg.fr>

<https://nicolasbaumards.org>

<https://sites.google.com/site/chambonvalerian/home>

*Corresponding author.

doi:10.1017/S0140525X24000542, e31

Abstract

Murayama and Jach raise a key problem in behavioral sciences, to which we suggest evolutionary science can provide a solution. We emphasize the role of adaptive mechanisms in shaping behavior and argue for the integration of hierarchical theories of goal-directed cognition and behavioral flexibility, in order to unravel the motivations behind actions that, in themselves, seem disconnected from adaptive goals.

We fully agree with Murayama and Jach's advocacy for a better characterization of the *mental computational processes* underlying motivated behavior. Their article rightly highlights the limitations of high-level motivational constructs and the necessity of opening the black boxes within which these constructs operate.

Evolutionary psychology has long endeavored to decode the functional aspects of what might initially appear to be mental black boxes. This approach conceptualizes motivations not as abstract high-level constructs, but as adaptive mechanisms shaped by evolutionary pressures to regulate behavior (Tooby, Cosmides, Sell, Lieberman, & Sznycer, 2008). This approach offers a clear solution to the black-box problem: The study of *input-output specifications* (proximate level) in a way that is consistent with *design-function fit* (ultimate level).

This approach has, in our view, already clarified the concept of motivation by dissecting the evolutionary functions behind specific motivations (Al-Shawaf, 2024; Del Giudice, 2023), and by introducing the concept of regulatory variables (i.e., cognitive parameters that allow value computation; Sznycer, 2022). To take the example given by Tooby et al. (2008), to explain hunger, it is not enough to say that humans approach food. This black box can be unpacked by studying the variables which, in the case of

hunger, are calculated and valued by the human mind (e.g., calorie density, package size, search time).

What evolutionary psychologists have done is precisely to unveil the input, regulatory variables, and output of many other motivations (Al-Shawaf, & Shackelford, 2024), such as the motivations not to be socially devalued (shame; Sznycer et al., 2018), to bargain (anger; Sell & Sznycer, 2024), to pair-bond (love; Fletcher, Simpson, Campbell, & Overall, 2015), to respect one's duties (morality; André, Debove, Fitouchi, & Baumard, 2022), or to avoid predators (fear; Öhman & Mineka, 2001). This framework brings such high-level motivations closer to basic motivational constructs such as hunger or thirst.

How do these innate high-level motivational systems and associated regulatory variables initiate the concrete, context-dependent actions of organisms? An answer is to be found in cognitive theories of goal-oriented cognition, whose developments in philosophy of action (Pacherie, 2008), evolutionary biology (Del Giudice, 2023), developmental psychology (Goddu & Gopnik, 2024), and comparative psychology (Tomasello, 2022) have all emphasized its hierarchical nature. Our view is that adaptive motivations are superordinate goals that shape and prioritize lower-level instrumental goals, with a cascading effect on the selection of immediate tasks and the execution of motor actions. This suggestion is consistent with an observation often made in the field of goal hierarchies, namely that higher-order goals determine the motivational value of lower-order goals (Carver & Scheier, 1982; Diefendorff & Seaton, 2015; Höchli, Brügger, & Messner, 2018).

Now, what about actions that could not have possibly been the original target of such evolved motivations? What about, for example, filling a form to apply for a job? This action does not seem to have been initiated by an evolved motivation, as administrative forms are very recent inventions. Here we want to raise a case for behavioral flexibility (Tomasello, 2022). As flexible causal agents (Goddu & Gopnik, 2024; Kelso, 2016), humans can invent new associations between their own actions and goals at multiple levels of the hierarchy (Chu, Tenenbaum, & Schulz, 2024; see “instrumental learning” in Tomasello, 2022). As a matter of fact, humans have specific motivational systems to reward such adaptive rearrangements of the goal hierarchy, through the practice of novel action–outcome associations without consequence (i.e., play; Pellis, Pellis, Pelletier, & Leca, 2019) or even simulated (Tooby & Cosmides, 2001).

We hypothesize that these new associations between actions and goals, whether experienced, observed, played, or simulated, are rewarded not by a general reward function, but by the evolved motivational systems themselves. This constraint is fully compatible with the idea that this type of learning is open-ended (i.e., it is possible to learn an almost infinite number of new action–outcome associations; Sigaud et al., 2024). The proximate means are open-ended, but the ultimate ends are highly constrained and limited in number. As Tomasello (2022) puts it, the means for achieving adaptive goals are left to the individual's discretion, since these means always depend on the context. In other words, we propose that *open-ended instrumental goals are means to a limited number of adaptive goals* (Baumard, Fitouchi, André, Nettle, & Scott-Philipps, 2024). Without these higher-order, adaptive goals, there would be no sense of fulfillment or effectiveness for lower-level, instrumental goals (Singh, 2022; Tomasello, 2022).

As an illustration, writing this commentary could be said to be the direct outcome of one or more evolved motivations (even if the activity itself is clearly evolutionarily novel), such as (1) the motivation to appear competent (i.e., pride; Sznycer et al., 2017), (2)

the motivation to learn new knowledge that makes a difference (curiosity; Goddu & Gopnik, 2024; Murayama, 2022), or (3) the motivation to reciprocate (i.e., for the payment we receive as public workers; André et al., 2022; Trivers, 1971). Specifically, these motivations have evolved to reinforce the value of new actions the result of which leads to (1) an increase in perceived competence, (2) the generation of new difference-making information, and (3) reciprocal cooperation, each of which is associated with specific regulatory variables. Behavioral flexibility is the key solution to this problem: Our minds can connect the action of writing this commentary to low-level goals (e.g., re-reading some papers, writing a draft) and up to the higher-level adaptive goals that make these instrumental goals ultimately motivating.

In closing, we want to emphasize two key points. First, behavioral flexibility is by no means specific to humans: It can be found in mammals and even reptiles (Wilkinson & Huber, 2012). As always, the difference between humans and non-human animals is a matter of degree. Second, adaptive motivations need not be conscious: There is no evolutionary reason why the ultimate functions of motivational systems should be explicit or accessible to introspection, as long as they can regulate the learning and implementation of concrete chains of actions that fulfill adaptive goals. As a matter of fact, one of the recurring problems of evolutionary psychology as a field is that these adaptive motivations are often profoundly counter-intuitive.

Acknowledgments. The authors thank Valentin Thouzeau.

Financial support. FrontCog funding (ANR-17-EURE-0017).

Competing interest. None.

References

- Al-Shawaf, L. (2024). Levels of analysis and explanatory progress in psychology: Integrating frameworks from biology and cognitive science for a more comprehensive science of the mind. *Psychological Review*. <https://sites.google.com/site/chambonvalerian/homehttps://doi.org/10.1037/rev0000459>
- Al-Shawaf, L., & Shackelford, T. K. (2024). *The Oxford handbook of evolution and the emotions*. Oxford University Press.
- André, J.-B., Debove, S., Fitouchi, L., & Baumard, N. (2022). Moral cognition as a Nash product maximizer: An evolutionary contractualist account of morality. [Preprint]. *PsyArXiv*. <https://doi.org/10.31234/osf.io/2hngu>
- Baumard, N., Fitouchi, L., André, J.-B., Nettle, D., & Scott-Philipps, T. (2024). The gene's-eye view of culture. *To Appear in the Handbook of Evolutionary Psychology*. <https://ecoevovx.org/repository/view/6303/>
- Carver, C. S., & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality-social, clinical, and health psychology. *Psychological Bulletin*, 92(1), 111–135. <https://doi.org/10.1037/0033-2909.92.1.111>
- Chu, J., Tenenbaum, J. B., & Schulz, L. E. (2024). In praise of folly: Flexible goals and human cognition. *Trends in Cognitive Sciences*, 28(7), 628–642. <https://doi.org/10.1016/j.tics.2024.03.006>
- Del Giudice, M. (2023). Motivation, emotion, and personality: Steps to an evolutionary synthesis. <https://doi.org/10.31234/osf.io/3ry7b>
- Diefendorff, J. M., & Seaton, G. A. (2015). Work motivation. In *International encyclopedia of the social & behavioral sciences* (pp. 680–686). Elsevier. <https://doi.org/10.1016/B978-0-08-097086-8.22036-9>
- Fletcher, G. J. O., Simpson, J. A., Campbell, L., & Overall, N. C. (2015). Pair-bonding, romantic love, and evolution: The curious case of homo sapiens. *Perspectives on Psychological Science*, 10(1), 20–36. <https://doi.org/10.1177/1745691614561683>
- Goddu, M. K., & Gopnik, A. (2024). The development of human causal learning and reasoning. *Nature Reviews Psychology*, 3(5), 319–339. <https://doi.org/10.1038/s44159-024-00300-5>
- Höchli, B., Brügger, A., & Messner, C. (2018). How focusing on superordinate goals motivates broad, long-term goal pursuit: A theoretical perspective. *Frontiers in Psychology*, 9, 1879. <https://doi.org/10.3389/fpsyg.2018.01879>
- Kelso, J. A. S. (2016). On the self-organizing origins of agency. *Trends in Cognitive Sciences*, 20(7), 490–499. <https://doi.org/10.1016/j.tics.2016.04.004>
- Murayama, K. (2022). A reward-learning framework of knowledge acquisition: An integrated account of curiosity, interest, and intrinsic-extrinsic rewards. *Psychological Review*, 129(1), 175–198. <https://doi.org/10.1037/rev0000349>
- Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, 108(3), 483–522. <https://doi.org/10.1037/0033-295X.108.3.483>
- Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, 107(1), 179–217. <https://doi.org/10.1016/j.cognition.2007.09.003>
- Pellis, S. M., Pellis, V. C., Pelletier, A., & Leca, J.-B. (2019). Is play a behavior system, and, if so, what kind? *Behavioural Processes*, 160, 1–9. <https://doi.org/10.1016/j.beproc.2018.12.011>
- Sell, A., & Sznycer, D. (2024). The recalibrational theory: Anger as a bargaining emotion. In L. Al-Shawaf & T. K. Shackelford (Eds.), *The Oxford handbook of evolution and the emotions* (pp. 135–144). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780197544754.013.6>
- Sigaud, O., Baldassarre, G., Colas, C., Doncieux, S., Duro, R., Perrin-Gilbert, N., & Santucci, V. G. (2024). A definition of open-ended learning problems for goal-conditioned agents (arXiv:2311.00344). arXiv. <http://arxiv.org/abs/2311.00344>
- Singh, M. (2022). Subjective selection and the evolution of complex culture. *Evolutionary Anthropology: Issues, News, and Reviews*, 31(6), 266–280. <https://doi.org/10.1002/evan.21948>
- Szzyner, D. (2022). Value computation in humans. *Evolution and Human Behavior*, 43(5), 367–380. <https://doi.org/10.1016/j.evolhumbehav.2022.06.002>
- Szzyner, D., Al-Shawaf, L., Bereby-Meyer, Y., Curry, O. S., De Smet, D., Ermer, E., ... Tooby, J. (2017). Cross-cultural regularities in the cognitive architecture of pride. *Proceedings of the National Academy of Sciences*, 114(8), 1874–1879. <https://doi.org/10.1073/pnas.1614389114>
- Szzyner, D., Xygalatas, D., Agey, E., Alami, S., An, X.-F., Ananyeva, K. I., ... Tooby, J. (2018). Cross-cultural invariances in the architecture of shame. *Proceedings of the National Academy of Sciences*, 115(39), 9702–9707. <https://doi.org/10.1073/pnas.1805016115>
- Tomasello, M. (2022). *The evolution of agency: Behavioral organization from lizards to humans*. The MIT Press.
- Tooby, J., & Cosmides, L. (2001). Does beauty build adapted minds? Toward an evolutionary theory of aesthetics, fiction and the arts. *SubStance*, 30(1/2), 6. <https://doi.org/10.2307/3685502>
- Tooby, J., Cosmides, L., Sell, A., Lieberman, D., & Szzyner, D. (2008). Internal regulatory variables and the design of human motivation: A computational and evolutionary approach. In A. Elliot (Ed.), *Handbook of approach and avoidance motivation* (pp. 251–271). Lawrence Erlbaum Associates. <https://doi.org/10.4324/9780203888148.ch15>
- Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46(1), 35–57. <https://doi.org/10.1086/406755>
- Wilkinson, A., & Huber, L. (2012). Cold-blooded cognition: Reptilian cognitive abilities. In J. Vonk & T. K. Shackelford (Eds.), *The Oxford handbook of comparative evolutionary psychology* (pp. 129–143). Oxford University Press. <https://psycnet.apa.org/record/2012-04454-008>

Expectancy value theory's contribution to unpacking the black box of motivation

Jacquelynne S. Eccles^a and Allan Wigfield^{b,c,*}

^aSchool of Education, University of California Irvine, Irvine, CA, USA;

^bDepartment of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, USA and ^cDepartment of Education and Brain and Motivation Research Institute (bMRI), Korea University, Seoul, Korea
Jseccles@uci.edu
awigfiel@umd.edu

*Corresponding author.

doi:10.1017/S0140525X24000475, e32

Abstract

Although in basic agreement with Murayama and Jach's call for greater attention to the black boxes underlying motivated behavior, we provide examples of our published suggestions regarding how subjective task value (and ability self-concepts) "gets into people's knowledge structures." We suggest additional mental computational processes to investigate and call for a developmental and situated individual differences approach to this work.

We agree with Murayama and Jach's (M&J) claims that unpacking the mental black box underlying motivated behavior and paying greater attention to the emerging properties of key latent constructs is critical precisely "because ... it would provide a new landscape of understanding these concepts." We think we have spent considerable time over the last 50 years on these goals. Contrary to M&J's claim that "Expectancy-value theory does not specify how value is incorporated and represented into the existing knowledge structure," our model of subjective task value and ability self-concepts/expectations for success is based on: (1) Social cognitive theories including attribution theory of achievement behavior and its links to mentally stored emotional responses and causal inferences regarding the nature of the self and reality; self-schema and identity-development theories; and the cognitive integration over time of one's experiences in the creation of concepts and cognitive algorithms; as well as cognitive developmental theories potentially underlying age-related changes in the kinds of cognitive algorithms one might use to make "wise" behavioral choices; and (2) several of the major classic reward theories including classical, operant, and observational learning, and mental rewards for successful enactment of new behaviors. We have articulated a wide array of mental processes linked to the formation and storage of value-related information that are linked to both reward systems and the self-systems in the medial prefrontal and posterior parietal cortex. Like M&J, we agree that seeking information is likely to be rewarded by the brain; we also believe that using cognitive algorithms to make short- and long-term behavioral choices is rewarding.

It is important to note that Eccles trained under Weiner – an achievement motivational theorist who explicitly replaced needs-based theories of motivation in favor of more mentally informed information processing and achievement-related problem solving. Building on his socio-emotional cognitive perspective on motivation, we focus specifically on those mental calculations related to the formation of relative expectancy and subjective task value beliefs and then the use of this information in making behavioral decisions. To the extent that such decisions have a direct influence on survival and reproduction, it is likely that the mental processes associated with such choices have developed and are inherently rewarding. The existence of specific brain regions linked to self-related mental processes supports this hypothesis.

Furthermore, although we do not fully understand these mental processes, we have specifically proposed some, such as those listed above, and have discussed the importance of uncovering others such as the varying algorithms used by people of different histories and ages in aggregating various pieces of information relative to subjective task value across different situations (i.e., Eccles & Wigfield, 2020). Right now, we are interested in individual differences in the mental computational mechanisms underlying the various subjective task values people place on the array of options available in making high-level behavioral plans related to long-term motivated behavioral choices, for example, occupational, and recreational choices. We also have discussed how different kinds of comparison processes, for example, temporal, social comparative, or dimensional across different activity domains, influence the development of both expectancies and subjective task values (Wigfield, Eccles, & Möller, 2020).

Moving beyond our own theory we turn to some broader comments regarding M&J's article. We were somewhat surprised by their choice of need for competence as the exemplar construct

on which to focus. There are constellations of "higher-level" constructs having to do with competence and competence perception, for example, expectancies, self-efficacy, self-concept of ability, and perceived control. Obviously, need for competence is related, but Ryan and Deci's (2017) notion that there is a *need* to be competent is different than having *beliefs* about one's competences and engaging in behaviors to improve competence. Needs imply something more basic and fundamental, whereas beliefs are formed through experience, and therefore more likely to be determined by the mental computational processes proposed by M&J.

We also think it is important to expand on M&J's focus on "reward-learning models of information seeking behavior." What other reward systems might be operative in motivated behavior? We mentioned several earlier. Additionally, we think it is important to consider the processes linked to seeking specific kinds of information. It seems likely that the mental processes involved in forming expectancies for success are different than those involved in forming one's subjective task values. Both the kinds of information used and the associated affective experiences pertaining to each of these constructs are likely to differ. It is also likely that the kinds of information used to make specific behavioral decisions vary across different contextual settings and developmental ages. Thus, we need to explore the nature of such processes behind different "higher-order" motivation constructs in different contextual settings and time frames.

Finally, it will be important to understand the development of the relevant mental computation processes. Cognitive maturation is undoubtedly involved in the developmental changes we see across the life span in the responses people make to information related in making "wise" motivated-behavioral choices. For example, young children (up to about 8) continue to have very high success expectations despite repeated failure in lab studies (e.g., Parsons & Ruble, 1977); after 8 children show a very linear decline in expectations for success following such failure information. Similarly, children's theory of mind changes over childhood. Finally, according to Baltes' (1997) SOC model, people of different ages should weight opportunities to "select, optimize, and compensate" differently as they manage their motivated behaviors differently due to age-related changes in their cognitive and physical resources. Research is needed to understand age differences in such motivated behavioral choices.

We thank M&J for helping us to think harder about unpacking the black box.

Financial support. None.

Competing interests. None.

References

- Baltes, P. B. (1997). On the incomplete architecture of human ontogeny: Selection, optimization, and compensation as foundation of developmental theory. *American Psychologist*, 52(4), 366–380. <https://doi.org/10.1037/0003-066X.52.4.366>
- Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, 61. <https://doi.org/10.1016/j.cedpsych.2020.101859>
- Parsons, J. E., & Ruble, D. N. (1977). The development of achievement-related expectancies. *Child Development*, 48(3), 1075–1079.
- Ryan, R. M., & Deci, E. (2017). *Self-determination theory: Basic human needs in motivation, development, and wellness*. Guilford.
- Wigfield, A., Eccles, J. S., & Möller, J. (2020). How dimensional comparisons help to understand linkages between expectancies, values, performance, and choice. *Educational Psychology Review*, 32, 657–680.

Needed: Clear definition and hierarchical integration of motivation constructs

Andrew J. Elliot^{a*} and Nicolas Sommet^b

^aUniversity of Rochester, Rochester, NY, USA and ^bSwiss National Centre of Competence in Research LIVES, University of Lausanne, Lausanne, Switzerland
Andrew.Elliot@rochester.edu
Nicolas.Sommet@unil.ch
<https://www.sas.rochester.edu/psy/research/apav/index.html>
<https://www.nicolassommet.com/>

*Corresponding author.

doi:10.1017/S0140525X24000487, e33

Abstract

Murayama and Jach offer a thoughtful and timely critique of motivation constructs. We largely concur with their basic premises, but offer additional input and clarification regarding the importance of carefully considering the energization and direction components of motivation, and fully attending to the hierarchical aspect of motivation rather than prioritizing particular levels of analysis.

The target article by Murayama and Jach (M&J) does exactly what one wants such a piece to do – it makes one step back and rethink the broad assumptions and premises that guide one's work. Such a meta-level piece, especially one so thoughtfully and even provocatively articulated, can be extremely helpful in clarifying one's perspective and laying out guidelines and priorities for how to proceed. The piece fits in a long and admirable tradition of internal criticism of the way motivation constructs are conceptualized and utilized to explain behavior (for other noteworthy examples, see Bindra, 1959; Bolles, 1978; Brown, 1961; Cofer, 1972; Kantor, 1942; Kleinginna & Kleinginna, 1981). Although this form of critique is not new per se, we believe it is important and needed at present in motivation science.

The authors' critique centers on high-level motivation constructs, and two core premises of the critique are that (1) high-level motivation constructs are not clearly defined and conceptualized, and (2) the nature of the influence of high-level motivation constructs on behavior is not well-understood. That is, "what they are" and "how they work" are not clearly specified (p. 25). Regarding the first premise, we fully agree and simply offer elaboration. Good conceptualization, in the motivation literature and beyond, requires clear construct definition and clear articulation of the construct's functional role. Slippage on the conceptual (as well as operational) front produces what we sometimes see in the motivation literature – jingle-jangle fallacies and resultant inconsistent empirical literatures that are difficult to summarize and interpret. We think clarity here begins with drawing an explicit and precise distinction between the two basic components of motivation – energization and direction. We define *energization* as the initial instigation of behavior that orients in a general way (the "why";

e.g., needs/motives) and *direction* as the channeling of this energization toward a specific end state (the "how"; e.g., goals/tactics; Elliot, 2023). Critically, once separated and carefully conceptualized, energization and direction must be put back together for a full and complete motivational explanation of behavior. For example, neither the need for competence (energization) nor achievement goals (direction) are sufficient to account for achievement behavior; both are needed in combination. We have called this combined construct a "goal complex" (Elliot & Thrash, 2001) – the integrated representation of a focal goal (direction) and the reason behind the goal that prompted its adoption (energization) – and a growing body of research attests to the theoretical and empirical utility of this concept (for reviews, see Liem & Senko, 2023; Sommet, Elliot, & Sheldon, 2021).

Regarding the second premise, M&J call for a focus on lower-level processes in motivational analyses of behavior, arguing that constructs at this level are optimally suited to explain behavior. Here we both agree and disagree. We agree that comprehensive motivational explanations must include lower-level processes. Indeed, we view motivation as decidedly hierarchical (Elliot, 2006; see also Cacioppo & Berntson, 1994; Carver & Scheier, 2001; Gallistel, 1982), encompassing myriad constructs at many levels across the neuraxis. In the main, the authors seem to argue that lower-level processes are understudied and such processes are needed to complement the explanatory value provided by high-level constructs (e.g., "No level of understanding should be dismissed as 'wrong' [i.e. one level of explanation should not be replaced with a lower-level explanation], because they just explain the behavior for different purposes," p. 21); we fully agree. At other points, however, they seem to argue for lower-level processes as a replacement for high-level constructs (e.g., "It is the mental computational processes, not the motivation constructs themselves, that are necessary to understand human behavior," p. 15); here we disagree. We see the value of explicating lower-level processes, but not at the expense of high-level constructs. *Many* levels of explanation needed – from rudimentary exteroceptive reflexes to subcortical computations to cortical appraisals to the emergent high-level constructs that the authors critique. We think all of these levels of analysis are worthy of study and provide added value to motivational explanations of behavior. Selecting the lower-level as the key to explanation seems to run the risk of reductionism (Sheldon, 2004), if not infinite regress (Boden, 1972). This level of analysis issue is reminiscent of Tolman's (1932) critique of behaviorism's sole focus on the "molecular" and his advocating for an additional, "molar," level of analysis that incorporates purpose; he described this molar level as "emergent" and as being "more than and different from the sum of" its molecular parts (p. 7). Categorizing high-level constructs as emergent or even psychologically constructed (as M&J do) does not necessarily mean they are epiphenomenal (as they seem to imply). Such high-level constructs can have an important, independent influence on behavior, often via evocation or recruitment of lower-level processes (e.g., for empirical evidence regarding the aforementioned goal complexes, see Sommet & Elliot, 2017). In short, each level of analysis has value, and the optimal level at any given time depends on one's overarching aim (e.g., to acquire a deeper understanding of underlying processes, to discover when and how to intervene, to explain a phenomenon to laypeople, etc.).

Psychological constructs are scientific tools used to describe, categorize, and organize collections of observations, and theories represent integrated combinations of such constructs. Human beings and the behavior they emit are extraordinarily intricate and complex; good theories must, by necessity, match this intricacy and complexity (i.e., be “level adequate,” see Berridge, 2004, p. 17). In motivation science, we believe the best theories will be those that are unabashedly hierarchical, thoroughly inclusive, and deeply integrative (Elliot & Sommet, 2023), comprised of both well-defined high-level motivation constructs and multiple levels of lower-level psychological processes. We concur with M&J that a major focus of research in motivation science moving forward needs to be on lower-level processes (of many sorts at many levels). We hasten to add that this work will only advance motivation science to the degree that the findings from it are carefully and thoughtfully integrated into the existing work on high-level constructs.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sector.

Competing interests. None.

References

- Berridge, K. C. (2004). Motivation concepts in behavioral neuroscience. *Physiology & Behavior*, 81, 179–209.
- Bindra, D. (1959). *Motivation*. The Ronald Press Co.
- Boden, M. A. (1972). *Purposive explanation in psychology*. Harvard University Press.
- Bolles, R. C. (1978). Whatever happened to motivation? *Educational Psychologist*, 13, 1–13.
- Brown, J. S. (1961). *The motivation of behavior*. McGraw-Hill.
- Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin*, 115, 401–423.
- Carver, C. S., & Scheier, M. F. (2001). *On the self-regulation of behavior*. Cambridge University Press.
- Cofer, C. N. (1972). *Motivation and emotion*. Scott, Foresman, & Co.
- Elliot, A. J. (2006). The hierarchical model of approach-avoidance motivation. *Motivation and Emotion*, 30, 111–116.
- Elliot, A. J. (2023). Energization and direction are both essential parts of motivation. In M. Bong, J. M. Reeve, & S. I. Kim (Eds.), *Motivation science: Controversies and insights* (pp. 10–14). Oxford University Press.
- Elliot, A. J., & Sommet, N. (2023). Integration in the achievement motivation literature and the hierarchical model of achievement motivation. *Educational Psychology Review*, 35, 77.
- Elliot, A. J., & Thrash, T. M. (2001). Achievement goals and the hierarchical model of achievement motivation. *Educational Psychology Review*, 13, 139–156.
- Gallistel, C. R. (1982). *The organization of action: A new synthesis*. Psychology Press.
- Kantor, J. R. (1942). Toward a scientific analysis of motivation. *The Psychological Record*, 5, 225–275.
- Kleinginna Jr, P. R., & Kleinginna, A. M. (1981). A categorized list of motivation definitions, with a suggestion for a consensual definition. *Motivation and Emotion*, 5, 263–291.
- Liem, G.A.D., & Senko, C. (2023). Goal complexes: A new approach to studying the coordination, consequences, and social contexts of pursuing multiple goals. *Educational Psychology Review*, 34, 2167–2195.
- Sheldon, K. M. (2004). *Optimal human being: An integrated multi-level perspective*. Psychology Press.
- Sommet, N., & Elliot, A. J. (2017). Achievement goals, reasons for goal pursuit, and achievement goal complexes as predictors of beneficial outcomes: Is the influence of goals reducible to reasons? *Journal of Educational Psychology*, 109, 1141–1162.
- Sommet, N., Elliot, A. J., & Sheldon, K. M. (2021). The “what” and “why” of achievement motivation: Conceptualization, operationalization, and consequences of self-determination derived achievement goal complexes. In R. Robbins & O. John (Eds.), *Handbook of personality psychology: Theory and research* (4th ed., pp. 104–121). Guilford Press.
- Tolman, E. C. (1932). *Purposive behavior in animals and men*. University of California Press.

Almost, but not quite there: Research into the emergence of higher-order motivated behavior should fully embrace the dynamic systems approach

Christophe Gernigon^{a*}, Rémi Altamore^a,
Robin R. Vallacher^b, Paul L. C. van Geert^c and
Ruud J. R. Den Hartigh^c

^aEuroMov Digital Health in Motion, Université de Montpellier, IMT Mines Alès, Montpellier, France; ^bDepartment of Psychology, Florida Atlantic University, Boca Raton, FL, USA and ^cDepartment of Psychology, University of Groningen, TS, Groningen, The Netherlands
christophe.gernigon@umontpellier.fr
remi.altamore@umontpellier.fr
vallacher@fau.edu
p.l.c.van.geert@rug.nl
j.r.den.hartigh@rug.nl
<https://www.researchgate.net/profile/Christophe-Gernigon>
<https://fr.linkedin.com/in/r%C3%A9mi-altamore-6788111a8>
<https://psy.fau.edu/people/vallacher.php>
<https://www.paulvangeert.nl>
<https://www.rug.nl/staff/j.r.den.hartigh>

*Corresponding author.

doi:10.1017/S0140525X24000384, e34

Abstract

Murayama and Jach rightfully aim to conceptualize motivation as an emergent property of a dynamic system of interacting elements. However, they do not embrace the ontological and paradigmatic constraints of the dynamic systems approach. They therefore miss the very process of emergence and how it can be formally modeled and tested by specific types of computer simulation.

We concur with Murayama and Jach (M&J) that motivation is an emergent property of a collective dynamic system of interacting elements. However, the principles and the model these authors develop do not fall within the ontological and paradigmatic framework of dynamic systems and emergent phenomena. This ambiguity needs to be clarified as it has important implications for how motivational processes can and should be conceptualized and investigated.

By considering that their model lends itself to testing its various parts, as well as the classic antecedents and outcomes of motivation, M&J seem to conceptualize motivational processes as driven by *component-dominant* dynamics, that is, as decomposable into isolable parts (e.g., Hausman & Woodward, 1999). However, according to the dynamic systems perspective, psychological phenomena are patterns that emerge from *interaction-dominant* dynamics (Den Hartigh, Cox, & van Geert, 2017; Van Orden, Holden, & Turvey, 2003; Wallot & Kelty-Stephen, 2018) that are *non-decomposable* and *non-isolable* (Bechtel & Richardson, 2010). Thus, the emergent properties of a dynamic

system cannot be deduced from the properties of its components, just as the fluidity, viscosity, and transparency of water cannot be deduced from the aggregate properties of oxygen and hydrogen (Bunge, 1977).

Moreover, the principle of M&J's reward-learning model is a reinforcement loop consisting of a causal chain that unfolds among its components, with very few interactions to modulate the causal relationships. The system is self-boosting in that an interest-based engagement promotes a positive feedback loop that sustains long-lasting information-seeking behavior. Strictly speaking, this behavior cannot be considered emergent, since it can be predicted on the basis of the value of its immediate determinants in the causal chain. Unlike a causal chain, even in loop form, a dynamic system involves complex interactions among components, which lead – through a process of self-organization – to the emergence of a global behavior pattern for the system. This pattern tends to stabilize by contributing to the formation of an attractor landscape, which in turn constrains the states of the system's components and their interactions, and so forth (e.g., Kelso, 1995). This attractor dynamics implies non-proportionality between variations of the system's components and those of the emergent behavior, which results in nonlinear dynamics of that behavior. This nonlinear dynamics can account for some well-documented typical motivational patterns, such as persistence of effort despite negative experiences, oscillation between motivated and unmotivated states, and abrupt shifts in motivation following a tiny variation in one of its putative determinants (Carver & Scheier, 1998; Gernigon, Vallacher, Nowak, & Conroy, 2015; López-Pernas & Saqr, 2024). In its current form, M&J's feedback loop could neither explain nor simulate such dynamics.

How motivational processes are conceptualized has, in turn, important consequences for how they can be investigated. M&J consider mathematical formulations of mental computational processes to be useful, but neither necessary nor sufficient. Surprisingly, however, they do refer to van der Maas et al. (2006) as a case example, whose dynamic model of the emergence of general intelligence is typically based on mathematical formalizations of interactions – governed by evolution rules – among many components that evolve over time. As the example of van der Maas et al. illustrates, patterns emerging from dynamic systems typically follow evolution rules that can be expressed mathematically, generally with logistic equations, or with coupled differential or difference equations (e.g., Guastello & Liebovitch, 2009). Whether they are parameterized directly or indirectly via software interfaces, these equations model the underlying processes and can thus account for the emergence of higher-level motivational patterns. Unlike M&J's reinforcement loop, specific computer simulation methods, such as dynamic networks, dynamic field models, agent-based models, cellular automata, and genetic algorithms, are designed to implement the self-organization processes that lead to the emergence of particular psychological phenomena (Gernigon, Den Hartigh, Vallacher, & van Geert, 2024). Hence, these methods make it possible to observe and test how these phenomena identifiable at a higher-order level emerge from rules modeled at a lower-order level (Nowak, 2004; Smaldino, 2023; Vallacher, Read, & Nowak, 2017).

A reinforcement loop and the modeling of self-organization processes are also substantially different in terms of the type of prediction that can be tested. M&J contend that their “*precise process model*” can help researchers make more fine-grained predictions about how different types of assessments or manipulations result in different outcomes. This contention reflects an

interventionist or *manipulative* conception of causality that is currently prevalent (e.g., Hausman & Woodward, 1999), but which yields poorly reproducible results (Open Science Collaboration, 2015) and which cannot account for a *process* causality based on principles of emergence (e.g., Gernigon et al., 2024; van Geert & de Ruiter, 2022). More realistically, though perhaps frustratingly for some researchers, the complexity of emergence processes and their idiosyncratic nature casts doubt on any promise of fine-grained prediction in terms of interventionist causality. The predictions permitted by process causality, on the other hand, concern typical statistical signatures of complexity and (nonlinear) dynamics that can be detected at the idiosyncratic level by specific time series analyses, such as Recurrence Quantification Analysis and Detrended Fluctuation Analysis, of both simulation and ecological data. In addition, dynamic models of individual cases are expected to yield, at population level, comparable descriptive statistics between simulation and ecological data. Ultimately, for the sake of convergent validity, a dynamic model of motivation should account for both the consistencies and inconsistencies of the field's literature (e.g., Gernigon et al., 2015, 2024). In doing so, we may come one step closer to understanding how intra-individual processes can give rise to different motivational trajectories.

To conclude, we agree with M&J that an explanatory account of motivation requires a focus on the lower-level mechanisms that give rise to higher-order motivated behavior. The lens of dynamic systems is best suited to providing this focus, as it captures the complexity of motivational processes better than traditional approaches. However, embracing this perspective is a paradigmatic choice that is conceptually and methodologically constraining. While still to be made, this promising choice is within reach of M&J and other motivational researchers.

Funding statement. This commentary is not linked to any specifically funded research program.

Competing interests. None.

References

- Bechtel, W., & Richardson, R. C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT Press.
- Bunge, M. (1977). Emergence and the mind. *Neuroscience*, 2(4), 501–509. [https://doi.org/10.1016/0306-4522\(77\)90047-1](https://doi.org/10.1016/0306-4522(77)90047-1)
- Carver, C. S., & Scheier, M. F. (1998). *On the self-regulation of behavior*. Cambridge University Press.
- Den Hartigh, R. J., Cox, R. F., & van Geert, P. L. (2017). Complex versus complicated models of cognition. In L. Magnani & T. Bertolotti (Eds.), *Springer handbook of model-based science* (pp. 657–669). Springer.
- Gernigon, C., Den Hartigh, R. J. R., Vallacher, R. R., & van Geert, P. L. C. (2024). How the complexity of psychological processes reframes the issue of reproducibility in psychological science. *Perspectives on Psychological Science*, 19(6), 952–977. <https://doi.org/10.1177/17456916231187324>
- Gernigon, C., Vallacher, R. R., Nowak, A., & Conroy, D. E. C. (2015). Rethinking approach and avoidance in achievement contexts: The perspective of dynamical systems. *Review of General Psychology*, 19(4), 443–457. <https://doi.org/10.1037/gpr0000055>
- Guastello, S. J., & Liebovitch, L. S. (2009). Introduction to nonlinear dynamics and complexity. In S. J. Guastello, M. Koopmans & D. Pincus (Eds.), *Chaos and complexity in psychology: The theory of nonlinear dynamical systems* (pp. 1–40). Cambridge University Press.
- Hausman, D. M., & Woodward, J. (1999). Independence, invariance and the causal Markov condition. *The British Journal for the Philosophy of Science*, 50(4), 521–583. <https://doi.org/10.1093/bjps/50.4.521>
- Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. MIT Press.
- López-Pernas, S., & Saqr, M. (2024). How the dynamics of engagement explain the momentum of achievement and the inertia of disengagement: A complex systems

- theory approach. *Computers in Human Behavior*, 153, 108126. <https://doi.org/10.1016/j.chb.2023.108126>
- Nowak, A. (2004). Dynamical minimalism: Why less is more in psychology. *Personality and Social Psychology Review*, 8(2), 183–192. https://doi.org/10.1207/s15327957pspr0802_12
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. <https://doi.org/10.1126/science.aac4716>
- Smaldino, P. E. (2023). *Modeling social behavior: Mathematical and agent-based models of social dynamics and cultural evolution*. Princeton University Press.
- Vallacher, R. R., Read, S. J., & Nowak, A. (Eds.) (2017). *Computational social psychology*. Routledge/Taylor & Francis.
- van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, 113(4), 842–861. <https://doi.org/10.1037/0033-295X.113.4.842>
- van Geert, P., & de Ruiter, N. (Eds.) (2022). *Toward a process approach in psychology: Stepping into Heraclitus' River*. Cambridge University Press.
- Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, 132(3), 331–350. <https://doi.org/10.1037/0096-3445.132.3.331>
- Wallot, S., & Kelty-Stephen, D. G. (2018). Interaction-dominant causation in mind and brain, and its implication for questions of generalization and replication. *Minds and Machines*, 28(2), 353–374. <https://doi.org/10.1007/s11023-017-9455-0>

Don't throw motivation out with the black box: The value of a good theory revisited

Jutta Heckhausen^{a*} and Falko Rheinberg^b

^aDepartment of Psychological Science, University of California, Irvine, CA, USA and ^bDepartment Psychologie, Universität Potsdam, Potsdam, Germany
heckhaus@uci.edu
motivation@psych.uni-potsdam.de
<https://faculty.sites.uci.edu/heckhausen/>

*Corresponding author.

doi:10.1017/S0140525X24000347, e35

Abstract

Murayama and Jach claim that current motivational constructs do not specify causal processes (*black-box problem*) and that *mental computational processes* solve this problem. We argue, process-focused research requires theoretical frameworks addressing situational variations, individual differences, and their interaction. Classic achievement motivation theory provides comprehensive models with empirically measurable process-related constructs and predictions. Recent developments build on this, addressing motivation, action, and their socio-cultural and lifespan context. Theory-free *mental computational processes* cannot do any of that.

A focus on process is laudable, but...

All psychological research should aim at identifying the specific processes that underly human behavior. However, “*mental computational processes*” championed by Murayama and Jach (M&J) lack a theoretical framework that is comprised of propositions about how these processes emerge, what influences them, and how they affect experience and behavior. Neuronal activity and the “*mental computations*” they can compose are not

functional in and of themselves. They emerge as functional processes in the context of an individual's interactions with the environment.

Nothing is as process-focused as a good theory

Good theories comprise theoretical constructs that are clearly defined and operationalized, and that have a specific function in the theoretical model. An excellent example of a theory comprising elaborated processes that are integrated into a person-by-situation interaction framework following Lewin's (1946) axiom, is achievement motivation theory (Atkinson, 1957; Heckhausen, 1977; McClelland, Atkinson, Clark, & Lowell, 1953) with its risk-taking model in terms of anticipated self-evaluation after success and failure. This theory solves the alleged *black-box problem* by making distinct predictions about the effects of situational incentives on behavior, such as task choice, effort, and task performance for individuals with predominant hope for success versus predominant fear of failure. Achievement situations with intermediate challenge are the most informative about one's own competence and therefore the most attractive for success seekers and the most threatening for failure avoiders (Brunstein & Heckhausen, 2018). Their task choices, effort investment, emotional responses, causal attributions, and performance vary accordingly along predicted patterns. Such person-by-situation interactions are based on the theoretically proposed and empirically tested process variables of goal setting, causal attribution, and self-evaluation (Brunstein & Heckhausen, 2018; Heckhausen, 1977; Stiensmeyer-Pelster & Heckhausen, 2018). These processes develop as mutually stabilizing individual differences in cognitive and emotional predispositions during goal-directed interactions with the environment in childhood (Heckhausen, 1975; Heckhausen & Heckhausen, 2018). Further compelling evidence for achievement motivation theory comes from the effective use of its models for intervention and change of motive dispositions and associated biases in cognition, emotion, and behavior (Rheinberg & Engeser, 2010).

Hence, even several decades ago, motivation was not at all a black box. Instead, theoretical models of specific and functionally interrelated processes interfacing person and situation were developed, operationalized, and thus made accessible for empirical research. More recent developments in achievement motivation research associated with Eccles and Wigfield's Situated Expectancy-Value Theory (Eccles & Wigfield, 2002, 2020) differentiate among different value components as a function of social developmental context and individual preference. This allows for more differentiated and developmentally as well as culturally informed modeling of achievement-motivated behavior.

“Mental computation” or functionality of motivational mindsets

Mental computation and cognitive processes are important, but the energizing and directional function of motivation critically relies on experienced and anticipated affect and its change (e.g., anticipated enjoyment of activity, anticipated pride about own competence). According to the Rubicon model of action-phases, cognition and affect work hand in hand, but have shifting priorities depending on whether the individual is trying to determine the optimal goal (deliberative phase of action cycle) or pursuing an already chosen goal (implemental phase) (Gollwitzer, 1990; Heckhausen & Gollwitzer, 1987). During the pre-decision phase, objectivity and

breadth of mental computation is essential and thus a deliberative mindset is activated, whereas during the post-decision phases, biased information processing following an implementation mindset shields the intended action from distractions and conflicting tendencies (Achtziger & Gollwitzer, 2018).

Motivation is a product of evolution and development

M&J address the evolution of motivation but ignore individual differences. Motivational mental processes are a product of phylogenetically evolved pre-adaptations and ontogenetically developed strategies and patterns in a specific individual. They are only in part universal outcomes of behavioral evolution at a phylogenetic scale (e.g., classical and operant conditioning and mastery striving, Heckhausen, 2000). The non-universal motivational processes reflect the ontogeny of individuals and their unique experiences with affective change (self and other-regulated), bearing a strong influence of preverbal exposure to affect–change patterns in the context of the parent–child dyad (Heckhausen & Heckhausen, 2018; Kuhl & Völker, 1998). Formative developmental conditions are associated with transitions, for instance the transition from intra-individual to inter-individual reference frames when starting school with its dominant evaluative framework of social comparison. As individuals become more self-regulated in adolescence and adulthood, their potential to follow the motivational predispositions acquired earlier increases exponentially and further stabilizes them. In this process, individuals become increasingly nimble in regulating and optimizing their own motivation in a given situational set of incentives, based on their extensive experience with self-regulation of motivation (Rheinberg & Engeser, 2010).

Levels of analyses: From micro to macro to micro

As M&J demand, we do need to get closer to the actual processes that are at play in specific situations. However, we should not throw out the conceptual and empirical progress we already made, just because we cannot yet successfully capture the more molecular processes. On the other hand, having such higher order constructs should not spare us from digging deeper into the more molecular processes. In our discovery endeavor to identify and link up processes at different levels of analysis, different approaches can complement each other. An example is the combination of situated expectancy-value theory (Eccles & Wigfield, 2002, 2020) which addresses goal choice, and motivational theory of lifespan development (Heckhausen, Wrosch, & Schulz, 2010, 2019), which addresses how long-term goals are pursued and changed across the life course (von Keyserlingk, Rubach, Lee, Eccles, & Heckhausen, 2022).

Conclusion

We can best examine the specific processes that bring about motivational experiences and behavior, if we have a theoretical framework that is guided by the function of the phenomena to be explained. A computational analysis per se solves no problems. The good news is, there is no *black-box problem* and no need to reinvent the wheel. We can follow the guidance of motivational scholars in the 1970s and 80s who pioneered and developed achievement motivation research. Uncovering person by situation interactions puts us motivational psychologists ahead of many in more unidimensional fields such as personality or social psychology.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

References

- Achtziger, A., & Gollwitzer, P. M. (2018). Motivation and volition in the course of action. In J. Heckhausen & H. Heckhausen (Eds.), *Motivation and action* (pp. 485–528). Springer.
- Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review*, 64, 359–372.
- Brunstein, J. C., & Heckhausen, H. (2018). Achievement motivation. In J. Heckhausen & H. Heckhausen (Eds.), *Motivation and action* (pp. 221–304). Springer.
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53, 109–132.
- Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, 61, 101859.
- Gollwitzer, P. M. (1990). Action phases and mind-sets. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of social behavior* (Vol. 2, pp. 53–92). Guilford.
- Heckhausen, H. (1975). Fear of failure as a self-reinforcing motive system. In I. G. Sarason & C. Spielberger (Eds.), *Stress and anxiety* (Vol. II, pp. 117–128). Hemisphere.
- Heckhausen, H. (1977). Achievement motivation and its constructs: A cognitive model. *Motivation and Emotion*, 1, 283–329.
- Heckhausen, J. (2000). Evolutionary perspectives on human motivation. *American Behavioral Scientist*, 43, 1015–1029.
- Heckhausen, H., & Gollwitzer, P. M. (1987). Thought contents and cognitive functioning in motivational and volitional states of mind. *Motivation and Emotion*, 11, 101–120.
- Heckhausen, J., & Heckhausen, H. (2018). Development of motivation. In J. Heckhausen & H. Heckhausen (Eds.), *Motivation and action* (pp. 679–743). Springer.
- Heckhausen, J., Wrosch, C., & Schulz, R. (2010). A motivational theory of life-span development. *Psychological Review*, 117, 32–60.
- Heckhausen, J., Wrosch, C., & Schulz, R. (2019). Agency and motivation in adulthood and old age. *Annual Review of Psychology*, 70, 191–217.
- Kuhl, J., & Völker, S. (1998). Entwicklung und Persönlichkeit [Development and personality]. In H. Keller (Ed.), *Lehrbuch der entwicklungspsychologie* (pp. 207–240). Huber.
- Lewin, K. (1946). Behavior and development as a function of the total situation. In L. Carmichael (Ed.), *Manual of child psychology* (pp. 791–844). John Wiley & Sons.
- McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). *The achievement motive*. Appleton-Century-Crofts.
- Rheinberg, F., & Engeser, S. (2010). Motive training and motivational competence. In O. C. Schultheiss & J. C. Brunstein (Eds.), *Implicit motives* (pp. 510–548). Oxford University Press.
- Stiensmeyer-Pelster, J., & Heckhausen, H. (2018). Causal attribution of behavior and achievement. In J. Heckhausen & H. Heckhausen (Eds.), *Motivation and action* (pp. 623–678). Springer.
- von Keyserlingk, L., Rubach, C., Lee, H. R., Eccles, J. S., & Heckhausen, J. (2022). College students' motivational beliefs and use of goal-oriented control strategies: Integrating two theories of motivated behavior. *Motivation and Emotion*, 46, 601–620.

Higher-order motivational constructs as personal-level fictions: A solution in search of a problem

Marko Jurjako* 

Faculty of Humanities and Social Sciences, Department of Philosophy and Division of Cognitive Sciences, University of Rijeka, Rijeka, Croatia
mjurjako@ffri.uniri.hr
<https://cogsci.uniri.hr/people/marko-jurjako/>

*Corresponding author.

doi:10.1017/S0140525X24000360, e36

Abstract

I argue that Murayama and Jach's claim that higher-order motivational constructs face the "black-box" problem is misconceived because it doesn't clearly distinguish between personal and subpersonal explanations. To solve it they propose interpreting motivations as causal effects of mental computational processes. I suggest that their solution might be more compellingly presented as providing a fictionalist perspective on some personal-level constructs.

In the target article, Murayama and Jach (M&J) argue that explanations of complex behaviors using higher-order motivational constructs, such as the need for competence or the achievement motive, face the "black-box" problem: While these constructs may explain why people act in certain situations, they don't explain the essence of motivation and *how* motivational processes arise. Moreover, such explanations supposedly face the *motivational homunculus* problem, where explaining one motivational construct in terms of another risks circular reasoning or infinite regress. To avoid this, the authors suggest that higher-order motivational constructs are psychologically constructed from more fundamental mental computational processes. I will argue that these objections are misguided because they fail to properly distinguish between levels of psychological explanation, particularly those involving personal and subpersonal explanations.

Higher-order motivational constructs elucidate behavior at the personal level, involving the references to whole persons and their psychological states (e.g., Dänzer, 2023; Dennett, 1969; Drayson, 2012). As pointed out by several philosophers, at this level explanations run out sooner than one might think (see, Dennett, 1969, p. 95). Here explanations often conclude when we understand the reasons behind people's actions, that is, when we can rationalize their behavior (see, e.g., Queloz, 2017). If we want to know what enables reasons to motivate action, we should transition to the subpersonal level, where behavioral, cognitive, and motivational processes are explained by underlying biological, physiological, and computational processes (Drayson, 2012).

Crucially, however, searching for subpersonal explanations (and opening the "black box") doesn't imply that the initial personal explanation was incomplete. Instead, we are seeking a different kind of explanation for the same thing. Personal explanations are supposed to illuminate what the person was doing and *why*, that is, in the light of what reasons they were acting, whereas subpersonal explanations primarily offer insights into *how* these processes were implemented at the computational (i.e., algorithmic) and/or physical levels (Wilkinson, 2014; see, also Marr, 1982). M&J come close to recognizing this when they assert that their "perspective indicates that high-level motivation constructs reflect higher-level explanations whereas mental computational processes represent lower-level explanations" and add that "[n]o level of understanding should be dismissed as 'wrong' (...), because they just explain the behavior for different purposes (...)" (p. 22). However, in formulating the objection to motivational constructs in terms of the "black-box" problem and associating it with the homunculus fallacy, they seem to overlook the fact that this perspective doesn't require that higher-level motivational constructs come prespecified with internal properties that might connect them with physical, biological, or computational variables. This is a job for subpersonal explanation (Drayson, 2012).

So, what problem do M&J manage to solve by providing a subpersonal computational solution to the black-box problem? To answer this question, it should be noted that there are different views on how to understand the relation between the personal and the subpersonal. In the philosophy of psychology, this is labeled the interface problem (Bermúdez, 2005, p. 35).

I believe that M&J's solution could be understood as providing a specific perspective on the interface problem. For instance, traditional functionalism posits that constructs at the personal level have causal roles in generating actions. According to such approaches, advancements in cognitive science should enable us to identify computational procedures implementing these roles and eventually discern their implementation in diverse brain processes (for discussion, see, Colombo & Fabry, 2021; Jurjako, 2022). In contrast, M&J propose that personal-level motivational constructs do not play these "energizing" causal roles in generating action; rather, they are the effects of underlying computational processes that actually cause action. This perspective is in line with insights from social psychology suggesting that the reasons people give for their actions, instead of reflecting the actual causal factors giving rise to actions, often serve as *post-hoc* rationalizations that in a particular cultural context may make sense of observed behaviors (Nisbett & Ross, 1980; see, also Cushman, 2020).

However, there is tension in M&J's account. On the one hand, they claim that "high-level motivation constructs are an emergent property of underlying mental computational processes" (p. 12). On the other hand, they claim that such constructs are "a consequence of psychological construction" (p. 12), meaning that they result from "interpreting and categorizing the regularities that exist in behavioral patterns and subjective experiences" (p. 13). If these constructs refer to emergent properties involving behavioral and experiential regularities, then, although they presumably lack energizing causal powers, they denote something real. However, if they are constructed from subjective interpretations of behavioral regularities and these interpretations falsely attribute energizing causality to them, then these constructs lack objective reality. Thus, should we understand these constructs as emergent properties devoid of energizing causal powers, or as products of subjective interpretations?

To avoid this ambiguity, I propose to reinterpret M&J's position as endorsing a kind of fictionalism about the personal/motivation-level constructs (Toon, 2023; see, also Dennett, 2022; Tollon, 2023). Such constructs could be understood as referring to useful fictions that form parts of our narratives about what we think typically causes our actions. Being fictive in this context doesn't mean that such constructs don't play significant roles in our lives and psychological theories (see, Cushman, 2020). They certainly do, just as constructs such as the equator, the average person, and the ideal gas law play significant roles in scientific theorizing and ordinary practices, even though they provide an idealized and thus not completely accurate view of real physical systems. Similarly, we could think of higher-level motivational constructs as referring to idealizations that enable us to capture and predict behavioral regularities at a more abstract level of description (e.g., Dennett, 1989), and also shape such regularities by embodying culturally based prescriptions for desirable behavior (e.g., McGeer, 2015). This view seems to be compatible with M&J's core claim that mental computational processes more precisely capture the causal underpinnings of action. Moreover, it simultaneously resolves the ambiguity in their account and avoids the reification of motivational constructs as emergent properties that lack energizing causal powers.

Acknowledgments. I wish to thank Miguel Núñez de Prado Gordillo, Sam Wilkinson, and Luca Malatesti for their helpful comments on an earlier version of this commentary.

Financial support. This paper is an outcome of project TIPPS, funded by the Croatian Science Foundation (grant HRZZ-IP-2022-10-1788) and is also supported by the University of Rijeka (grant uniri-iskusni-human-23-14).

Competing interest. None.

References

- Bermúdez, J. L. (2005). *Philosophy of psychology: A contemporary introduction*. Routledge.
- Colombo, M., & Fabry, R. E. (2021). Underlying delusion: Predictive processing, looping effects, and the personal/sub-personal distinction. *Philosophical Psychology*, 34(6), 829–855. <https://doi.org/10.1080/09515089.2021.1914828>
- Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences*, 43, e28. <https://doi.org/10.1017/S0140525X19001730>
- Dänzer, L. (2023). The personal/subpersonal distinction revisited: Towards an explication. *Philosophy*, 98(4), 507–536. <https://doi.org/10.1017/S0031819123000220>
- Dennett, D. C. (1969). *Content and consciousness*. Routledge.
- Dennett, D. C. (1989). *The intentional stance*. MIT Press.
- Dennett, D. C. (2022). Am I a fictionalist? In T. Demeter, T. Parent, & A. Toon (Eds.), *Mental fictionalism* (pp. 352–361). Routledge.
- Drayson, Z. (2012). The uses and abuses of the personal/subpersonal distinction. *Philosophical Perspectives*, 26(1), 1–18. <https://doi.org/10.1111/phpe.12014>
- Jurjako, M. (2022). Can predictive processing explain self-deception? *Synthese*, 200(4), 303. <https://doi.org/10.1007/s11229-022-03797-6>
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.
- McGeer, V. (2015). Mind-making practices: The social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations*, 18(2), 259–281. <https://doi.org/10.1080/13869795.2015.1032331>
- Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Prentice-Hall.
- Queloz, M. (2017). Two orders of things: Wittgenstein on reasons and causes. *Philosophy*, 92(3), 369–397. <https://doi.org/10.1017/S0031819117000055>
- Tollon, F. (2023). Free will as an epistemically innocent false belief. *European Journal of Analytic Philosophy*, 19(2), (A2)1–15. <https://doi.org/10.31820/ejap.19.2.2>
- Toon, A. (2023). *Mind as metaphor: A defence of mental fictionalism*. Oxford University Press. <https://doi.org/10.1093/oso/9780198879626.001.0001>
- Wilkinson, S. (2014). Levels and kinds of explanation: Lessons from neuropsychiatry. *Frontiers in Psychology*, 5, 373. <https://doi.org/10.3389/fpsyg.2014.00373>

When unpacking the black box of motivation invites three forms of reductionism

Agnes Moors* 

Faculty of Psychology and Educational Sciences, KU Leuven – University of Leuven, Leuven, Belgium
agnes.moors@kuleuven.be
http://ppw.kuleuven.be/okp/people/Agnes_Moors/

*Corresponding author.

doi:10.1017/S0140525X24000426, e37

Abstract

In their proposal for unpacking the black box of motivation, Murayama and Jach (M&J) propose three types of reductions: From high-level to low-level motivational constructs, from motivation to cognition, and from contentful to contentless explanations. Although these reductions come with the promise of parsimony, they carry the risk of losing vital explanatory power.

The motivation literature typically distinguishes between process theories and content theories. Process theories seek to provide a mechanistic explanation of behavior, specifying a mechanism between input (stimulus) and output (behavior) that includes motivation constructs such as goals. Content theories address questions about what our fundamental goals are (e.g., autonomy, competence, and connectedness in self-determination theory; Ryan & Deci, 2017), or they apply existing process theories to a specific type of goal (e.g., achievement, affiliation, power; Murray, 1938). Unlike content theories, existing process theories have already taken steps toward opening the black box that sits between observable stimulus input and behavioral output. The critique of M&J therefore seems to be more directed toward content theories than toward process theories. Alternatively, their critique could be directed toward process theories, suggesting that these theories do not unpack their mechanisms in sufficient detail, that is, at a low enough level of analysis.

The authors go on to propose a way to do this unpacking. Despite cautioning that their exercise is not a reductionist attempt, they appear to promote three types of reductions. I will argue how each of these types poses risks for throwing away the baby with the bathwater.

A first type of reduction is a shift from high-level (abstract) motivational constructs to low-level (concrete) motivational constructs. For instance, they propose replacing the goal for “competence” with the goal for “information seeking.” The ideal seems to shift to motivational constructs that are as close as possible to the behavior itself. Thus, they argue that information-seeking behavior is caused by the goal to have information rather than by the goal for competence. They further propose that such a low-level goal is caused by a gap in this goal. Thus, the goal to have information is caused by uncertainty, that is, a gap between a current and desired level of information.

The ultimate step in this shift toward low-level constructs is the shift from more extrinsic to more intrinsic motivation. People seek information because they value the act of information seeking (or at least the immediate outcome of this act: information) rather than because this information is instrumental for reaching other goals. Such a shift seems to downplay the role of extrinsic motivation, however. There may certainly be cases in which people act for the sake of it, but much of our behavior, including information seeking, is done to reach other goals. The explanatory power of existing process theories resides in the fact that they propose a goal hierarchy in which behavioral goals that are the proximal causes of behavior can be considered as subordinate goals that have a high expectancy for reaching other, superordinate goals.

This feature is preserved in several existing process theories, including in my own goal-directed theory of behavior causation (Moors, 2022; Moors, Boddez, & De Houwer, 2017). Here, a gap between a high-level goal and a current state activates the goal to reduce this gap. This can be done by different types of behaviors (i.e., assimilation), but also by adjusting the high-level goal (i.e., accommodation) or by changing the interpretation of the current state (i.e., immunization). To illustrate, the gap between the current state and the high-level goal to become popular activates the goal to reduce this gap. One way to reduce this gap is via behavior (e.g., wearing nice clothes, making jokes; i.e., assimilation), another is to give up the goal to become popular (i.e., accommodation), and still another is to reinterpret the current state as one in which you are already popular (i.e., immunization).

It might be argued that the high-level goals proposed in content theories such as the goals for autonomy and competence are

not well chosen, but then the solution would be to come up with better ones, not to do away with high-level goals altogether. An alternative solution would be to view the goals for autonomy and competence as meta-goals that are at the service of, or assist in, the attainment of other, low-level goals. In this vein, the goal for autonomy can be considered as the goal to be allowed to *choose* one's own (low-level) goals and the goal for competence or control can be considered as the goal to *achieve* these (low-level) goals.

A second type of reduction that M&J seem to promote consists in a shift from motivation to cognition. The authors admit that in their computational process model, there are still rewards, which are representations of valued outcomes and hence motivational constructs. However, once the unpacking of the black box has arrived at its most concrete level of motivation, the authors argue that it makes little sense to continue calling this motivation. This reveals that the ideal to which they aspire is to ultimately reduce motivation to cognition. This is reminiscent of the attempt of predictive processing theory to reduce the explanation of behavior (and other phenomena such as perception and affect) to the confirmation and disconfirmation of expectations (Clark, 2013; Friston, 2009).

A third type of reduction that the authors seem to promote consists in gradually explaining away content or semantics. In standard mechanistic explanations in psychology, mechanisms between stimulus input and behavioral output are composed of representations with a content and a format (i.e., the structural parts) and operations acting on these representations (i.e., the activities or working parts) (Bechtel, 2008). The authors' ideal seems to be to reduce mechanistic explanations that consist of both contentful representations and operations to explanations that consist only of operations.

In conclusion, even if existing process theories of motivation have already made progress in unpacking the black box, it may be argued that there is always room for further unpacking at lower levels of analysis. Whether this shift to lower levels of analysis should include the rejection of high-level goals, the reduction of motivation to cognition, and the evolution toward content-less mechanistic explanations is open to debate, as is the level of analysis that will prove to be most fruitful for predicting and regulating behavior (Karoly, 1999).

Financial support. Preparation of this article was supported by grant C14/23/062 of the Research Fund of KU Leuven.

Competing interests. None.

References

- Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Routledge. <https://doi.org/10.2478/disp-2008-0012>
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204. <https://doi.org/10.1017/S0140525X12000477>
- Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13(7), 293–301. <https://doi.org/10.1016/j.tics.2009.04.005>
- Karoly, P. (1999). A goal systems-self-regulatory perspective on personality, psychopathology, and change. *Review of General Psychology*, 3(4), 264–291. <https://doi.org/10.1037/1089-2680.3.4.264>
- Moors, A. (2022). *Demystifying emotions: A typology of theories in psychology and philosophy*. Cambridge University Press. <https://doi.org/10.1017/9781107588882>
- Moors, A., Boddez, Y., & De Houwer, J. (2017). The power of goal-directed processes in the causation of emotional and other actions. *Emotion Review*, 9, 310–318. <https://doi.org/10.1177/1754073916669595>
- Murray, H. A. (1938). *Explorations in personality*. Oxford University Press.
- Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford. <https://doi.org/10.1521/978.14625/28806>

Motivational whack-a-mole: Foundational boxes cannot be unpacked

Ezgi Ozgan  and Jedediah W. P. Allen* 

Department of Psychology, Bilkent University, Ankara, Turkey
ezgi.ozgan@bilkent.edu.tr
jallen@bilkent.edu.tr

*Corresponding author.

doi:10.1017/S0140525X24000517, e38

Abstract

The proposed “black-box” problem and its solution are drawn from the same substance-oriented framework. This framework's assumptions have consequences that re-create the black-box problem at a foundational level. Specifically, Murayama and Jach's solution fails to explain novel behavior that emerges through an organism's development. A process-oriented theoretical shift provides an ontological explanation for emergent behavior and eliminates the black-box problem altogether.

Murayama and Jach (M&J) critically evaluate psychology's explanatory use of high-level motivations as causes of complex behavior (i.e., the “black-box” problem). Their critique presents a valuable case for the need to focus on the concrete dynamics and causal relations of cognitive processes. The critical side of their argument helps clarify how descriptions of motivation interpreted as causal explanations are only apparent; however, their positive proposal simultaneously risks a continuation of the illusion through a new iteration of the problem. That is, their proposed solution seems to be built on the same theoretical foundations as the problem, and this might just exchange one large black box for several smaller ones.

M&J point to an *explanatory* illusion that there are properties being attributed to motivation that it does not possess. Instead, they propose to eliminate those properties from motivation altogether (Witherington, 2014). In turn, motivation is interpreted as a label for the composition of the causal relations amongst lower-level constructs that do the actual work of energizing (and explaining) behavior. Thus, motivation is merely a *container* with no causal (or explanatory) power over its contents and associated behavior – the motivation itself cannot explain behavior beyond its contents (Witherington, 2011). Consequently, M&J render motivation as an epiphenomenal outcome of the causal structures amongst lower-level constructs. Although they use the term “emergent” to describe motivation, it does not seem to be *ontological* emergence – because their definition of motivation lacks novel qualities that can causally affect the relations amongst the lower-level constructs (i.e., no downward causation; Witherington, 2011). There is degree-wise merit in M&J's solution since their proposed constructs – compared to motivations – have a more direct causal relation to the unfolding changes observed throughout a behavior. However, their solution assumes a foundational version of the same black-box problem – because the constructs and motivations are “just” foundational atoms at different scales, and the problems at the motivation level are

inherited by the lower levels. In other words, M&J's attempt to resolve the *explanatory* illusion of motivation results in a "solution illusion."

The key to their commitment to *foundationalism* and *epiphenomenal emergence* is an underlying substance-oriented framework (Bickhard, 2006). This framework is evident when they describe the "energization" aspect of motivation at the center of the *explanatory* illusion. This involves dependence on an external (or internal) impetus to initiate behavior (Bickhard, 2003). This assumption aligns with the inertness of atomism and sets the stage for M&J to assert two corollaries of a substance-oriented framework: *Compositional emergence* and *instrumental relations*. Atomism establishes the foundationalism part of their solution, where constructs are considered to possess greater explanatory power than any emergent qualities that motivation might offer (Allen & Bickhard, 2022). However, it is the two corollaries that ultimately make the solution to the black-box problem more apparent than real.

First, the *compositional ontology* of atoms underlies the lack of ontological emergence (Bickhard, 2006; Witherington, 2011). For M&J, due to the foundational atoms' surface togetherness, novelty is structural. This compositional quality is evident in the assertion that the causal relations of constructs can be a substitute for motivation (i.e., the [re]arrangement of foundational parts is the reason for the manifested difference among high-level concepts; Seibt, 2009). Based on this assumption, motivation does not possess any emergent qualities that could explain behavior beyond the foundational constructs; the entirety of the explanation takes place at the foundational level. Second, the assumption of *instrumental relations* is about the missing ontological ties amongst foundational parts. The foundation is the only existential reality, and no real phenomena could emerge through the relations of the parts (an implication of compositional ontology, Allen & Bickhard, 2022; Seibt, 2009). Thus, foundational parts can continue their existence in isolation and any relations they possess are strictly instrumental. This corollary is evident in the re-interpretation of the high-level motivation "need for competence" as a reward-learning model. Reward-learning models are developed within a computationalist approach – which explicitly assumes *instrumental relations* to govern the communication amongst parts to explain how behavior unfolds (Bickhard & Terveen, 1995).

Based on these substance-oriented corollaries, M&J adopt an epiphenomenal re-interpretation of motivation that precludes their solution from enabling qualitative emergence. That is, any computational substitute for motivation does not have the flexibility to explain novelty in behavior – that is, the presuppositions underlying the proposed causal relations cannot undergo development through constructive emergence (Allen & Bickhard, 2022). The explanatory power of any proposed model is constrained by the qualities of the foundational constructs – both ontological and relational. However, ontological emergence and constitutive relations (i.e., where the relation is intrinsic to the organization's existence and necessary for the continuity of the "parts") are necessary at the higher-level phenomena to explain any behavior that develops through learning (e.g., developmental changes in social understanding). Therefore, the limitation of the proposed solution to explain novel behavior leads to a *solution* illusion, a foundation-level black box that can never be unpacked (Allen & Bickhard, 2013).

The alternative solution to the black-box problem is a paradigm shift away from a substance-oriented framework. This would eliminate the black-box problem at all levels of behavioral complexity by replacing the atom-ontology of physical phenomena with process (van Geert & de Ruiter, 2022). Since processes

are inherently active, they must interact with each other (Bickhard, 2003). Thus, the "need" to energize behavior is an illusion since living organisms constantly behave due to their existence as processes. Consequently, motivation is not a *trigger* for behavior but a *selection* amongst potential ways of reorganizing the lower-level processes that constitute the organism. This definition of motivation is similar to the "direction" aspect of motivation mentioned by M&J. Motivation is part of the flow of control in terms of how the organism changes its processes through which the behavior itself emerges. In this sense, a process-oriented framework offers a form of explanation that renders both the black-box issue and its purported solution superfluous.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interest. None.

References

- Allen, J. W., & Bickhard, M. H. (2013). Stepping off the pendulum: Why only an action-based approach can transcend the nativist–empiricist debate. *Cognitive Development*, 28(2), 96–133. <http://doi.org/10.1016/j.cogdev.2013.01.002>
- Allen, J. W., & Bickhard, M. H. (2022). Emergent constructivism: Theoretical and methodological considerations. *Human Development*, 66(4–5), 276–294. <http://doi.org/10.1159/000526220>
- Bickhard, M. H. (2003). An integration of motivation and cognition. In L. Smith, C. Rogers, & P. Tomlinson (Eds.), *Development and motivation: Joint perspectives* (pp. 41–56). British Psychological Society, Monograph Series II. <http://doi.org/10.53841/bpsmono.2003.cat529.5>
- Bickhard, M. H. (2006). Developmental normativity and normative development. In L. Smith, & J. Voneche (Eds.), *Norms in human development* (pp. 57–76). Cambridge University Press. <http://doi.org/10.1017/CBO9780511489778.003>
- Bickhard, M. H., & Terveen, L. (1995). *Foundational issues in artificial intelligence and cognitive science: Impasse and solution*. Elsevier Scientific.
- Seibt, J. (2009). Forms of emergent interaction in general process theory. *Synthese*, 166(3), 479–512. <http://doi.org/10.1007/s11229-008-9373-z>
- van Geert, P., & de Ruiter, N. (Eds.). (2022). *Toward a process approach in psychology: Stepping into Heraclitus' river*. Cambridge University Press. <http://doi.org/10.1017/9781108859189>
- Witherington, D. C. (2011). Taking emergence seriously: The centrality of circular causality for dynamic systems approaches to development. *Human Development*, 54(2), 66–92. <http://doi.org/10.1159/000326814>
- Witherington, D. C. (2014). Self-organization and explanatory pluralism: Avoiding the snares of reductionism in developmental science. *Research in Human Development*, 11(1), 22–36. <http://doi.org/10.1080/15427609.2014.874763>

Connecting theories of personality dynamics and mental computational processes

Juliette L. Ratchford^{a,b*}  and

Eranda Jayawickreme^{a,b} 

^aDepartment of Psychology, Wake Forest University, Winston-Salem, NC, USA and ^bProgram for Leadership and Character, Wake Forest University, Winston-Salem, NC, USA

julietteratchford@gmail.com

jayawide@wfu.edu

<https://jayawide.sites.wfu.edu/>

*Corresponding author.

doi:10.1017/S0140525X24000499, e39

Abstract

Whole Trait Theory (and other dynamic theories of personality) can illuminate the process by which motivational states become traits. Mental computational processes constitute part of the explanatory mechanisms that drive trait manifestations. Empirical work on Whole Trait Theory may inform future research directions on mental computational processes.

Murayama and Jach (M&J) suggest that one important direction for future work is that “to understand motivational states and traits, we need to develop a theory of mental computational processes that explicitly addresses how intra-individual processes translates into long-term development” (target article, sect. 5, para. 6). One route that may illuminate this future direction is to engage recent accounts of personality dynamics. Here we propose Whole Trait Theory (Fleeson & Jayawickreme, 2021a; Jayawickreme, Fleeson, Beck, Baumert, & Adler, 2021) as a guiding framework. Whole Trait Theory suggests that personality traits are composed of two parts: (1) An explanatory aspect which captures social-cognitive mechanisms that cause trait manifestations and (2) a descriptive aspect which captures manifestations of personality traits and states in daily life (Fleeson & Jayawickreme, 2021a). Whole Trait Theory suggests that the explanatory component is important to understanding meaningful intra-individual variations in behavior. In our view, this aligns with M&J’s account; specifically, we argue that mental computational processes represent an example of the explanatory processes underlying trait enactments. Given these parallels, empirical work engaging Whole Trait Theory and other dynamic accounts of personality may inform future directions of research on mental computational processes.

One particularly illuminating line of research using Whole Trait Theory has examined the relation between momentary goal pursuits and trait manifestations. In a 10-day experience sampling study of extraversion, participants reported both their state extraversion and the extent to which they were trying to accomplish 18 goals related to facets of extraversion (e.g., sociable) in the last 30 minutes (McCabe & Fleeson, 2012). These momentary goal pursuits predicted nearly three quarters of the variance in state extraversion, suggesting that motivation plays a role in state manifestations of traits (McCabe & Fleeson, 2012). In subsequent research, momentary goals explained nearly half the variance in extraversion and conscientiousness trait enactments (McCabe & Fleeson, 2016). Additionally, in an experiment where participants were assigned either an extraversion or conscientiousness goal to enact for the next 45 minutes, participants reported higher state levels of that respective trait (McCabe & Fleeson, 2016). These studies demonstrate a connection between motivation and trait enactment that may elucidate how intra-individual processes transition to long-term development.

The transition from intra-individual processes to long-term development may be related to the descriptive aspect of Whole Trait Theory. The descriptive aspect is represented as density distributions of states. These density distributions refer to the unique distribution formed over time of a person’s states (Fleeson & Jayawickreme, 2015). Characteristics of density distributions include the location of the distribution on a dimension (i.e., different people have different means; Fleeson & Nettle, 2008) and the width of the distribution (i.e., how much variability in intra-individual states; Fleeson & Gallagher, 2009). Distributions vary

from person to person; however, individual’s distributions are typically stable in terms of their location (r s around 0.8) from week to week (Fleeson, 2001). Additionally, as indicated in the studies by McCabe and Fleeson (2012, 2016), density distributions are shaped by the motivations and goals of an individual, which are expressed as part of the explanatory aspect of Whole Trait Theory. Continued manifestation of trait expressions form density distributions of states.

The explanatory aspect of Whole Trait Theory includes “the set of cognitive, affective, biological, and motivational processes that influence the degree to which a person manifests the trait at any given moment” (Fleeson & Jayawickreme, 2021b, p. 99). The explanatory aspect, as its name suggests, explains the distributions of states; it explains the variations in behaviors, why people enact different trait contents at different times (Fleeson & Jayawickreme, 2021b). The manner in which people understand and react to the situation they are in leads to changes in their trait enactment and behavior; people adapt their behavior to their context. Additionally, these behaviors can be reinforced or undermined by their external environment. For example, when a person is more talkative, others may engage with them more. Additionally, state extraversion has been found to cause more state positive affect (e.g., McNiel, Lowman, & Fleeson, 2010), which may in turn reinforce extraverted behavior. There are a host of processes that undergird the explanatory aspect of Whole Trait Theory – social-cognitive, affective, biological, interpretative, temporal – but a key process is the motivational process (Fleeson & Jayawickreme, 2021b). As indicated in the goal-related research described above, momentary goal pursuits – momentary motivations – explained between half and three quarters of the variance in trait enactments (McCabe & Fleeson, 2012, 2016). The explanatory aspect is thus well-positioned to include mental computational processes as a specific subtype within the explanatory aspect of Whole Trait Theory.

Mental computational processes are described by M&J as momentary motivation manifestations. If we consider mental computational processes as part of the explanatory aspect, then we would expect that people’s mental computational processes change based on their situation and on their momentary goal pursuits. Additionally, changes in mental computational processes would lead to a shift in the distributions of motivational states. To the extent to which these changes are reinforced by a person’s environment, these changes would become more fixed and stable, perhaps accounting for more stable motivational traits (e.g., intrinsic motivation).

M&J lay the foundation for mental computational processes and their relation to motivation. Future research into mental computational processes could draw inspiration from research on Whole Trait Theory and other dynamic accounts of personality. Some avenues for future directions include investigating: (1) The density distributions of momentary motivational “states” and their relation to stable motivational traits, (2) how mental computational processes are reinforced or challenged by changes in people’s environments, (3) the situational influences on the manifestations of mental computational processes, and (4) the role of specific goal pursuits in the selection of mental computational processes. Such research into how, when, and why mental computational processes are engaged will help to further unpack the black box of motivation.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interest. None.

References

- Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, 80(6), 1011–1027. <https://doi.org/10.1037/0022-3514.80.6.1011>
- Fleeson, W., & Gallagher, P. (2009). The implications of Big Five standing for the distribution of trait manifestation in behavior: Fifteen experience-sampling studies and a meta-analysis. *Journal of Personality and Social Psychology*, 97(6), 1097–1114. <https://doi.org/10.1037/a0016786>
- Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, 56, 82–92.
- Fleeson, W., & Jayawickreme, E. (2021a). Whole trait theory puts dynamics at the core of structure. In *The handbook of personality dynamics and processes* (pp. 579–599). Academic Press.
- Fleeson, W., & Jayawickreme, E. (2021b). Whole traits: Revealing the social-cognitive mechanisms constituting personality's central variable. In *Advances in experimental social psychology* (Vol. 63, pp. 69–128). Academic Press.
- Fleeson, W., & Nofle, E. (2008). The end of the person–situation debate: An emerging synthesis in the answer to the consistency question. *Social and Personality Psychology Compass*, 2(4), 1667–1684. <https://doi.org/10.1111/j.1751-9004.2008.00122.x>
- Jayawickreme, E., Fleeson, W., Beck, E. D., Baumert, A., & Adler, J. M. (2021). Personality dynamics. *Personality Science*, 2, 1–18. <https://doi.org/10.5964/ps.6179>
- McCabe, K. O., & Fleeson, W. (2012). What is extraversion for? Integrating trait and motivational perspectives and identifying the purpose of extraversion. *Psychological Science*, 23(12), 1498–1505. <https://doi.org/10.1177/0956797612444904>
- McCabe, K. O., & Fleeson, W. (2016). Are traits useful? Explaining trait manifestations as tools in the pursuit of goals. *Journal of Personality and Social Psychology*, 110(2), 287–301. <https://doi.org/10.1037/a0039490>
- McNiel, J. M., Lowman, J. C., & Fleeson, W. (2010). The effect of state extraversion on four types of affect. *European Journal of Personality*, 24(1), 18–35. <https://doi.org/10.1002/per.738>

The role of metacognitive feelings in motivation

Rolf Reber* , Josefine Haugen† and

Liva J. Martinussen†

Department of Psychology, University of Oslo, Oslo, Norway
rolf.reber@psykologi.uio.no
l.j.a.haugen@psykologi.uio.no
l.j.martinussen@psykologi.uio.no
<https://www.sv.uio.no/psi/english/people/academic/rolf/reb/>; <https://www.sv.uio.no/psi/english/?vrtx=person-view&uid=lindajos>; <https://www.sv.uio.no/psi/english/people/external-phds/livajm/index.html>

*Corresponding author.

doi:10.1017/S0140525X24000359, e40

Abstract

Metacognitive feelings are an integral part of mental computational processes and influence the outcome of computations. We review supporting evidence on affect inherent in perceptual processes, fluency in study decisions, metacognitive feelings in aha-experiences and intuition, and affect in early phases of interest development. These findings connect to recent theories that combine metacognitive feelings with computational models.

Murayama and Jach (M&J) include subjective experiences as input and output of mental computational mechanisms in their sophisticated model of motivation (their Figs. 1 and 2). However, the role of subjective experiences remains underspecified. Our commentary highlights the central role of metacognitive feelings in mental computational processes. We argue that subjective feelings are not just input and output but an integral part of mental computational processes, and they influence the outcome of computations. Motivation could be seen in terms of mental computational processes continuously monitored and regulated by metacognitive feelings. Metacognitive feelings are subjective experiences that inform an individual about cognitive processes and include affect, subjective certainty, and fluency, which is the ease with which a mental process is executed (Efklides, 2006; Schwarz & Clore, 2007). Such feelings provide continuous information about an individual's interaction with the environment. Positive affect, high fluency, and high certainty indicate that the interaction with the environment proceeds smoothly; thus, individuals do not need to change their behavior. By contrast negative affect, low fluency, and low certainty indicate a problem whose solution needs new information or behavior change.

The following evidence supports the notion that mental computational processes are interwoven with metacognitive feelings. First, affect is inherent in perceptual processes. Brief exposure to a coherent but non-recognizable outline yields more positive affect than a non-coherent outline, as measured by EMG activity of Zygomaticus Major, the “smiling muscle” (Erle, Reber, & Topolinski, 2017; Flavell, Tipper, & Over, 2018; Topolinski, Erle, & Reber, 2015). Cognitive involvement in these evaluations was minimal, which means that affect may occur before the mental computational processes. Similarly, success of later perceptual processes, such as identification of objects or solving mental rotation tasks, yields positive affect (Lindell, Zickfeld, & Reber, 2022). These studies show that there never is “no affect,” in contrast to models where affect only serves as input or output. The interesting question will be how the ongoing dynamics of affect in perception that guide attention influence mental computational processes.

Second, feelings of knowing and judgments of learning guide study decisions (Brooks, Yang, & Köhler, 2021; Hanczakowski, Zawadzka, & Cockcroft-McKay, 2014; Metcalfe & Finn, 2008). One underlying experience is fluency, either at retrieval or encoding. For example, the easier it is to retrieve fragments of the study materials, the more learners feel they know (Koriat, 1993); the easier it is to encode materials, the higher learners will judge the learning outcomes (Benjamin, Bjork, & Schwartz, 1998; Koriat & Ma'ayan, 2005; Rawson & Dunlosky, 2002). Interestingly, judgments of learning may contradict learning outcomes because items that are easy to generate (e.g., answers to general knowledge questions) are often more difficult to retrieve later than items that are difficult to generate (Benjamin et al., 1998). Indeed, in a training program to learn typing, a schedule that made learning difficult but yielded superior learning outcomes was liked less than a schedule that made learning easy but yielded inferior learning outcomes. Not surprisingly, the former group wanted to change to the easier schedule because they erroneously believed that they would learn faster (Baddeley & Longman, 1978). As we understand it, M&J's model cannot explain motivational effects where metacognitive feelings are not in line with the outcome of mental computational processes. Any theory of motivation needs to explain results where metacognitive feelings seem to play an independent role in learning decisions.

†These authors contributed equally to this work.

Third, research on intuitive problem solving suggests that perceived solution progress deviates from actual progress, which M&J's model again has difficulties to explain. Studies combining a stepwise problem-solving paradigm (Bowers, Regehr, Balthazard, & Parker, 1990) with "feelings of warmth" indicating closeness to the solution (Metcalfe & Wiebe, 1987) showed that participants felt far from the solution until right before they solved the task, even though their actual progress was closer to the solution than they were aware of (Bowers, Farvolden, & Mermigis, 1995; Reber, Ruch-Monachon, & Perrig, 2007). Evidence about the role of metacognitive judgments in learning decisions (e.g., Metcalfe & Finn, 2008) suggests that problem solvers are more likely to leave a task unsolved if it feels difficult, even if the underlying computational processes linearly progress toward the solution.

Fourth, in an aha-experience, mental processes that lead to sudden insight are accompanied by metacognitive feelings (e.g., Skaar & Reber, 2020; Webb, Little, & Cropper, 2018; for a review, see Wiley & Danek, 2024). As cognitive insight and metacognitive feelings appear simultaneously, an aha-experience is a unified construct, and metacognitive feelings do not just serve as input or output.

Finally, affect plays a major role in early phases of interest development (see Hidi & Renninger, 2006). Although individuals may not be consciously aware of their interest especially in the earliest phases of interest development (Hidi & Renninger, 2006; Renninger, Talian, & Kern, 2022), affective experiences may have a crucial role in influencing preferences and behavior during engagement with an object (Krapp, 2007). Again, affect is not just input and output, but an integral part of engagement with a subject.

These findings connect to recent theories that combine computational models, such as reinforcement learning (Briellmann & Dayan, 2022) and predictive coding (Brouillet & Friston, 2023; Fernández Velasco & Loev, 2024; Yoo, Jasko, & Winkielman, 2024) with metacognitive feelings, mainly fluency. Fluency could be seen as a parameter in computational processes, for example, short-term value in reinforcement learning or prediction precision in predictive coding. Moreover, feelings may help determine the course of action, as Fernández Velasco and Loev (2024) propose. According to this hypothesis, mental computational processes and metacognitive feelings take different roles in knowledge acquisition; the former compute the predictive dynamics whereas feelings guide action, akin to action tendencies inherent in emotions (e.g., Frijda, 1988).

Including algorithms of predictive coding and reinforcement learning would be a promising avenue to develop M&J's proposed model. Performance predictions based on metacognitive feelings play a major role in learning decisions. Thus, predictive coding accounts may refine the proposed model. Reinforcement learning (Sutton & Barto, 2018) seems promising because it builds on similar assumptions of recursive processes, including reward, as the proposed model.

Financial support. This work was supported by Research Council of Norway, #283540 and #289516.

Competing interest. None.

References

- Baddeley, A. D., & Longman, D. J. A. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics*, 21(8), 627–635. <https://doi.org/10.1080/00140137808931764>
- Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, 127(1), 55–68. <https://doi.org/10.1037/0096-3445.127.1.55>

- Bowers, K. S., Farvolden, P., & Mermigis, L. (1995). Intuitive antecedents of insight. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The creative cognition approach* (pp. 27–51). MIT Press.
- Bowers, K. S., Regehr, G., Balthazard, C., & Parker, K. (1990). Intuition in the context of discovery. *Cognitive Psychology*, 22(1), 72–110. [https://doi.org/10.1016/0010-0285\(90\)90004-N](https://doi.org/10.1016/0010-0285(90)90004-N)
- Briellmann, A. A., & Dayan, P. (2022). A computational model of aesthetic value. *Psychological Review*, 129(6), 1319–1337. <https://doi.org/10.1037/rev0000337>
- Brooks, G., Yang, H., & Köhler, S. (2021). Feeling-of-knowing experiences breed curiosity. *Memory*, 29(2), 153–167. <https://doi.org/10.1080/09658211.2020.1867746>
- Brouillet, D., & Friston, K. (2023). Relative fluency (unfelt vs felt) in active inference. *Consciousness and Cognition*, 115, 103579. <https://doi.org/10.1016/j.concog.2023.103579>
- Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process?. *Educational Research Review*, 1(1), 3–14. <https://doi.org/10.1016/j.edurev.2005.11.001>
- Erle, T. M., Reber, R., & Topolinski, S. (2017). Affect from mere perception: Illusory contour perception feels good. *Emotion*, 17(5), 856–866. <https://doi.org/10.1037/emo0000293>
- Fernández Velasco, P., & Loev, S. (2024). Metacognitive feelings: A predictive-processing perspective. *Perspectives on Psychological Science*, in press. <https://doi.org/10.1177/17456916231221976>
- Flavell, J. C., Tipper, S. P., & Over, H. (2018). Preference for illusory contours: Beyond object symmetry, familiarity, and nameability. *Emotion*, 18(5), 736–738. <https://doi.org/10.1037/emo0000386>
- Frijda, N. H. (1988). The laws of emotion. *American Psychologist*, 43(5), 349–358. <https://doi.org/10.1037/0003-066X.43.5.349>
- Hanczakowski, M., Zawadzka, K., & Cockcroft-McKay, C. (2014). Feeling of knowing and restudy choices. *Psychonomic Bulletin & Review*, 21, 1617–1622. <https://doi.org/10.3758/s13423-014-0619-0>
- Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, 41(2), 111–127. https://doi.org/10.1207/s15326985ep4102_4
- Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, 100(4), 609–639. <https://doi.org/10.1037/0033-295X.100.4.609>
- Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, 52(4), 478–492. <https://doi.org/10.1016/j.jml.2005.01.001>
- Krapp, A. (2007). An educational-psychological conceptualisation of interest. *International Journal for Educational and Vocational Guidance*, 7(1), 5–21. <https://doi.org/10.1007/s10775-007-9113-9>
- Lindell, T. A. E., Zickfeld, J. H., & Reber, R. (2022). The role of affect in late perceptual processes: Evidence from bi-stable illusions, object identification, and mental rotation. *Journal of Experimental Psychology: Human Perception and Performance*, 48(12), 1347–1361. <https://doi.org/10.1037/xhp0001059>
- Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, 15(1), 174–179. <https://doi.org/10.3758/PBR.15.1.174>
- Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & Cognition*, 15(3), 238–246. <https://doi.org/10.3758/BF03197722>
- Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(1), 69–80. <https://doi.org/10.1037/0278-7393.28.1.69>
- Reber, R., Ruch-Monachon, M.-A., & Perrig, W. J. (2007). Decomposing intuitive components in a conceptual problem solving task. *Consciousness & Cognition*, 16, 294–309. <https://doi.org/10.1016/j.concog.2006.05.004>
- Renninger, K. A., Talian, M. E., & Kern, H. M. (2022). Interest: How it develops and why it matters. In D. Fisher (Ed.), *Routledge encyclopedia of education: Educational psychology*. Routledge. <https://doi.org/10.4324/9781138609877-REE193-1>
- Schwarz, N., & Clore, G. L. (2007). Feelings and phenomenal experiences. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (2nd ed., pp. 385–407). The Guilford Press.
- Skaar, Ø. O., & Reber, R. (2020). The phenomenology of aha-experiences. *Motivation Science*, 6(1), 49–60. <https://doi.org/10.1037/mot0000138>
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.
- Topolinski, S., Erle, T. M., & Reber, R. (2015). Necker's smile: Immediate affective consequences of early perceptual processes. *Cognition*, 140, 1–13. <https://doi.org/10.1016/j.cognition.2015.03.004>
- Webb, M. E., Little, D. R., & Cropper, S. J. (2018). Once more with feeling: Normative data for the aha experience in insight and noninsight problems. *Behavior Research Methods*, 50(1), 2035–2056. <https://doi.org/10.3758/s13428-017-0972-9>
- Wiley, J., & Danek, A. H. (2024). Restructuring processes and Aha! experiences in insight problem solving. *Nature Reviews Psychology*, 3(1), 42–55. <https://doi.org/10.1038/s44159-023-00257-x>
- Yoo, J., Jasko, K., & Winkielman, P. (2024). Fluency, prediction and motivation: How processing dynamics, expectations and epistemic goals shape aesthetic judgements. *Philosophical Transactions of the Royal Society B*, 379(1895), 20230326. <https://doi.org/10.1098/rstb.2023.0326>

Mental computational processes have always been an integral part of motivation science

Michael Richter^{a*} and Guido H. E. Gendolla^b

^aFaculty of Health, Effort Lab, School of Psychology, Liverpool John Moores University, Liverpool, UK and ^bGeneva Motivation Lab, FPSE, Section of Psychology & Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland

m.richter@ljamu.ac.uk

guido.gendolla@unige.ch

www.effortlab.website

www.unige.ch/motivation

*Corresponding author.

doi:10.1017/S0140525X24000414, e41

Abstract

Some constructs in motivation science are certainly underdeveloped and some motivation researchers may work with under-specified constructs, as suggested by Murayama and Jach (M&J). However, this is not indicative of a general problem in motivation science. Many motivation theories focus on specific mechanisms underlying motivated behavior and thus have already adopted the computational process perspective that M&J call for.

Murayama and Jach (M&J) raise an important point by highlighting that some constructs in motivation science are underdeveloped: They are used as an end point of research and not as its beginning. We agree that some researchers and theorists do not show enough curiosity to fully specify the constructs that they use and do not show sufficient vigor in detailing specific motivational mechanisms or do not make them explicit enough. However, the conclusion that this is a general problem of motivation science is not warranted. Motivation science has always looked at filling the “black box” by specifying constructs and explaining how behavioral tendencies are generated. Computational models of motivation have been around for decades – even if they have not been labelled “computational.”

For instance, building on Lewin, Dembo, Festinger, and Sears’s (1944) formal theory of resulting valence, Atkinson’s risk-taking model (1957) suggested that the direction of achievement behavior – whether one approaches or avoids a specific task – depends on the relative strength of two competing motivational forces: The motivation to achieve success and the motivation to avoid failure. These two high-level constructs were further specified by postulating that they are determined by the subjective probabilities and incentive values of success and failure and weighted by individuals’ motives to achieve success and to avoid failure. Atkinson and Lewin et al.’s models thus did not only suggest high-level constructs that determine the direction of behavior, but also elaborated on the mechanisms underlying these constructs using an approach that would nowadays be called “computational.” Another early motivation theory specifying high level constructs – drive and habit – and offering an explicit computational model outlining the mechanisms determining the direction and intensity

of behavior is Hull’s (1943) drive reduction theory. More recently, Kruglanski et al.’s (2012) model of cognitive energetics has used the higher-order constructs potential driving force, restraining force, and effective driving force to explain the energization of behavior, the selection of goals, and the likelihood of goal attainment. This model also elaborates on the mechanisms underlying the postulated high-level constructs: Potential driving force is suggested to be a function of goal importance and the amount of available resources, and restraining force is predicted to be determined by resource conservation tendency, task difficulty, and the salience and importance of alternative goals. Importantly, Kruglanski et al.’s model also includes a process perspective by suggesting that the strength of potential driving force and restraining force are computed and compared before a decision about whether to engage in goal pursuit and how much effort to invest is taken. Another example is motivational intensity theory (Brehm & Self, 1989), which has been explicitly acknowledged by M&J as a model that elaborates on the mechanisms underlying motivated behavior. Motivational intensity theory suggests that task difficulty and success importance – which are postulated to be a function of need state, level of instrumentality, and incentive value – jointly determine the effort that is invested in goal pursuit. The specific process by which difficulty and success importance determine effort is predicted to be a function of the clarity of task difficulty (Richter, 2013). Like Kruglanski et al.’s model, motivational intensity theory suggests a specific sequence in which the computations are executed: Clarity of task difficulty information is processed first, followed by an assessment and comparison of task difficulty and success importance that determines whether one engages in the task at hand, and a final decision about how much effort is exerted.

The models described in the preceding paragraph constitute only a subset of the motivation-related theories that have already done what M&J ask for. Carver and Scheier’s (1981) control theory, Lewin’s (1939) field theory, Locke and Latham’s (1990) goal setting theory, Kukla’s (1972) attributional theory of performance, or Vroom’s (1964) valence-instrumentality-expectancy theory constitute further examples of models that are not limited to high-level constructs but unpack the “black box” by describing specific mechanisms that underlie the high-level constructs and clarify how they motivate behavior. Moreover, in many of these models, motivation is not considered as the initial cause of behavior but the result of a multitude of processes. It is also of note that some models of motivated behavior that seem to offer only high-level constructs often implicitly postulate more complex mechanisms underlying motivated behavior. For instance, self-determination theory’s (Ryan & Deci, 2017) high-level concept of autonomous motivation seems to constitute at first sight one of the high-level, “black box” concepts that M&J criticize. However, even if it is not frequently explained in work on self-determination, autonomous motivation is not considered to be a direct determinant of behavior. For instance, autonomous motivation is supposed to influence performance via the intervening variables perceived locus of causality, perceived volition, and freedom of choice (Cerasoli, Nicklin, & Nassreelgrawi, 2016; Reeve, 2009). Based on this theorizing, one could even argue that autonomous motivation is not considered to be the initial driving force of performance but only one of many variables that are used as input for the computational mechanisms underlying performance.

The preceding paragraphs demonstrate that motivation science has always been concerned with mechanisms underlying motivated behavior. It is certainly true that in some work the underlying

mechanisms did not get the attention that they deserve. It is also true that focusing on the mechanisms underlying high-level motivation constructs provides a great opportunity to advance our understanding of how the direction and intensity of behavior are determined. However, we disagree with M&J's position that motivation science in general avoids specifying what is inside the "black box" of high-level motivation constructs. Considering mental computational processes in motivation theories is neither new, nor something that motivation scientists need to begin to focus on. It has been an integral part of motivation science for decades. We therefore consider it of lesser relevance to remind motivation scientists that they should examine specific mechanisms underlying motivated behavior. The more important question to us is why not all motivation scientists focus on these mechanisms and why some researchers seem to be satisfied with not looking into the "black box."

Financial support. This work received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

References

- Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review*, 64, 359–372. <https://doi.org/10.1037/h0043445>
- Brehm, J. W., & Self, E. A. (1989). The intensity of motivation. *Annual Review of Psychology*, 40, 109–131. <https://doi.org/10.1146/annurev.ps.40.020189.000545>
- Carver, C. S., & Scheier, M. F. (1981). *Attention and self-regulation: A control-theory approach to human behavior*. Springer.
- Cerasoli, C. P., Nicklin, J. M., & Nassreldgrawi, A. S. (2016). Performance, incentives, and needs for autonomy, competence, and relatedness: A meta-analysis. *Motivation and Emotion*, 40, 781–813. <https://doi.org/10.1007/s11031-016-9578-2>
- Hull, C. L. (1943). *Principles of behavior*. Appleton-Century.
- Kruglanski, A. W., Bélanger, J. J., Chen, X., Köpertz, C., Pierro, A., & Mannetti, L. (2012). The energetics of motivated cognition: A force-field analysis. *Psychological Review*, 119, 1–20. <https://doi.org/10.1037/a0025488>
- Kukla, A. (1972). Foundations of an attributional theory of performance. *Psychological Review*, 79, 454–470. <https://doi.org/10.1037/h0033494>
- Lewin, K. (1939). Field theory and experiment in social psychology: Concepts and methods. *The American Journal of Sociology*, 44, 868–896. <https://doi.org/10.1086/218177>
- Lewin, K., Dembo, T., Festinger, L., & Sears, P. S. (1944). Level of aspiration. In J. M. Hunt (ed.), *Personality and the behavior disorders* (pp. 333–378). Ronald Press.
- Locke, E.A., & Latham, G.P. (1990). *A theory of goal setting and task performance*. Prentice-Hall.
- Reeve, J. (2009). *Understanding motivation and emotion*. Wiley.
- Richter, M. (2013). A closer look into the multi-layer structure of motivational intensity theory. *Social and Personality Psychology Compass*, 7, 1–12. <https://doi.org/10.1111/spc3.12007>
- Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guildford.
- Vroom, V. H. (1964). *Work and motivation*. Wiley.

Motivational constructs: Real, causally powerful, not psychologically constructed

Andrea Scarantino* 

Department of Philosophy, Georgia State University, Atlanta, GA, USA
ascarantino@gsu.edu
<https://sites.google.com/site/andreascarantinoswebsite/>

*Corresponding author.

doi:10.1017/S0140525X24000530, e42

Abstract

Murayama and Jach criticize the use of high-level motivational constructs in psychology, urging psychologists to “unpack” the black box. These constructs are alleged to be “psychological constructions” with no causal powers of their own. I argue that this view is mistaken, and that high-level motivational constructs are causal even when unpacked in terms of underlying computational, algorithmic, and implementational processes.

The positing of motivational constructs results from *black-box inferences*, which consist of surmising, upon observing causes generating effects in a system, that an intervening variable mediates between them (Sober, 1998). These inferences allow psychologists to posit that an organism does what it does because of *thirst*, *fear*, or *need for competence*, just as they allow biologists to posit *genes*, physicists to posit *protons*, and sociologists to posit *socioeconomic status* in their explanatory practices.

Murayama and Jach (M&J) criticize the use of high-level motivational constructs in psychology like *need for competence* or *need to belong*, urging psychologists to “unpack” the black box. What is the rationale for this unpacking? Marr (1982) distinguished three equally important levels of analysis: A *computational level* describing the function computed by a system, an *algorithmic level* describing the representations and algorithm used to compute such function, and an *implementation level* describing how the representations and algorithm are physically realized.

The authors declare allegiance to this framework, but add that, once a computational analysis has been offered, high-level motivational concepts lose their causal relevance – high-level motivational constructs are mere “psychological constructions” with no causal powers of their own. Given the centrality of the notion of psychological construction, the target article says regrettably little about it. The authors claim to echo constructionism about emotions, so we are left with the impression that what makes emotions psychologically constructed (if anything) is what makes motivational concepts psychologically constructed.

Some brief remarks on psychological construction appear in a discussion on *affiliative motivation*. Ordinary people observe one another, note a tendency to affiliate with certain social groups and make a black-box inference that people “have an affiliative...motivation.” The problem is that affiliative motivation is not “itself represented in our mental computational processes,” but results from “people’s subjective construction of...mental processes, and should not be considered as the determinant of behavior.” In other words, affiliative motivation is a psychological construction rather than a real cause, that is, just a convenient way for an external observer to interpret behavior. This analogy is problematic, because ordinary people and scientists are not engaged in the same activity when they make black-box inferences. Scientists posit motivational constructs which aspire to be scientifically explanatory; ordinary people posit motivational constructs which aspire to be explanatory in the folk psychological sense. The fact that a construct has its origin in ordinary language, as *affiliative motivation* does, does not settle the question of whether it has “theoretical status” in science.

Consider the ordinary language constructs of *water* and *air*. The difference between them is that water – defined as H₂O – can be embedded in chemically interesting generalizations, whereas air is too heterogeneous for that purpose, from which it follows that water has theoretical status in chemistry, and air

does not (Moors, 2022). This is precisely what psychological constructionists have argued for emotion concepts. On their view, the trouble with emotion concepts is not that they are naïve concepts, but that they are naïve concepts like *air* rather than like *water*. They do not share physical properties or mechanisms of interest to affective scientists and therefore are not natural kinds (Barrett, 2006; see Scarantino, 2015 for a response).

This constructionist punchline is missing entirely from the target article: No evidence is provided that motivational constructs do not allow for interesting psychological generalizations. On the contrary, the authors acknowledge that “motivation constructs...have great utility in that they can make generalizable predictions,” they can “make our explanation parsimonious,” and they “can inform researchers of potential intervention programs.” This reads like a *prima facie* case for giving motivational concepts natural kind status in psychology, contrary to what psychological constructionists have claimed about emotion concepts.

But M&J add that “we should not interpret these empirical findings as evidence that high-level motivation constructs directly cause behavior.” They seem to assume that if there is a lower-level computational explanation available, the higher-level motivational explanation stops being causal. But what does it mean for A to cause B? The authors appear to endorse an *interventionist account* of causation, according to which A causes B just in case intervening on A is an *effective strategy* for changing B.

On this interventionist view, motivational constructs are straightforwardly causal: If you intervene on *thirst* (a low-level naïve motivational concept they have no qualms with), you can change drinking behaviors, and if you intervene on *need for competence*, you can change exploratory behaviors. Even if we follow M&J in presupposing that *need for competence* motivates by virtue of a *computational process* which pursues the rewarding value of information, it remains true that intervening on the need for competence is an *effective strategy* for changing exploratory behaviors.

M&J’s argument also proves too much: It could be used to undermine the causal powers of computational variables themselves. If lower-level causal explanations exclude higher-level ones, we would have to conclude that the algorithmic and implementation processes underlying reward maximization deprive the computational variables of causal powers. Physical processes at the subatomic level may end up being the only genuinely causal processes on this view, assuming that there are no lower-level physical processes below them. I assume the authors would not welcome this implication.

I suggest that the trouble with *need for competence* is not that it lacks causal powers, but rather that it is too heterogeneous as a motivational construct, because it purports to explain behaviors as diverse as the exploration of potential majors by a university student and the exploration of a maze by a rat. Such behaviors are fundamentally different not necessarily at the computational level – if we accept the ubiquity of reinforcement-learning models – but certainly at the algorithmic and implementation levels (cf. Piccinini, 2020).

To conclude, I agree that we should not limit our explanations of behavior to the mere positing of motivational constructs, and we should thoroughly investigate their lower-level realizers. The reason is that a full understanding of how causally powerful motivational constructs cause behavior demands figuring out their computational, algorithmic, and implementation dimensions, which can guide us to discovering the most fruitful natural kinds at different levels of behavioral explanation.

Acknowledgements. I want to thank Gualtiero Piccinini for helpful feedback on a previous draft.

Financial support. None.

Competing interests. None.

References

- Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science*, 1(1), 28–58.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W.H. Freeman.
- Moors, A. (2022). *Demystifying emotions: A typology of theories in psychology and philosophy*. Cambridge University Press.
- Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford University Press.
- Scarantino, A. (2015). Basic emotions, psychological construction, and the problem of variability. In L. F. Barrett & J. A. Russell (Eds.), *The psychological construction of emotion* (pp. 334–376). The Guilford Press.
- Sober, E. (1998). Black box inference: When should intervening variables be postulated? *The British Journal for the Philosophy of Science*, 49(3), 469–498.

Adopt process-oriented models (if they’re more useful)

Brendan A. Schuetze^{a,b*} and Luke D. Rutten^c

^aDepartment of Education Science, University of Potsdam, Potsdam, Germany;

^bDepartment of Educational Psychology, University of Utah, Salt Lake City, UT, USA and ^cDepartment of Educational Psychology, The University of Texas at Austin, Austin, TX, USA

brendan.schuetze@gmail.com

luke_rutten@utexas.edu

<https://schu.ete.co>

*Corresponding author.

doi:10.1017/S0140525X24000372, e43

Abstract

Though we see the potential for benefits from the development of process-oriented approaches, we argue that it falls prey to many of the same critiques raised about the existing construct level of analysis. The construct-level approach will likely dominate motivation research until we develop computational models that are not only accurate, but also broadly usable.

Tradeoffs between accuracy and parsimony are inherent in most scientific endeavors. Murayama and Jach (M&J) argue that contemporary motivation theories, which operate nearly exclusively at the construct level, are making the wrong tradeoffs between accuracy and parsimony. They challenge the idea that motivation constructs directly cause behavior. Instead, they argue (1) constructs are essentially epiphenomenal byproducts, and (2) process-oriented computational models are necessary to unpack mechanisms of motivated behavior and advance the field of motivation science.

We agree with the first premise. Motivation constructs are not literal causes of behavior; researchers will never find a need-for-competence dial in the brain (at least, not in the way that latent variable models assume). Rather, the need-for-competence and other motivational constructs act as labels that summarize

patterns emerging from yet to be defined processes. M&J argue that the best path forward is to begin attempting to “unpack the black box” and define these processes. Their proposed solution of computational modeling, however, shares many of the same faults as the construct-level approach they critique; they simply trade one black box for another. Where construct-focused approaches assume a need-for-competence drives behavior, their exemplar knowledge acquisition model assumes an intrinsic-reward-for-knowledge drives behavior. Importantly, we could ask the same question of the reward-for-knowledge as M&J ask of the need-for-competence: What process creates this drive? As acknowledged by the authors themselves, their suggestion to move down a level of analysis does not solve the black-box problem. It merely changes the black boxes we use.

Because neither construct nor computational models address the black-box problem, we need an alternative way to evaluate between them. One might even argue that the black-box language obfuscates the real point: That the best measure of a model is its accuracy and that process-oriented models provide a way toward greater accuracy. We disagree with the former notion. The question of which level of mental process is best to model is not a question of “Which level is most accurate?” Of course, the answer to that will *always* be the next level down.

Rather, we believe it better to ask “Which level is most useful?” Given our background in applied psychology, and specifically our experiences training pre-service teachers, we know that one of the ultimate goals of motivation theory is to generate insights with practical importance for teachers, students, bosses, workers, and so on. These experiences have led us to adopt a more pragmatist philosophy of science, wherein a key feature of any worthwhile theory is that it can be used to make an impact (Elder-Vass, 2022; James, 1907/2001). History tells us this is where computational models struggle.

For example, educational researchers including Carroll (1963) and Bloom (1976) made a strong push for computational and process-oriented “models of school learning” in the mid-twentieth century (see also Bjork, 1973). However, those models were difficult to understand, even for researchers. Consequently, these theories received minimal adoption and faded in importance (Harnischfeger & Wiley, 1978). More recently, researchers in self-regulated learning – ourselves included (Schuetze, 2024) – have put forth a number of non-construct-focused models based on discrepancy reduction (e.g., Ackerman, 2014; Carver & Scheier, 1990; Thiede & Dunlosky, 1999). Many researchers have testified to the benefits of building these sorts of models for the purpose of theory development (e.g., Aubé, 1997; Guest & Martin, 2021; van Rooij & Blokpoel, 2020). However, due to their complexity and relatively narrow areas of focus, these process-oriented models have struggled to make the same impact on school systems and business leaders as construct-focused understandings of human behavior, such as self-efficacy and growth mindset.

Our contention here is that even if we create highly accurate theories of motivated behavior, if they are not usable or interpretable by those who are in positions to apply them, more needs to be done. Applied to M&J’s proposal, we believe that computational models of motivation can be useful to the world at large – but this will require additional work. Part of this work may mean finding creative places to implement motivational theories, such as in intelligent tutors, where theoretical complexity is managed by a technical system, as opposed to by teachers or managers (Yan, Sana, & Carvalho, 2024). Other parts of this work may mean creating a hierarchy of mutually compatible theories

operating at different levels of analysis. Insights derived from lower levels can be distilled and moved up to higher (perhaps construct) levels that require less time and expertise to put into practice (Anderson, 2002; Donoghue & Horvath, 2016). In essence, the researcher’s theory of motivation doesn’t necessarily need to be the same as the practitioner’s. Different groups can understand the same phenomenon in distinct levels of detail. Indeed, divisions of this sort help molecular biologists, pharmacists, and medical doctors create and administer medical treatments that work despite having very different levels of focus. Similarly, the average driver does not need to understand how their car’s engine functions. The mechanic, however, must. Again, theories at different levels should aim not for the utmost accuracy, but for the utmost utility to those who are using it.

With this in mind, we find ourselves agreeing with M&J’s second premise as well: That computational and process-oriented models can help us better understand motivation. Not because they help us solve the black-box problem, but because they show promise in helping us make more useful and applicable theories of motivated behavior. However, pitfalls accompany this promise, and they must be kept in mind. Given the history of computational modeling and the associated increases in time, effort, and expertise required to use these models, we believe that great care will be needed to translate them into practically applicable formats. Until we achieve broadly usable computational models, we cannot fault motivation researchers or practitioners for sticking to tried-and-true construct-oriented models. Pending that development, the construct level of analysis will continue to be the primary lever of change outside the lab.

Acknowledgement. The authors thank Veronica Yan for comments on a draft of this commentary.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

References

- Ackerman, R. (2014). The diminishing criterion model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General*, 143(3), 1349–1368. <https://doi.org/10.1037/a0035098>
- Anderson, J. R. (2002). Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science*, 26(1), 85–112. https://doi.org/10.1207/s15516709cog2601_3
- Aubé, M. (1997). Toward computational models of motivation: A much needed foundation for social sciences and education. *Journal of Artificial Intelligence in Learning*, 8(1), 43–75.
- Bjork, R. A. (1973). Why mathematical models? *American Psychologist*, 28(5), 426–433. <https://doi.org/10.1037/h0034623>
- Bloom, B. S. (1976). *Human characteristics and school learning*. McGraw-Hill.
- Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, 64(8), 1–9. <https://doi.org/10.1177/0161468163064008>
- Carver, C. S., & Scheier, M. F. (1990). Origins and functions of positive and negative affect: A control-process view. *Psychological Review*, 97(1), 19–35. <https://doi.org/10.1037/0033-295X.97.1.19>
- Donoghue, G. M., & Horvath, J. C. (2016). Translating neuroscience, psychology and education: An abstracted conceptual framework for the learning sciences. *Cogent Education*, 3(1), 1267422. <https://doi.org/10.1080/2331186X.2016.1267422>
- Elder-Vass, D. (2022). Pragmatism, critical realism and the study of value. *Journal of Critical Realism*, 21(3), 261–287. <https://doi.org/10.1080/14767430.2022.2049088>
- 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>
- Harnischfeger, A., & Wiley, D. E. (1978). Conceptual issues in models of school learning. *Journal of Curriculum Studies*, 10(3), 215–231. <https://doi.org/10.1080/0022027780100304>

- James, W. (1907/2001). What pragmatism means. In A. Delbanco (Ed.), *Writing new England* (pp. 80–93). Harvard University Press. <https://doi.org/10.4159/harvard.9780674335486.c18>.
- Schuetze, B. A. (2024). A computational model of school achievement. *Educational Psychology Review*, 36, 18. <https://doi.org/10.1007/s10648-024-09853-6>
- Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(4), 1024–1037. <https://doi.org/10.1037/0278-7393.25.4.1024>
- van Rooij, I., & Blokpoel, M. (2020). Formalizing verbal theories: A tutorial by dialogue. *Social Psychology*, 51(5), 285–298. <https://doi.org/10.1027/1864-9335/a000428>
- Yan, V. X., Sana, F., & Carvalho, P. F. (2024). No simple solutions to complex problems: Cognitive science principles can guide but not prescribe educational decisions. *Policy Insights from the Behavioral and Brain Sciences*, 11(1), 59–66. <https://doi.org/10.1177/23727322231218906>

Beyond reductionism: Understanding motivational energization requires higher-order constructs

Kennon M. Sheldon^{a*} and Richard M. Ryan^{b,c}

^aPsychological Sciences, University of Missouri, Columbia, Missouri; ^bAustralian Catholic University, Sydney, Australia and ^cEwha Womans University, Seoul, Korea

Sheldonk@missouri.edu

Richard.Ryan@acu.edu.au

missouri.edu

selfdeterminationtheory.org

*Corresponding author.

doi:10.1017/S0140525X24000438, e44

Abstract

We argue that the target article's computational/reductionistic approach to motivation is insufficient to explain the energization of human behavior, because such explanation requires broad consideration of "what people are trying to do." We illustrate what is gained by retaining (rather than jettisoning) higher-order motivation constructs and show that the authors' approach assumes, but fails to name, such constructs.

Human motivation theories are supposed to explain the energization and direction of behavior. In their target article, Murayama and Jach (M&J) argue that theories involving broad constructs such as goals, motives, and needs are "pitched too high" for this purpose. Specifically, they assert that traditional motivational concepts, like intrinsic motivation (IM) or needs for competence, autonomy, or belongingness, play no role in energizing behavior. What *really* causes behavior is some array of simpler cognitive mechanisms as addressed by the authors' reinforcement learning paradigm (Figure 2). Understanding motivation is thus not about understanding people, as they attempt to discern and meet their own needs; rather, it is about understanding cognitive processes that are largely inscrutable to the people that they run.

These are bold conclusions, given that decades of research have established the explanatory and practical utility of higher-order motivation constructs. For example, psychological autonomy (i.e., feeling ownership of one's behavior), measured in different ways and contexts, has emerged as critical for persistent

engagement and mental health (Ryan et al., 2022). In education, meta-analyses show that interventions boosting teachers' autonomy support reliably enhance student IM, engagement, and performance (e.g., Reeve, Ryan, Cheon, Matos, & Kaplan, 2022). Yet such social- and personality-level factors do not count as causes from these authors' computational perspective.

So, what motivational energizers are identified by their reinforcement learning paradigm? Scrutiny of Figure 2 reveals some puzzles in this regard. In that figure, first note that all downstream processes are driven (starting at the top of the figure) by "awareness of a knowledge gap." But doesn't such awareness imply a pre-existing motivation to know or learn something? If the person has no such desire, they will not perceive or care about the gap. Notably, IM (wanting to do an activity because it is interesting) is known to enhance people's sensitivity to knowledge gaps.

Befitting the article's reductionistic stance, however, no motives are represented in Figure 2. Still, arrows lead from "existing knowledge network" (bottom) to "the generation of new questions" (middle right), to "awareness of a knowledge gap" (top). This sequence appears to assume that people want to increase their knowledge of the world (a hallmark of IM theories), but this is not explained. Instead, we are informed that "In the reward-learning framework, there is an implicit assumption that people choose to seek information that has a high reward value." This is the circular problem most reinforcement theories have had, as they are devoid of content regarding experiential value and reward. In the center-right of Figure 2 one finds "rewarding experiences," but these energize information-seeking behavior only via a side-loop affecting "the expected reward value of new information," which is said to moderate the path from knowledge gap to information-seeking. In short, M&J's model replaces a relatively straightforward scheme (a person pursues knowledge in domains in which they have interest or value, and that motivation can be enhanced or diminished by experienced supports and obstacles) with a less intuitive scheme, in which the energizers of behavior are either unspecified, or split up amongst an array of low-level process variables.

Why do we need higher-order motivation constructs? There are many possible justifications. One is that cybernetic/hierarchical models of action control assert that much behavior is energized and directed by the abstract or long-term goals a person adopts (Carver & Scheier, 1981). For example, the goal "I will become a researcher!" has likely organized the daily behaviors of many BBS readers. Similarly, broad motive dispositions (e.g., nAchievement) are known to result from childhood environments and affordances (McClelland, Koestner, & Weinberger, 1989) that set parameters for what people strive for throughout their lives (Sheldon & Schuler, 2011). As a third example, self-determination theory (Ryan & Deci, 2017) shows that people better internalize and sustain motivation for activities in which psychological needs for competence, relatedness, and autonomy can be satisfied, predicting learning and engagement over time.

In short, we would argue that causality is not invariably bottom-up as M&J imply. Instead, the "low level mental computations" they highlight may better be thought of as part of the *how* of motivation (i.e., the ways in which our preferences and motives are executed), not the *why* of motivation (i.e., its energizers). They are mechanisms which *serve* our varied goals and motives, rather than always *determining* them.

Indeed, Figure 1B contains a very relevant arrow, which M&J don't discuss, that leads *down* from "subjective experiences" to "mental computational mechanisms." We suggest that this top-down path

illustrates how a desired goal can shape specific mechanisms within the cognitive machinery, in service of approaching a goal or future state. Once we decide we really want something, we have impressive capabilities that can serve those wants (Sheldon, 2014).

Thus, our preferred model of science is not computational reductionism, but rather *consilience* (Wilson, 1998), in which scientists coordinate multiple levels of description using appropriate organizing constructs. We are interested in every level of analysis, from the social and interpersonal to the mechanistic. Of course, computational models may emerge as important research tools, but they do not “replace” or fully explain other levels of description. As Ryan and Deci (2017) argued, psychological theories are not distinct from biological accounts, and can be coordinated with them. Yet psychological events are lawful and important and are “typically the most practical level at which we can intervene in human affairs (p. 7).” In contrast, Figure 2’s mechanistic model provides little practical leverage for affecting real-world behaviors.

In conclusion, although the target article’s point is well-taken (beware of over-reifying concepts), we think it is a mistake to throw out higher motivational constructs altogether. These are not just illusions or post-behavioral constructions; they reflect real causal propensities and persistent regularities in the dynamics of human striving. They help us understand both what people are trying to do in life and the social conditions that support or thwart these motives. Without them, we may be stranded in a world of mechanisms, having lost sight of the real people who deploy them.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

References

- Carver, C. S., & Scheier, M. F. (1981). *Attention and self-regulation: A control theory approach to human behavior*. Springer-Verlag.
- McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review*, 96(4), 690–702.
- Reeve, J., Ryan, R. M., Cheon, S. H., Matos, L., & Kaplan, H. (2022). *Supporting students’ motivation: Strategies for success*. Routledge.
- Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Press.
- Ryan, R. M., Duineveld, J. J., Di Domenico, S. I., Ryan, W. S., Steward, B. A., & Bradshaw, E. L. (2022). We know this much is (meta-analytically) true: A meta-review of meta-analytic findings evaluating self-determination theory. *Psychological Bulletin*, 148(11–12), 813–842.
- Sheldon, K. M. (2014). Becoming oneself: The central role of self-concordant goal selection. *Personality and Social Psychology Review*, 18(4), 349–365.
- Sheldon, K. M., & Schuler, J. (2011). Needing, wanting, and having: Integrating motive disposition theory and self-determination theory. *Journal of Personality and Social Psychology*, 101, 1106–1123.
- Wilson, E. O. (1998). *Consilience: The unity of knowledge*. Alfred Knopf.

Postcard from inside the black box

David Spurrett* 

Philosophy, University of KwaZulu-Natal, Durban, South Africa
spurrett@ukzn.ac.za
<https://philpeople.org/profiles/david-spurrett>

*Corresponding author.

doi:10.1017/S0140525X24000578, e45

Abstract

There are indeed questionable motivation constructs in psychology. The diagnosis and proposed remedies in the target article both neglect the crucial consideration that all tendencies to behaviour compete for the same finite set of degrees of freedom. Action selection also has irreducibly economic aspects which should constrain motivation constructs and already inform healthy research programmes.

The target article is correct that some motivation constructs in psychology are questionable, and that something needs doing. The dubious construct problem is not unique to motivation. The rate at which constructs are being introduced in psychology generally is rising and the number of times each is deployed empirically simultaneously falling (Elson, Hussey, Alsalti, & Arslan, 2023). In the case of motivation, the diagnostic parts of the target article and its constructive proposal both neglect the crucial consideration that all tendencies to behaviour compete for the same finite set of degrees of freedom, primarily of the body, which have alternative uses.

While the target article correctly notes that motivation is a “determinant of behaviour” but fails to take what we know about behaviour control and production sufficiently seriously. Most actions or behaviours involve some deployment of the body which has a finite number of joints and muscles. Some deployments are mutually exclusive: Nobody can stand and lie down at once. Some aren’t: Many people can walk and chew gum at once. The finite number of joints and muscles, along with facts about the orientation of the body and relevant surfaces and media, set a finite number of available degrees of freedom. These deployments stand, furthermore, in heterogeneously structured relations of mutual exclusivity.

Sherrington (1906) introduced the notion of a final common path, referring to the last neural stage at which competition between incompatible deployments of combinations of muscles can be resolved. There is no definitive and total final common path for the whole body because the selection of possible movements that is available at any time is sensitive to factors including orientation, gravitation, inertia, and the arrangement and properties of nearby surfaces. As Gallistel (1980) has explained this variability implies that control of skeletal muscles must be expressed through a “lattice hierarchy” in which the level at or before which competition over deployment of degrees of freedom must be resolved is not fixed, but depends on the properties of candidate actions and the situation of the body (Spurrett, 2021b). Since the available bodily means are scarce and have alternate uses, the problem of selecting between deployments of them is irreducibly amenable to an economic analysis in terms of efficiency of goal attainment (Spurrett, 2021a).

There can be various types of explanation for the movements of bodies. Some refer to motivation whether basic (she was hungry) or higher level (she needs “competence”) while others don’t and might involve reflexes or habits. While the problem of allocating scarce means with alternative uses is essentially economic in character, the solutions need not be, and theorists have contemplated both processes that are sensitive to returns and opportunity costs, and ones that aren’t. Any genuine cause of embodied activity must compete for control of the required degrees of freedom with the causes of other possible deployments. And genuine competition must happen at or before the applicable places in the lattice hierarchy. So a test of any motivation construct is what it has

to say about how the hypothesised factor joins this competition. For example, if a genuine motivation for “competence” is competing with fatigue over whether to get up from the couch to train, the two are in conflict over what the legs do next. So however “basic” or “higher-order” a motivational source might be it must interact in some way with any other process that could control the body.

These considerations provide methodologically significant constraints on satisfactory theorising about motivation. Worthwhile hypotheses about higher-order behaviour should have specific commitments regarding this interaction. The target article, however, makes no mention of bodies or competition of control for it and so misses an enormously valuable tool for evaluating motivation constructs. This lapse of diagnosis also applies to the prescription since the same constraints play no role in articulating or defending the offered remedies. This isn’t a merely theoretical criticism because some research programmes have been taking this seriously for decades. In neuroeconomics, focusing on circuits constituting or being upstream of “final common paths” for body control led to significant discoveries about neural processes of valuation and selection (e.g., Platt & Glimcher, 1999). Subsequent work has shown that the very same bottlenecks process rewards of widely varying modalities including money, food, drink, relief from pain, and social reputation, where rewards can be certain or risky, immediate or delayed, larger or smaller (Levy & Glimcher, 2012 and Bartra, McGuire, & Kable, 2013 are useful meta-analyses). These findings aren’t simply read off the brain but depend on behavioural estimates of subjective value. The behavioural data that are required to interpret the neural data include determining how subjects trade-off rewards in various modalities, that is, how much money would be given up for how much drink, or relief from pain, and so on. These considerations point to a more demanding notion of what “unpacking the black box” should amount to and show that work meeting those criteria has already engaged some “higher-order” rewards like social reputation. Opening the black box does mean “specifying mental computation” but that demands more than occasionally saying “emergent property” without providing criteria or empirical content. What it requires is identifying neural circuits in ways sensitive to the economic character of the embodied selection problem and constrained by empirical discoveries in neuroeconomics. It also requires studying activity in those circuits armed with choice data that meaningfully characterises subjective values.

The point here isn’t that neuroeconomics has definitively shown that all competition for action involves estimations of utility or expected subjective value. The regulative hypothesis that it does guides much neuroeconomic research and has not yet been refuted, but it remains an empirical claim. The point is that hypotheses about motivation, however “higher order” their postulates, are hypotheses about what make us do things or refrain from doing them. We do them with our bodies, and close attention to how bodies are controlled and to how their movements are selected simply isn’t optional.

Funding statement. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

References

- Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage*, 76, 412–427.
- Elson, M., Hussey, I., Alsalti, T., & Arslan, R. C. (2023). Psychological measures aren’t toothbrushes. *Communications Psychology*, 1(1), 1–4.
- Gallistel, C. R. (1980). *The organisation of action: A new synthesis*. Lawrence Erlbaum.
- Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, 22, 1027–1038.
- Platt, M. L., & Glimcher, P. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, 400, 233–238.
- Sherrington, C. S. (1906). *The integrative action of the nervous systems*. Yale University Press.
- Spurrett, D. (2021a). The descent of preferences. *British Journal for the Philosophy of Science*, 72(2), 485–510.
- Spurrett, D. (2021b). Time and the decider. *Behavioral and Brain Sciences*, 44, e133.

Predictive processing: Shedding light on the computational processes underlying motivated behavior

Lieke L. F. van Lieshout^a , Zhaoqi Zhang^a,
Karl J. Friston^b and Harold Bekkering^{a*} 

^aDonders institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands and ^bThe Wellcome Trust Centre for Neuroimaging, University College London, London, UK.

lieke.vanlieshout@donders.ru.nl

claire.zhang@donders.ru.nl

k.friston@ucl.ac.uk

harold.bekkering@donders.ru.nl

<https://www.ru.nl/en/people/lieshout-l-van>

<https://www.ru.nl/en/people/zhang-z-claire>

<https://profiles.ucl.ac.uk/2747-karl-friston>

<https://www.ru.nl/en/people/bekkering-h>

*Corresponding author.

doi:10.1017/S0140525X24000396, e46

Abstract

Integrating the predictive processing framework into our understanding of motivation offers promising avenues for theoretical development, while shedding light on the computational processes underlying motivated behavior. Here we decompose expected free energy into intrinsic value (i.e., epistemic affordance) and extrinsic value (i.e., instrumental affordance) to provide insights into how individuals adapt to and interact with their environment.

We agree with the authors that motivation is often viewed as a high-level construct, defined by many researchers as a causal determinant of behavior in a “black-box” fashion. We also agree that to understand motivation means to understand the “spring to action.” Beyond the conventional constructs of high-level motives, we need to define what and how this energization underwrites action selection or choice behavior.

We take this opportunity to rehearse the key arguments in Murayama & Jach, as seen through the lens of the predictive processing (a.k.a., active inference) account of motivated behavior. Active inference is sometimes portrayed as an account of sentient behavior, implying actions driven by the process of sense making. In this account, perception is formulated as a process of inference and, thereby, rests on a calculus of beliefs – sometimes referred to as Bayesian mechanics (Ramstead et al., 2023) or self-evidencing

(Hohwy, 2016). Meanwhile, active inference suggests that behavior can be reflexive or planned, depending on whether it aims to minimize prediction errors or anticipates future consequences. The concept of expected surprise, derived from this framework, encompasses both uncertainty resolution and avoidance of unexpected outcomes, guiding goal-directed behavior. This explanation of motivated behavior is grounded in statistical physics and highlights both the inherent role of intrinsic value (i.e., epistemic affordance) and extrinsic value (i.e., instrumental affordance). We begin by critically examining “reward learning models of information seeking,” followed by an outline *how* self-organization processes can drive motivated behavior.

The paper’s emphasis is on “reward learning models of information seeking behavior” suggesting that information gain has intrinsic value and serves as a source of motivation. We argue this view can be replaced – or at least be elaborated – under an active inference formulation. Arguably, the most crucial aspect is that prior preferences, which form the basis of expected value, encompass all aspects of sensory experience, and cannot be simplified to a mere reward function. In other words, these preferences function to prevent surprising outcomes that would diverge from an individual’s typical expectations, maintaining consistency with their self-concept (i.e., “kind of thing that I am”). Consequently, the expected free energy can be decomposed into intrinsic value (i.e., epistemic affordance) and extrinsic value (i.e., instrumental affordance).

More explicitly, the “kind of thing I am” refers to a necessary aspect of entities capable of self-organization; specifically, those with the ability to infer the consequences of their actions. The imperative is to maximize the evidence (a.k.a., marginal likelihood) for generative models of how observations are caused. In contexts where individuals are actively making decisions, their beliefs about what actions to take are influenced by the expected free energy associated with each possible choice. Essentially, they weigh the potential consequences of each option and choose the one with the most favorable expected outcome in terms of minimizing expected surprise or uncertainty. This is in contrast with expected utility theory, in which there is merely one specific kind of outcome that is considered the most desirable, determining the utility or reward function (e.g., monetary payoff). However, in active inference, this approach is replaced with a system where preferences guide decision-making rather than a singular utility function. This means that instead of aiming to maximize a monothematic payoff, individuals prioritize choices based on their preferences among various possible outcomes. Your view of who you are determines how you encounter *every* aspect of an observable outcome. Imagine you are a student of cognitive neuroscience who is to be examined about the content of this BBS article. You face a trade-off between time spent reading the article and making dinner for your friends. Quickly reading through the article might be enough to pass the exam while leaving you adequate time to make a meal. Your approach may vary depending on your self-perception, such as whether you view yourself as an exceptional student or not. Ultimately, your actions are likely guided more by personal preferences and considerations than by a single reward function.

The big question is now: How do mental computational processes self-organize? Mental computational processes have the capacity to infer the consequences of actions by minimizing surprise and prediction errors. This endows generative models with a future-pointing aspect and the notion of planning (as inference). This perspective suggests that humans and animals often exhibit behavior aimed at mastering the environment, driven by a combination of intrinsic and extrinsic values. This dual aspect

decomposition of affordances suggests that agents are compelled to explore their environments to maximize information gain – actively gather evidence to build and optimize world models – while being sensitive to the constraints of their preferences and goals. This interpretation aligns with the notion that humans can recognize regularities and creating mental categories from their own behaviors and subjective experiences. Formally, this can be expressed as minimizing an evidence bound called variational free energy (Winn & Bishop, 2005) that comprises complexity and accuracy (Ramstead et al., 2023):

$$\text{Variational free energy} = \text{model complexity} - \text{model accuracy}$$

Complexity scores the divergence between prior beliefs (before seeing outcomes) and posterior beliefs (after seeing outcomes), while accuracy corresponds to the goodness of fit. In short, complexity scores the information gain or cost of changing one’s mind in an information theoretic and thermodynamic sense, respectively. Through repeated interactions with the environment, the brain updates its models based on the prediction errors it encounters. This self-organized learning process allows the brain to proactively infer what actions will lead to expected outcomes, adaptively learning from mistakes and adjusting future behavior to minimize surprises. By understanding how different situations affect outcomes and how actions unfold over time, these cognitive systems can plan actions strategically to minimize surprises and errors in the long term, not just at that moment. This decomposition furnishes a complementary perspective on the complex interplay between extrinsic and intrinsic motivation.

In summary, integrating the predictive processing framework into our understanding of motivation offers promising avenues for theoretical development, shedding light on the computational processes underlying motivated behavior and providing insights into how individuals adapt to and interact with their environment.

Financial support. None.

Competing interest. None.

References

- Hohwy, J. (2016). The self-evidencing brain. *Noûs*, 50(2), 259–285.
- Ramstead, M. J., Sakthivadivel, D. A., Heins, C., Koudahl, M., Millidge, B., Da Costa, L., ... Friston, K. J. (2023). On Bayesian mechanics: A physics of and by beliefs. *Interface Focus*, 13(3), 20220029.
- Winn, J., & Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research*, 6(4), 661–694.

There’s no such thing as a free lunch: A computational perspective on the costs of motivation

Eliana Vassena^{a*}  and Jacqueline Gottlieb^b

^aBehavioural Science Institute, Radboud University, Nijmegen, The Netherlands and ^bDepartment of Neuroscience and the Mortimer Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA
eliana.vassena@donders.ru.nl
jg2141@columbia.edu

*Corresponding author.

doi:10.1017/S0140525X2400058X, e47

Abstract

Understanding the psychological computations underlying motivation can shed light onto motivational constructs as emergent phenomena. According to Murayama and Jach, reward-learning is a key candidate mechanism. However, there's no such thing as a free lunch: Not only benefits (like reward), but also costs inherent to motivated behaviors (like effort, or uncertainty) are an essential part of the picture.

Imagine you are offered 1 million euros to work for a year on an extremely high-stakes job in a remote location. Although this may sound like a worthwhile reward, it entails spending a long time away from loved ones in a high-pressure stressful environment. *Despite* the considerable rewards of this “once in a lifetime” opportunity, you may pass on it because of its high costs. How did you make this decision?

Scientists across multiple fields, from psychology, to neuroscience to robotics, have put a lot of effort into the challenge of defining motivation and its workings (ironically, one might add). However, as Murayama and Jach rightly emphasize, there is a pressing need for studies of motivation to move beyond the “black-box” approach and provide more precise definition, quantification, and implementation to the many concepts they have associated with motivation. We wholeheartedly support this constructivist view and contend that it applies not only to the rewards but also the costs of a situation.

A fundamental principle of decision-making in economics, neuroscience, and psychology is that individuals generate adaptive behavior by making trade-offs between the benefits and costs of alternative options (Camerer, 2008; Silvestrini, Musslick, Berry, & Vassena, 2023; Silveti, Vassena, Abrahamse, & Verguts, 2018; Westbrook & Braver, 2015). Illustrating this principle, the Motivational Intensity Theory, a classic theory of motivation (Brehm & Self, 1989; Silvestrini, 2017; Silvestrini et al., 2023), posits that effort investment is proportional to the importance of the outcome and the difficulty of the task. This implies a trade-off whereby individuals discount the benefits by the costs implied in obtaining them and, consequently, aim to minimize effort by selectively boosting it for a sufficiently valuable goal.

In the last decade, the idea of cost-benefit trade-offs was successfully married to the framework of reinforcement learning to explain how experience can drive learning of rewards as well as of costs (Sutton & Barto, 1998; Verguts, Vassena, & Silveti, 2015). In reinforcement learning, expectations are updated whenever an outcome is better or worse than expected – that is, a prediction error occurs. Importantly, by applying this learning to both costs and rewards, reinforcement learning can explain how decision-makers learn not only which actions lead to rewards but also how much effort they need to exert to obtain the reward, and what is the likelihood that the reward will arrive after completing the action. However, it is important to note that learning to optimize the effort involved in a task entails not only the monitoring of external rewards, but also a *meta-learning* mechanism that monitors and regulates the decision-maker's own internal state. This is because the effort involved in a task depends critically on internal computations entailed in performing the task – such as the difficulty of attending to relevant features, of planning ahead and/or generating vigorous actions – as well as on the decision-maker's levels of fatigue and arousal (Bijleveld,

2023; Dora, van Hooff, Geurts, Kompier, & Bijleveld, 2022; Matthews et al., 2023; Müller, Klein-Flügge, Manohar, Husain, & Apps, 2021). In turn, selecting the best decision strategy also requires tracking more complex features of the environment, such as volatility (i.e., how stable the environment is), average reward rate (i.e., how much reward is available in a given context), or the opportunity cost of time (i.e., whether time on this particular task is well spent or should better be allocated to an alternative task) (Kurzban, Duckworth, Kable, & Myers, 2013). In this framework, a decision-maker can adapt its decisions to the context, for example, by learning to be more flexible in a volatile situation, or by learning to exert more effort to obtain rewards in a favorable reward-rich environment.

A promising neuro-computational model of meta-level motivated behavior is the Reinforcement Meta-Learner (RML) developed by Silveti et al. (2018), which situates mathematically precise computations of meta-learning value and costs within biologically plausible neural circuits. The RML postulates that the dorsal anterior cingulate cortex receives dopaminergic inputs conveying the rate of rewards in a task and, upon perceiving a decline in rewards (a “need for control”), calls for a boost of noradrenaline from the locus coeruleus to enhance the efficiency of cognitive computations. However, a noradrenaline boost is perceived as a cost and the system learns through experience to choose the level of boost that maximizes rewards while minimizing the cost. This multi-level cost-benefit optimization allows a remarkable level of cross-validation and falsification across tasks, contexts, and modalities. For example, the RML can capture trade-offs in motivated behavior in the context of working memory, physical effort, or attentional effort driven by the need to gain information (Silveti, Lasaponara, Daddaoua, Horan, & Gottlieb, 2023; Silveti et al., 2018). The RML also reproduces the sensitivity to reward volatility, producing higher learning rates in volatile relative to stable environments – that is, specifically when quickly updating beliefs is beneficial given the situation at hand (Silveti, Seurinck, & Verguts, 2011, 2013). Finally, the RML conceptually squares with intriguing work in the motivation literature on persistence and giving up (goal disengagement; Gollwitzer, 2018; Kappes & Schattke, 2022), highlighting its ability to optimize effort exertion over longer time scales.

The RML thus offers a mathematically precise computation of the subjective benefits and costs involved in a task and implements this computation in a biologically plausible circuit. Because of its biological plausibility, the RML generates testable (falsifiable) predictions about brain and behavior, which less prone to typical pitfalls of verbal predictions such as oversimplification and lack of specificity. Crucially, the RML can be used to simulate the effects of impairments of the system on motivation. Motivational impairments are consistently observed across neuropsychiatric disorders (Caligiore, Silveti, D'Amelio, Puglisi-Allegra, & Baldassarre, 2020; Husain & Roiser, 2018; Silveti, Baldassarre, & Caligiore, 2019). A mechanistic understanding of the impaired computations may reveal dissociable underlying disease profiles that are virtually indistinguishable at the surface symptom levels, suggesting that motivation – if properly situated and specified – may be the key to capture clinically relevant phenotypes.

In sum, motivational constructs are characterized as emergent phenomena that stem from dynamic optimization of a cost-benefit trade-off of many decision-relevant variables within a reinforcement meta-learning framework (i.e., multivariate dynamic optimization). The meta-learning dimension allows

considering not only momentary simple trade-offs but explains how we flexibly adapt to our environment while considering our internal states. The meta-reinforcement learning framework (as implemented by the RML model) thus dismisses the “motivational homunculus,” in favor of a highly integrated, situated neurocomputational solution, whose building blocks are constructed based on existing and validated psychological and neurobiological knowledge (Silvetti et al., 2018, 2023), which constitute a significant advance toward the constructivist view advocated by M&J.

Financial support. Eliana Vassena was supported by an Open Competition Xs grant (NWO 406.XS.04.129) of the Netherlands Organisation for Scientific Research (NWO).

Competing interest. None.

References

- Bijleveld, E. (2023). The ebb and flow of cognitive fatigue. *Trends in Cognitive Sciences*, 27, 1109–1110.
- Brehm, J. W., & Self, E. A. (1989). The intensity of motivation. *Annual Review of Psychology*, 40, 109–131.
- Caligiore, D., Silvetti, M., D’Amelio, M., Puglisi-Allegra, S., & Baldassarre, G. (2020). Computational modeling of catecholamines dysfunction in Alzheimer’s disease at pre-plaque stage. *Journal of Alzheimer’s Disease*, 77, 275–290.
- Camerer, C. F. (2008). Neuroeconomics: Opening the gray box. *Neuron*, 60, 416–419.
- Dora, J., van Hooff, M. L. M., Geurts, S. A. E., Kompier, M. A. J., & Bijleveld, E. (2022). The effect of opportunity costs on mental fatigue in labor/leisure trade-offs. *Journal of Experimental Psychology: General*, 151, 695–710.
- Gollwitzer, P. M. (2018). The goal concept: A helpful tool for theory development and testing in motivation science. *Motivation Science*, 4, 185–205.
- Husain, M., & Roiser, J. P. (2018). Neuroscience of apathy and anhedonia: A transdiagnostic approach. *Nature Reviews Neuroscience*, 19, 470–484.
- Kappes, C., & Schattke, K. (2022). You have to let go sometimes: Advances in understanding goal disengagement. *Motivation and Emotion*, 46, 735–751.
- Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, 36(6), 661–679.
- Matthews, J., Pissaro, M. A., Jurgels, M., Müller, T., Vassena, E., Chong, T. T. J., & Apps, M. A. (2023). Computational mechanisms underlying the dynamics of physical and cognitive fatigue. *Cognition*, 240, 105603.
- Müller, T., Klein-Flügge, M. C., Manohar, S. G., Husain, M., & Apps, M. A. J. (2021). Neural and computational mechanisms of momentary fatigue and persistence in effort-based choice. *Nature Communications*, 12, 4593.
- Silvestrini, N. (2017). Psychological and neural mechanisms associated with effort-related cardiovascular reactivity and cognitive control: An integrative approach. *International Journal of Psychophysiology*, 119, 11–18.
- Silvestrini, N., Musslick, S., Berry, A. S., & Vassena, E. (2023). An integrative effort: Bridging motivational intensity theory and recent neurocomputational and neuronal models of effort and control allocation. *Psychological Review*, 130, 1081–1103.
- Silvetti, M., Seurinck, R., & Verguts, T. (2011). Value and prediction error in medial frontal cortex: Integrating the single-unit and systems levels of analysis. *Frontiers in Human Neuroscience*, 5, 75.
- Silvetti, M., Seurinck, R., & Verguts, T. (2013). Value and prediction error estimation account for volatility effects in ACC: A model-based fMRI study. *Cortex*, 49, 1627–1635.
- Silvetti, M., Vassena, E., Abrahamse, E., & Verguts, T. (2018). Dorsal anterior cingulate-brainstem ensemble as a reinforcement meta-learner. *PLoS Computational Biology*, 14, e1006370.
- Silvetti, M., Baldassarre, G., & Caligiore, D. (2019). A computational hypothesis on how serotonin regulates catecholamines in the pathogenesis of depressive apathy. In Cutsuridis, V. (Ed.), *Multiscale models of brain disorders* (pp. 127–134). Springer International Publishing. doi: 10.1007/978-3-030-18830-6_12
- Silvetti, M., Lasaponara, S., Daddaoua, N., Horan, M., & Gottlieb, J. (2023). A reinforcement meta-learning framework of executive function and information demand. *Neural Networks*, 157, 103–113.
- Sutton, R. S., & Barto, A. G. (1998) *Reinforcement learning: An Introduction* (Vol. 1). MIT press Cambridge.
- Verguts, T., Vassena, E., & Silvetti, M. (2015). Adaptive effort investment in cognitive and physical tasks: A neurocomputational model. *Frontiers in Behavioral Neuroscience*, 9, 57.
- Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive Affective & Behavioral Neuroscience*, 15, 395–415.

Definitional devils and detail: On identifying motivation as an animating dynamic

Rex A. Wright^{a,b*} , Simona Sciarra^c  and Giuseppe Pantaleo^c 

^aDepartment of Psychiatry and Behavioral Sciences, University of Texas Dell School of Medicine, Austin, TX, USA; ^bDepartment of Psychology, University of North Texas, Denton, TX, USA and ^cFaculty of Psychology, Vita-Salute San Raffaele University, Milan, Italy.

Rex.Wright@austin.utexas.edu

simona.sciarra@outlook.com

pantaleo.giuseppe@univr.it

<https://dellmed.utexas.edu/directory/rex-a-wright>

<https://www.univr.it/docenti/p/pantaleo-giuseppe>

*Corresponding author.

doi:10.1017/S0140525X2400044X, e48

Abstract

Murayama and Jach critically evaluate the idea that motivation is a dynamic that determines behavior and propose alternatively that it might be an emergent property that people construe through perceived regularities in experience and action. The critique has value but fails to appreciate the progress that has been made in moving beyond the idea of which the authors are critical.

Murayama and Jach (M&J) critically evaluate the idea that motivation is a dynamic that determines behavior and propose alternatively that motivation might be an emergent property that people construe through perceived regularities in experience and action, which themselves derive from underlying mental computations. Their critique is thought provoking and well taken in multiple respects. Moreover, their alternative proposal is well worth considering. However, in our view, the critique fails to appreciate the considerable progress that has been made in moving beyond the idea of which the authors are critical. The progress constrains the scope of the critique and suggests that motivation science can comfortably proceed assuming that motivation is more than a reflective construal.

M&J’s critique centers around the concern that when motivational constructs are identified as causal, they suffer a black-box problem. Conceptual black boxes can predict designated outcomes but they cannot tell us how the outcomes are generated. In other words, they cannot explain the outcomes that they predict. We endorse this concern but feel that it has been addressed to a greater degree than the authors might realize.

One way the concern has been addressed is through prior recognition of the black-box problem in the motivation sphere. Perhaps most visibly, the problem was recognized by Lewin (1931) in his landmark call for a transition from (static) Aristotelian thinking in psychology to (dynamic) Galilean thought. It also was recognized, for example, by Wicklund (1990) in his less well-known, but powerful, critique of “zero-variable” theories in psychology. Another way the concern has been addressed is through the development of motivational

theories that identify motivation not as an animating dynamic but rather as a state of goal-oriented animation. By casting motivation as something to be explained, instead of something that explains, these theories have allowed specification of compelling causal processes that generate subjective, physiological, and behavioral outcomes of interest.

Early examples of the motivational theories that do not suffer the black-box problem are Festinger's (1957) theory of cognitive dissonance and Atkinson's theory of achievement motivation (Atkinson, 1964; Atkinson & Feather, 1966). Festinger's theory articulated processes that lead people to alter their belief systems in predictable fashions. The theory has sometimes been understood to have assumed a driving (determinative) need for cognitive consistency. However, as M&J observe, it did not. Atkinson's theory articulated how emergent motives to achieve and avoid failure combine with expectancies of success to determine achievement striving. Like Festinger's theory, Atkinson's theory did not assume driving needs. Rather, it assumed trait-like tendencies to place different value on favorable and unfavorable performance outcomes. This stands contrary to M&J's suggestion that the theory is paradigmatic of "black box" reasoning.

More recent examples of theories that do not suffer the black-box problem are Wicklund and Gollwitzer's (1982) theory of symbolic self-completion and a general analysis of motivation that derives from Brehm's theories of motivation and emotion intensity (Brehm, 1999; Brehm & Self, 1989; Brehm, Wright, Solomon, Silka, & Greenberg, 1983). Symbolic self-completion theory articulates dynamic processes that influence tendencies to seek and display symbols of one's desired identity (Sciara, Contu, Regalia, & Gollwitzer, 2023; Sciara, Regalia, & Gollwitzer, 2022). The general analysis extends beyond specific goal pursuits (e.g., involving achievement or identity symbols) and is especially noteworthy here because it distinguishes motivational constructs that sometimes are muddled, and identifies mechanisms that can cause performers to be animated to different degrees and in different respects at different points in time (for relevant discussions, see also, e.g., Gollwitzer, 1990; Heckhausen & Gollwitzer, 1986).

In quick summary, as described by Wright (2016), the general analysis identifies *motives* as reasons to act that can vary in strength, or importance, and have the capacity to (1) be either active or inactive (quiescent), and (2) operate explicitly (consciously) or implicitly (non-consciously). When motives are active, they guide behavior; when they are inactive, they only hold potential for doing so. Motives are distinguished from *motivation* by virtue of their ability to be inactive. Active motives are states of motivation (animation), whereas inactive motives are not. Motive strength (importance) is proposed to be determined by a set of factors, including the perceived value of available (e.g., financial) incentives (e.g., 5 USD as compared to 500 USD) and perceived need with respect to those incentives (e.g., poverty as compared to wealth). The analysis holds that momentary *effort* (forceful exertion) can, but will not necessarily, correspond to motive strength and engenders physiological adjustments indicative of *energization* (energy mobilization). It also addresses the relationship between motive strength and *desire*, suggesting that the latter is not a simple linear function of the former.

Although we feel that the black-box concern that M&J raise has been addressed to a greater degree than they might realize,

this is not at all to say that we think their discussion of the concern has no value. Our thoughts in this regard are very much to the contrary. For one, we believe it is useful to draw attention to the black-box problem at intervals in hopes of discouraging new investigators from falling into old theoretical traps. For another, we believe that discussions along these lines draw attention to another serious problem in the field, the jingle-jangle terminological problem that M&J reference at points (Pekrun, 2023).

For a long while, motivation science has been conducted by scholars functioning in different academic units, such as business, economics, education, psychology, and neuroscience. This has been beneficial insofar as it has allowed questions to be addressed from diverse scholarly perspectives. However, it has been harmful insofar as it has fostered the development of insulated intellectual eco-systems that employ distinctive and, often confusing, lexicons. One contemporary challenge is to improve this state of affairs by developing a common language for discussing phenomena and processes of interest. Development of such a language would facilitate idea exchange and advance the scientific endeavor. Good science is fraught with definitional devils and detail. Linguistic congruency would address many of those that currently impede progress within the motivation scientific community.

Financial support. The work received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interest. None.

References

- Atkinson, J. W. (1964). *An introduction to motivation*. Van Nostrand.
- Atkinson, J., & Feather, N. (1966). *A theory of achievement motivation*. Wiley & Sons.
- Brehm, J. W. (1999). The intensity of emotion. *Personality and Social Psychology Review*, 3, 2–22. https://doi.org/10.1207/s15327957pspr0301_1
- Brehm, J. W., & Self, E. A. (1989). The intensity of motivation. *Annual Review of Psychology*, 40, 109–131. <https://doi.org/10.1146/annurev.ps.40.020189.000545>
- Brehm, J. W., Wright, R. A., Solomon, S., Silka, L., & Greenberg, J. (1983). Perceived difficulty, energization, and the magnitude of goal valence. *Journal of Experimental Social Psychology*, 19, 21–48. [https://doi.org/10.1016/0022-1031\(83\)90003-3](https://doi.org/10.1016/0022-1031(83)90003-3)
- Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.
- Gollwitzer, P. M. (1990). Action phases and mind-sets. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of social behavior* (Vol. 2, pp. 53–92). The Guilford Press.
- Heckhausen, H., & Gollwitzer, P. M. (1986). Information processing before and after the formation of an intent. In F. Klix & H. Hagendorf (Eds.), *In memoriam Hermann Ebbinghaus: Symposium on the structure and function of human memory* (pp. 1071–1082). Amsterdam: Elsevier/North Holland.
- Lewin, K. (1931). The conflict between Aristotelian and Galilean modes of thought in contemporary psychology. *Journal of General Psychology*, 5, 141–177. <https://doi.org/10.1080/00221309.1931.9918387>
- Pekrun, R. (2023). Jingle-Jangle fallacies in motivation science: Toward a definition of core motivation. In M. Bong, J. Reeve & S.-I. Kim (Eds.), *Motivation science: Controversies and insights* (pp. 52–58). Oxford University Press. <https://doi.org/10.1093/oso/9780197662359.001.0001>
- Sciara, S., Regalia, C., & Gollwitzer, P. M. (2022). Resolving incompleteness on social media: Online self-symbolizing reduces the orienting effects of incomplete identity goals. *Motivation Science*, 8, 268–275. <https://doi.org/10.1037/mot0000267>
- Sciara, S., Contu, F., Regalia, C., & Gollwitzer, P. M. (2023). Striving for identity goals by self-symbolizing on Instagram. *Motivation and Emotion*, 47, 965–989. <https://doi.org/10.1007/s11031-023-10039-w>
- Wicklund, R. A. (1990). *Zero-variable theories and the psychology of the explainer*. Springer.
- Wicklund, R. A., & Gollwitzer, P. M. (1982). *Symbolic self-completion*. Routledge.
- Wright, R. A. (2016). Motivation theory essentials: Understanding motives and their conversion into effortful goal pursuit. *Motivation and Emotion*, 40, 16–21. <https://doi.org/10.1007/s11031-015-9536-4>

The ins and outs of unpacking the black box: Understanding motivation using a multi-level approach

F. Wurm^{a,b}, I. J. M. van der Ham^{a,b} and

J. Schomaker^{a,b,*} 

^aHealth, Medical & Neuropsychology, Leiden University, Leiden, The Netherlands and ^bLeiden Institute for Brain and Cognition, Leiden, The Netherlands

f.r.wurm@fsw.leidenuniv.nl

c.j.m.van.der.ham@fsw.leidenuniv.nl

j.schomaker@fsw.leidenuniv.nl

<https://www.universiteitleiden.nl/en/staffmembers/franz-wurm#tab-1>

<https://www.universiteitleiden.nl/en/staffmembers/ineke-van-der-ham#tab-1>

<https://www.universiteitleiden.nl/en/staffmembers/judith-schomaker/publications#tab-1>

*Corresponding author.

doi:10.1017/S0140525X24000566, e49

Abstract

Although higher-level constructs often fail to explain the mechanisms underlying motivation, we argue that purely mechanistic approaches have limitations. Lower-level neural data help us identify “biologically plausible” mechanisms, while higher-level constructs are critical to formulate measurable behavioral outcomes when constructing computational models. Therefore, we propose that a multi-level, multi-measure approach is required to fully unpack the black box of motivated behavior.

Murayama and Jach (M&J) suggest that high-level constructs of motivation are often used to explain behavioral findings, but that the use of this abstract terminology jeopardizes our understanding of the actual mechanisms underlying these effects. The authors suggest that we should *unpack the black box* and look at motivated behavior as the outcome of mental computational processes. We agree that computational models are paramount for our understanding of motivation and other cognitive processes. However, we believe that a focus restricted to the mechanistic aspects is too limited, and that such a narrow focus poses a threat to (theoretical) advancements in the broader field of cognitive (neuro)science. Here, we identify crucial limitations of a purely mechanistic approach and propose ways to reconcile these.

We will discuss the limitations of the approach put forward by M&J by relying on a level-of-analysis approach (e.g., Bechtel & Richardson, 2010; Marr & Poggio, 1976; Sun, 2009). Notably, we will follow Marr’s (1982) framework for explaining complex information processing systems that involves three levels of explanation, suggesting that we need all three levels to gain a comprehensive understanding. These levels include: (1) A computational level addressing *what* the system does (i.e., what is the goal of the system), (2) the algorithmic level, relating to *how* the system achieves this, and (3) the implementation level relates to the *physical realization*. Although Marr formulated this framework in the context of visual processing, these levels have been applied over and above the originally intended domain. For example, the theory of reinforcement learning (RL) has been termed the “poster

child” of Marr’s framework as it spans all three levels from the computational goal of reward maximization to multiple algorithmic solutions and robust neural implementations in the brain (e.g., Niv & Langdon, 2016).

Following the rationale of a level-of-analysis approach, M&J criticize that current research on motivation is overly concerned with high-level computational accounts, while neglecting the algorithmic realization (i.e., “mental computational processes”). In line with Marr, we want to highlight the importance of describing a complex information system on *all* three levels to attain a comprehensive and complete understanding of how such a system works. Naturally, this also includes the implementation level.

One missing link in the proposed approach by M&J is the biological plausibility of the proposed mechanisms. Although claims of biological plausibility are often (rightly) labeled as “empty” and “inconsistent” (e.g., Love, 2021), we use the term to emphasize the need to test any proposed mechanisms against reality using appropriate behavioral and neural measures. Instead of limiting ourselves to just one level, a multi-level and multi-measure approach is required to provide a full perspective in order to not send us astray. We agree that if your aim is to mimic problem solving or mimic behavior – for example, as a computer scientist or robotics engineer – a purely algorithmic approach may be fruitful, as implementation can be achieved in various ways. However, as cognitive neuroscientists, we are actually strongly concerned with the exact implementation and neural mechanisms underlying cognitive, affective, and behavioral functions. Therefore, we must be careful in selecting computational models (Mars, Shea, Kolling, & Rushworth, 2012; Nassar & Frank, 2016; Nassar & Gold, 2013).

Noteworthy, an important method in our toolkit to support claims about cognitive functions is the falsification of computational models by testing specific predictions and identifying evidence that contradicts them (Palminteri, Wyart, & Koehlin, 2017). With algorithmic accounts of high-level concepts on the rise (e.g., Brielmann & Dayan, 2022; Gershman & Cikara, 2021; Shenhav et al., 2017) we propose such a falsification approach on the implementation to be extremely important. Again, the theory of RL provides a poignant example of where a purely mechanistic approach revealed its limitations. First, the representation of value is a central aspect in most RL models. Based on a plethora of model-driven neuroimaging studies, it was concluded that specific neurons or brain regions implement value-based RL algorithms. However, more recent and scrutinizing investigations suggest that behavioral and neural patterns are better explained by so-called policy-based RL algorithms (Hayden & Niv, 2021). The difference between these algorithms may seem subtle but has strong implications for theoretical accounts. Second, recent work even challenges the implementation of RL in the brain as such and proposes a model that can explain dopaminergic activity more readily (Jeong et al., 2022). This demonstrates that in order to make progress in the field of cognitive neuroscience, there is a necessity to link mechanistic algorithms to neural substrates, or else we may be explaining behavior, while failing to understand how the brain implements the solutions.

M&J acknowledge that other levels of explanation have merit, but they do not provide a framework on how to bridge and integrate these levels. We propose that for a mechanistic approach to be valuable, our models need be tested against empirical data. High-level constructs can act as tools to shape our thinking, to communicate our ideas to others, and define relevant input and output measures for our algorithms. Consistent with the suggestion of the authors, behavioral data are a first important step

from high-level concepts to low-level mechanisms. However, we should not stop there and continue to evaluate algorithms in light of neural measures. To reconcile the limitations of a purely mechanistic approach, we propose a multi-level, multi-measure approach. Lower-level information can help us identify “biologically plausible” models, and higher-level constructs can help us formulate measurable behavioral and neural outcomes when constructing computational models.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interest. None.

References

- Bechtel, W., & Richardson, R. C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT Press.
- Brielmann, A. A., & Dayan, P. (2022). A computational model of aesthetic value. *Psychological Review*, 129(6), 1319–1337. <https://doi.org/10.1037/rev0000337>
- Gershman, S. J., & Cikara, M. (2021). Structure learning principles of stereotype change. *PsyArXiv*, 2(1954), 1–27. <https://psyarxiv.com/52f9c/>
- Hayden, B. Y., & Niv, Y. (2021). The case against economic values in the orbitofrontal cortex (or anywhere else in the brain). *Behavioral Neuroscience*, 135(2), 192–201. <https://doi.org/10.1037/bne0000448>
- Jeong, H., Taylor, A., Floeder, J. R., Lohmann, M., Mihalas, S., Wu, B., ... Nambodiri, V. M. K. (2022). Mesolimbic dopamine release conveys causal associations. *Science*, 378(6626), eabq6740. <https://doi.org/10.1126/science.abq6740>
- Love, B. C. (2021). Levels of biological plausibility: Levels. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1815), 0–2. <https://doi.org/10.1098/rstb.2019.0632>
- Marr, D. (1982). *Vision: A computational approach*. Freeman & Co.
- Marr, D. C., & Poggio, T. (1976). *From understanding computation to understanding neural circuitry*. Massachusetts Institute of Technology, A.I. Memo 357.
- Mars, R. B., Shea, N. J., Kolling, N., & Rushworth, M. F. (2012). Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *Quarterly Journal of Experimental Psychology*, 65(2), 252–267. <https://doi.org/10.1080/17470211003668272>
- Nassar, M. R., & Frank, M. J. (2016). Taming the beast: Extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences*, 11, 49–54. <https://doi.org/10.1016/j.cobeha.2016.04.003>
- Nassar, M. R., & Gold, J. I. (2013). A healthy fear of the unknown: Perspectives on the interpretation of parameter fits from computational models in neuroscience. *PLoS Computational Biology*, 9(4), 1–6. <https://doi.org/10.1371/journal.pcbi.1003015>
- Niv, Y., & Langdon, A. J. (2016). Reinforcement learning with Marr. *Current Opinion in Behavioral Sciences*, 11(3), 67–73. <https://doi.org/10.1016/j.cobeha.2016.04.005>
- Palminteri, S., Wyart, V., & Koehlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, 21(6), 425–433. <https://doi.org/10.1016/j.tics.2017.03.011>
- Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, 40(1), 99–124. <https://doi.org/10.1146/annurev-neuro-072116-031526>
- Sun, R. (2009). Theoretical status of computational cognitive modeling. *Cognitive Systems Research*, 10(2), 124–140. <https://doi.org/10.1016/j.cogsys.2008.07.002>

Authors' Response

Response to the critiques (and encouragements) on our critique of motivation constructs

Kou Murayama^{a,b*} and Hayley Jach^{a,c}

^aHector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany; ^bResearch Institute, Kochi University of Technology, Kochi, Japan and ^cMelbourne School of Psychological Sciences,

The University of Melbourne, Parkville, VIC, Australia
k.murayama@uni-tuebingen.de; <https://motivationsciencelab.com/>
hayley.jach@unimelb.edu.au

*Corresponding author.

doi:10.1017/S0140525X24001353, e50

Abstract

The target article argued that motivation constructs are treated as black boxes and called for work that specifies the mental computational processes underlying motivated behavior. In response to critical commentaries, we clarify our philosophical standpoint, elaborate on the meaning of mental computational processes and why past work was not sufficient, and discuss the opportunities to expand the scope of the framework.

R1. Introduction

We sincerely appreciate the commentaries we received from a variety of disciplines for our target article which questioned the constructs of motivation explaining higher-order behavior (Murayama & Jach, 2024). It is a true pleasure as authors to see that our target article sparked robust discussion. This nevertheless brings challenges for any attempt to respond to such a heterogeneous set of opinions. In the responses, some showed endorsement with our argument whereas others exhibited strong disagreement; some indicated that our proposal to specify mental computational processes is not feasible whereas others pushed to go even further. Some made specific remarks on our reward-learning framework on knowledge acquisition whereas others discussed more general issues.

In the following, we made our best attempt to thoughtfully guide ourselves through the broad array of comments. We first attempted to correct some misunderstandings regarding the theoretical positioning (sect. R2), and then tackled the claims that our proposal has already been implemented in many existing motivation theories (sect. R3). We then discussed comments regarding the scope of what we called mental computational processes (sect. R4). Finally, we turned to various suggestions from commentators to build models of mental computational processes underlying high-level motivation constructs (sect. R5).

R2. Clarifying our theoretical positioning

Some commentaries (Moors; Sheldon & Ryan; Ozgan & Allen) rejected our theoretical position on the grounds that it is a reductionism and, therefore, *prima facie* untenable. We find it useful to respond to these comments first in order to clarify our theoretical positioning in the philosophy of mind. First of all, reductionism encompasses a wide range of perspectives and should not be dismissed solely on the basis of being reductionist. One extreme version of reductionism is reductionist atomism, which believes that the only scientific way to understand complex phenomenon is to analyze it into its component parts (Sawyer, 2002). Put in the context of our article, this position aims to *replace* or *eliminate* the constructs of motivation (i.e., higher-level explanation) by introducing mental computational processes (i.e., lower-level explanation). Sheldon & Ryan and Ozgan & Allen seem to interpret our proposal as the commitment to this extreme version of reductionism. But this is not our position (although we admit

that some sentences in our target article were misleading with this regard as noted by **Elliot & Sommet**). Our point is that theories have long taken motivation constructs for granted in the past and did not make sufficient effort to dig into the mental computational processes underlying motivated behavior. This is explicitly stated in the target article: “No level of understanding should be dismissed as ‘wrong’ (i.e., one level of explanation should not be replaced with a lower-level explanation ... but the problem of motivation literature is that most researchers are satisfied with higher-level explanations (i.e., supposing high-level motivation constructs to explain behavior) and little effort has been made to pursue lower-level explanations” (sect. 5.2).

At the same time, our position is, in fact, partly compatible with some forms of reductionism, because we argued that the function of motivation constructs (i.e., to cause behavior) is realized through lower-level mental computational processes. But this is not a controversial argument by itself – no scientist today doubts that everything in the mind is realized somehow through the brain and neural activities. If we accept this, it is unwarranted to dismiss our standpoint as reductionist. Importantly, if one rejects this standpoint by calling it reductionism, this could lead to holism, which holds that higher-order explanations are completely independent from lower-level explanations. This perspective is akin to dualism and does not have clear logical coherence in the current philosophy of mind (Sawyer, 2002).

To avoid holism, one must accept that higher-order processes emerge from lower-level units in some way, and we are arguing that a better understanding of the higher-level motivation constructs can be gained by investigating these lower-level processes. And there is more benefit than scientific knowledge and understanding: Better understanding the possible mechanisms of such a process can help to design better interventions. A helpful analogy may be drug trials. For many drugs, the process via which they have effects is a black box. Despite that, pharmaceutical medication is used to help people and save lives. But if we had a better understanding of the processes that lead to the drugs’ actions, we could create even better medication that would have greater effects and help more people.

Another point of clarification is that, while we agree that both levels (i.e., motivation constructs and mental computational processes) of explanation are important, it is the mental computational processes which directly cause so-called motivated behavior. Higher-level motivation constructs have a lot of utility to predict behavior (which we acknowledged in the target article) but are not the direct cause. **Jurjako** positioned such a position of ours as *mental fictionalism* in philosophy of mind (Toon & Toon, 2023) or *mental modelism* (see Crane & Farkas, 2022), which we appreciate and agree with. According to this standpoint, motivation constructs can be considered as a useful fiction or a hypothetical model in that they do not have direct causal effects on behavior but are used in daily narratives to explain causal effects. They also have, however, a significant role in theorizations, practical interventions, and understanding of human behavior in daily life settings. Jurjako used the equator, the average person, and the ideal gas law as examples of such mental fictions or models. This is critically different from the assertion that motivation is a post-hoc rationalization or epiphenomenon, void of meanings (e.g., Nisbett & Wilson, 1977).

Scarantino made a distinction between the inferences (or psychological construction) made by ordinary people and scientists. He argued that motivational constructs are the inferences made by scientists or experts, not by ordinary people; therefore, they

are useful as theoretical constructs. We actually agree (and we acknowledge that our example using ordinary people was misleading). He criticized that we did not provide any evidence that motivational constructs do not allow for interesting psychological generalizations, but we did not do so in the target article because we do believe that motivational constructs allow for generalizations and make useful theoretical predictions (see sect. 3.1 in the target article). But this is different from the assertion that motivation constructs have a direct causal effect. On this point we disagree. Scarantino stated “motivation concepts are plainly causal” by showing that intervening on higher-order motivation constructs changes behavior. **Sheldon & Ryan** also argued that the causal effects of needs are empirically demonstrated by interventions. However, even with interventions it is extremely challenging to conclude that the target constructs (e.g., need for autonomy) have a causal effect, when the target constructs are broad and not well-specified such as higher-order motivation constructs (Bailey et al., 2024). We can certainly establish the causal effects of intervention itself (e.g., attempts to change teachers’ autonomy support behavior) and the outcome, but it is a much harder job to demonstrate that the target motivation constructs intervene on the effect (see also Eronen, 2020).

Moors also indicated that our standpoint is reductionism in that we are trying to replace higher-order motivation constructs or goals with lower-order ones. Again, we do not. She takes the position that goals or motivation are hierarchically organized and questioned our proposal by indicating that higher-level goals play an important role in understanding lower-level goals or actions. Interestingly, **Dubourg, Chambon, & Baumard** (Dubourg et al.) brought in a similar idea of goal/motivation hierarchy from evolutionary psychology, but they saw it as consistent with our proposal. More specifically, they indicated that evolutionary psychology has seen it critical to specify lower-level variables in a way that is consistent with higher-level motivation. Our positioning is closer to Dubourg et al. Motivation constructs are hypothetical in that they are posited to conveniently explain patterns of behavior and subjective experiences, but they are still informative and play a critical instrumental role to think about mental computational processes. They are not completely detached in this regard – they inform each other.

However, one critical deviation from these “hierarchical models” (**Heckhausen & Rheinberg; Dubourg et al.; Sheldon & Ryan; Elliot & Sommet; Del Giudice**) is that we do not think these high-level “fundamental motivation” constructs (achievement motivation, affiliation motivation, etc.), which are allegedly created through evolution, have a top-down causal influence on mental computational processes and, thus, behavior. It may look like they do, but we do not need to think that way. It is sufficient to suppose that mental computational processes went through evolutionary processes. People show a broad range of social behavior not because evolution shaped a central motivation system ordering us to be social in general – it is these behaviors (and the processes that caused the behaviors) which were shaped by evolution. We are increasingly convinced by the idea based on recent accumulating neuroscientific literature suggesting that there is no single fixed brain area or system that is dedicated to a particular type of motivation or emotion (e.g., Meliss, Tsuchiyagaito, Byrne, van Reekum, & Murayama, 2024; Pessoa, 2017). If there is a top-down signal that orients various types of specific goal-directed behavior in a particular manner (e.g., in a manner that makes the organism competent), where does that come from? (see also commentaries by **Wurm, van der Ham,**

& Schomaker [Wurm et al.] for the importance of considering the constraints from the neural data). At least for now, we do not see good evidence of such non-specific, top-down signal of high-level motivation constructs.

Related to the comments on reductionism, André & Baumeister expressed a concern that we are attempting to reduce motivation to cognition. Other commentaries also equated mental computational processes with cognition (e.g., Moors; Sheldon & Ryan; Heckhausen & Rheinberg; Eccles & Wigfield). In fact, the history of motivation research has often been portrayed as the tension between motivation and cognition (Bem, 1967; Weiner, 1991), forming the sentiment that we should separate motivation from cognition. André & Baumeister especially indicated that motivation cannot be described by mental computational processes. Specifically, motivation is something inside us which makes rewards appealing; on the other hand, rewards cannot be appealing to computers because computers, as non-sentient objects, lack motivation. We are not sure if such a hard dichotomy helps us understand human functioning as a whole – the distinction between motivation and cognition is useful in many cases but is often blurred and creates unnecessary constraints in our explanation (Murayama, 2022b). Unless we take the position of dualism (which is generally rejected by scientific consensus and philosophy of mind), mental computational processes must logically mediate appealing feelings (i.e., a rewarding feeling). We can already observe some integration of mental computational processes and motivational properties in the literature: For example, “incentive salience” describes the feeling that something is rewarding and desirable (Bindra, 1974; Bolles, 1972), and researchers have investigated its computational basis (Berridge, 2023). And again, it is not the motivation itself which is shaped by evolution – we feel certain stimulus rewarding because evolutionary processes programmed us to feel that way. In fact, in research of information-seeking, there are some attempts to computationally understand *why* we feel particular stimuli/situations as appealing via simulations (e.g., Giron et al., 2023; Gruaz, Modirshanechi, & Brea, 2024) or rational analysis perspectives (Dubey & Griffiths, 2020; see also Ten, Oudeyer, Sakaki, & Murayama, 2024).

R3. Are there already many theories focusing on mental computational processes?

Several commentaries posited that motivation theories focusing on mental computational processes already exist, and are even the core agenda in motivation science (Custers, Eitam, & Higgins [Custers et al.]; Richter & Gendolla; Wright, Sciara, & Pantaleo [Wright et al.]; Heckhausen & Rheinberg; Eccles & Wigfield). Some authors are especially explicit in this regard: Heckhausen & Rheinberg stated “The good news is, there is no black-box problem and no need to reinvent the wheel,” and Richter & Gendolla claimed “Motivation science has always looked at filling the ‘black box’ by specifying constructs and explaining how behavioral tendencies are generated.”

We acknowledge that most of the examples suggested by these authors have tried to explain motivated behavior without resorting to the concepts of need or motivation. We appreciate the effort of these researchers to remove high-level concepts from explanation and the impact that their research has made in motivation science. However, we feel that many of these examples do not sufficiently address what we aspired for in the target article.

To respond to these commentaries, we first clarify what we meant by mental computational processes. As some commentaries correctly pointed out (Wurm et al.; Jurjako; Scarantino), we had Marr’s (1982) three level of analysis in mind when discussing mental computational processes. Marr proposed that, when analyzing the capacity of a system, we have three levels of questions: (1) The computational level, which asks what is the *function/nature* of the system, (2) the algorithmic level, which asks what are the *processes* by which the function is computed, and (3) the implementation level, which asks what is the *physical realization* of the process (van Rooij & Baggio, 2021). By mental computational processes, we meant the algorithmic level.¹ That is, beyond *what* it does (i.e., computational level), it should explain *how* it achieves what it does. Adopting the example by Cummins (2000; also cited by van Rooij & Baggio, 2021), let us consider a hypothetical psychological construct called *multiplication system*. The highest, computational level can explain that the function of this system is to multiply numbers. The algorithmic level can say that there are different ways to achieve the function – for example, the multiplication system may use a partial product algorithm or sequential addition. By clever empirical study, we can even tell which process is more likely to operate. For this example, by assessing reaction time for various numbers, we can test these two hypothesized processes realizing the multiplication system because sequential multiplication should take longer as a function of multiplier size.

With this distinction in mind, in our reading of the theories suggested by the commentators, while we acknowledge that these theories provide great insights into mental computational processes, they have two related issues.

First, some theories still rely on the concepts which are described at the computational level in our opinion. For example, several commentators (e.g., Del Giudice; Heckhausen & Rheinberg; Moors; Eccles & Wigfield; Wright et al.; Richter & Gendolla) mentioned various theories which include common terms in motivation research – expectancies, values, or goals (e.g., Atkinson, 1957; Brehm & Self, 1989; Eccles & Wigfield, 2020; Vroom, 1964). Here one common assumption is that people are more motivated to take action if the action has higher expectancy or value or fits with one’s goal. But this simply describes what it does (e.g., if you find X valuable, you will do Y). It does not explain the process underpinning the purpose. As noted in our original manuscript (sect. 5.4), there are many algorithmic level questions we can ask. For example, how does one find something valuable? Perhaps value is not a quantity but takes a form of mental representation – what, then, are the mechanisms of representational change and how can one translate multidimensional representation into subjective feeling of values? For example, (situated) expectancy-value theory (Eccles & Wigfield, 2020) posits that perceived value is formed by socio-cognitive factors (e.g., cultural milieu, goals, other’s beliefs and behaviors), but according to our perspective, specifying factors is important but not enough to unravel mental computational processes. The mechanism explaining how a factor influences these constructs is still a black box.

Vassena & Gottlieb provided a nice example in this regard. Effort and cost are critical components in many traditional motivation theories. Motivation intensity theory (Brehm & Self, 1989) argued that effort investment is proportional to the importance of the outcome and the difficulty of the task (Richter, Gendolla, & Wright, 2016). This is a simple but powerful theory explaining a variety of motivated behavior without using the concepts of motivation/needs, but it critically depends on the concepts of

task difficulty and outcome importance. However, finding the optimal amount of effort by learning task difficulty and outcome importance is not a trivial job, and we need to specify how people can achieve this. In their computational model (Silvestrini, Musslick, Berry, & Vassena, 2023), the authors showed that it is critical to incorporate meta-learning mechanisms in the mental computational processes to explain the relevant motivated behavior, especially in complex environments (e.g., environments with volatility). This example illustrates how we can unpack the black box of motivation theories which are not reliant on the concepts of needs or motivation.

Second, even when these theories could be classed as mechanistic, they tend to be underspecified relative to typical algorithmic-level explanations in the literature. For example, several commentators mentioned theories which adopt the basic idea of a goal/motivation hierarchy (Del Giudice; Elliot & Sommet; Moors; Dubourg et al.). For example, Del Giudice's General Architecture of Motivation (GAM) model provides a comprehensive picture of human motivated behavior and takes a hierarchical position: Higher-level motivational systems (i.e., organisms' core biological goals such as physical safety, mating, and offspring care) send emotional signals to (and receive feedback from) a lower-level instrumental goal pursuit system which manages narrower, more specific goals in an open-ended manner. We already provided a criticism of the assumption that higher-level motivation constructs influence lower-level goals in such hierarchical models (see sect. R2). Indeed, these models do not explain where these higher-level goals come from. But let us assume they do.² While we agree that such a hierarchical organization represents mechanisms to explain motivated behavior, there are many underspecified parts in the model. How are different levels of goals and actions organized or represented? What kind of information is carried from a high-level goal to the low-level ones? How do these goals constrain each other to produce a single action output? In the reinforcement-learning (reward-learning) literature, hierarchical reinforcement learning (Botvinick, 2012) provides a nice algorithmic framework to concretely pin down how agents can manage such an action/goal hierarchy; however, we do not yet see an implementation of similar frameworks in the motivation literature.

A similar point can be made for the whole trait theory (Fleeson & Jayawickreme, 2015) suggested by Ratchford & Jayawickreme. The theory assumes that environmental and cognitive factors as well as individual differences in how they are processed result in certain specific distributional patterns of affective, behavioral, and cognitive states. The theory suggests that these specific distributional patterns give rise to what we call traits (Fleeson & Jayawickreme, 2015). It is indeed consistent with our perspective that the theory treats traits simply as patterns of various states and indicates that they do not directly cause behaviors (although they also seem to treat traits as real on some occasions, which is reflected in some phrases like "trait enactment"). At the same time, how these environmental and cognitive factors interact remain unspecified. It is this process that our proposal called for in our target article; and promisingly, the investigation of these factors is now an emerging area in personality psychology (e.g., Horstmann, Rauthmann, Sherman, & Ziegler, 2020; Kuper et al., 2022; Roemer, Horstmann, & Ziegler, 2021).

We acknowledge that the issues are a matter of degree: We do not intend to say that the theories suggested by commentaries do not address mental computational processes at all. They may do so to some extent, but we can and should dig deeper. Some of

these theories can be a great first step toward this aim. For example, some commentators (Custers et al.; Elliot & Sommet) mentioned classic incentive theories of motivation (Bindra, 1974; Bolles, 1972; Toates, 1986). These theories are underspecified according to our perspective. At the same time, these theories can also be deemed as the foundation for contemporary reinforcement-learning theories (Dayan & Balleine, 2002), which we believe address mental computational processes to a much greater extent.

The reward-learning framework of knowledge acquisition (Murayama, 2022a), which we presented as an example model describing mental computational processes, is also underspecified. As indicated by some commentators (Sheldon & Ryan; Schuetz & Rutten; which we also noted in the target article), the framework has an implicit assumption that awareness of a knowledge gap initiates information-seeking behavior. But how can one be aware of a knowledge gap? How should we describe the knowledge representation – via belief states (Golman, Gurney, & Loewenstein, 2021) or a knowledge network (Murayama, 2022a; Sizemore, Karuza, Giusti, & Bassett, 2018)? How can we define the knowledge gap with that representation, and what kind of methods are used to compute it? As noted in the target article, this is a hot area in the field and we need to unpack the presented framework further in future studies.

It is also important to clarify that describing a theory in a mathematic form does not necessarily mean that it describes mental computational processes. Atkinson's expectancy and value theory of achievement motivation (Atkinson, 1957) mentioned by Heckhausen & Rheinberg, Richter & Gendolla, and Wright et al., for example, is a clear mathematical theory but it does not explain the nature of expectancy as we discussed earlier (and the resultant value, which is an inverse of expectancy). The model of cognitive energetics mentioned by Richter & Gendolla (Kruglanski et al., 2012), which adopted Lewin's (1942) force-field approach, is a theory which explains social judgement and self-control with the concepts of driving force and restraining force. The model is described in mathematical forms, but like Lewin's formulation, the equations are described at a very general level. In addition, these mathematical theories critically lack time dynamics describing how the concepts causally influence each other over time. According to our view, any mental computational processes described in these theories are still underspecified.

R4. Should we push more?

Another set of commentaries challenge our article by indicating that the proposal we put forth (i.e., to specify mental computational processes), as it stands, is not enough and we need to step further. Gernigon, Altamore, Vallacher, van Geert, & Den Hartigh (Gernigon et al.) argued that the reward-learning framework of knowledge acquisition is driven by component-dominant dynamics which are fundamentally different from the interaction-dominant dynamics of the dynamic system approach (Den Hartigh, Cox, & van geert, 2017; Van Orden, Holden, & Turvey, 2003). They argued that what we called "emergent property" is not, strictly speaking, emergent, because we specify the factors and causal relations underlying the phenomenon. Ozgan & Allen made a similar point by considering our stance as a substance-oriented framework.

Indeed, the reward-learning model of knowledge acquisition is not a dynamic systems model in a strict sense, which is

characterized by the interaction of elements producing phenomena that cannot be predicted by examining the elements themselves. By looking at the specified reward-learning mechanisms, it is easy to see how an agent acts as if it had the need for competence. But it is important to emphasize that we presented the framework as one example of how mental computational processes can be specified. Regardless of whether the proposed mechanisms are component- or interaction-dominant, we welcome models and frameworks that seek to specify computational mechanisms underlying motivated behavior. Perhaps one important question is how much of motivated behavior can be explained by interaction-oriented dynamics (i.e., dynamic systems phenomena) and how much can be explained by component-dominant dynamics. There are certainly phenomena which can be better captured by dynamic systems perspective (Gernigon, Vallacher, Nowak, & Conroy, 2015; Kaplan & Garner, 2017; Laskar & van der Maas, 2024), but we do not believe that this perspective encompasses all motivation constructs. This can be tested if more work emerges focusing on mental computational processes in the future.

Alexander also provides a similar critical comment, although from a different perspective. Specifically, she suggested that the mental computational processes as illustrated by our reward-learning framework assume linearity and directionality, disregarding the dynamic and complicated nature of mental functioning. As noted above, we describe mental computational processes rather broadly, and they can accommodate complexity, dynamics, and non-linearity. The reward-learning model of knowledge acquisition is just one example, and we do not limit ourselves to it. However, unlike Gernigon et al., Alexander seems to have a more pessimistic view of whether we can truly specify such mental computational processes to explain motivated behavior (see also Heckhausen & Rheinberg). We do agree that mental computational processes underlying behavior are complicated and dynamic, especially when we leave the laboratory and attempt to study real-life behaviors. At the same time, recent years have seen a dramatic rise of modelling approaches in cognitive science and increased availability of real-life data via digital technologies (Allen et al., 2024). We feel that the time is ripe to take this bold step to further our understanding of motivated behavior in real life, rather than accumulating empirical evidence solely using broader motivational constructs. Our proposal aimed to encourage scholars to take such an endeavor.

Schuetze & Rutten brought another important perspective to our proposal – it is not the accuracy of mental computational processes *per se* that defines the best model but how useful the model is in practice. However accurate a model is, if the model is not interpretable, it is not useful for practitioners or policymakers. They provided interesting examples from the field of education, in which two prominent education researchers, Carroll (1963) and Bloom (1976), pushed forward computational process models of school learning but attracted minimum attention due to these models' complexity (Harnischfeger & Wiley, 1978).³ We partly agree and indeed, investigating increasingly fine-grained levels of analysis may not be helpful to think about the best intervention for the phenomenon. At the same time, we believe that identifying the mental computational processes in many cases could provide the right bite-sized chunks for practice – not too broad, but not too intricate. When we say that the need for autonomy is important for well-being, self-determination theory provides several important practical suggestions, such as providing choices, encouraging self-initiation, and offering rationales for why

autonomy is important (e.g., Pintrich & Schunk, 2002). But these practices themselves do not offer a universal solution (hence why it is difficult to promote people's well-being!). To find a more effective intervention, we need to understand how the provision of choices leads to increased well-being by identifying the underlying processes that give rise to the feeling of autonomy (see also Yan, Sana, & Carvalho, 2024). Even the complex systems perspective, which has been criticized for its lack of usefulness for applications, could provide valuable insights into when and under what conditions an intervention works (Gernigon, Den Hartigh, Vallacher, & van Geert, 2024; van der Maas, 2024; see also the concept of process causality suggested by Gernigon et al. and Ozgan & Allen).

Schuetze & Rutten's commentary provides an interesting contrast to the commentary by Wurm et al. They demanded that we should also consider the physiological implementation of the mental computational processes, that is, neural mechanisms (see also the commentary by Vassena & Gottlieb). They argued that neural level analyses can be an empirical tool to falsify or modify the proposed mental computational mechanisms, indicating the importance of taking multi-level perspective, considering different levels of analysis altogether. In a similar vein, Spurrett argued that, when we consider mental computational processes, we cannot avoid the role played by the bodies and actions (and their neural implementations). In fact, all the motivated decision making comes down to physical actions which compete with each other, which places substantial limitations on what we can do at a time. For us, such bodily constraints are also part of the level of physiological implementation by Marr (1982).

We agree with the point, and this can be also a great response to Schuetze & Rutten's commentary. Even when a certain theory turns out to be useful in practice, lower-level analysis still serves as a tool to empirically constrain theory (Marr, 1982). At the same time, we also feel that seeking to understand brain mechanisms (and associated bodily mechanisms) adds further complexity to our already-challenging endeavor, especially given the limitations of currently available neuroscientific methods. We acknowledge that these methods have made substantial progress in recent years, but many challenges remain before we can specify the neural mechanisms underlying motivated behavior.

R5. Critical factors when considering mental computational mechanisms

Several commentaries raised additional critical factors concerning theories of mental computational mechanisms underlying high-level motivation constructs. We appreciate the suggestions. This kind of exchange would have been impossible if we stopped our thinking at the higher-level motivation construct, and illustrates the fruitfulness of our suggested direction.

van Lieshout, Zhang, Friston, & Bekkering (van Lieshout et al.) suggested that integrating the predictive processing framework would enrich our understanding of motivated behavior. These commentators state that the predictive processing framework encompasses all aspects of sensory experience, not a mere reward function, and agents choose actions according to the expected free energy – to maximize prediction of the world (Friston & Kiebel, 2009). Reber, Haugen, & Martinussen (Reber et al.) also indicated the utility of this framework (but see the commentary by Moors for a critical remark). The reward-learning framework of knowledge acquisition is not inconsistent with the predicting processing framework

(Fitzgibbon & Murayama, 2022). Although not explained in the target article, Murayama (2022a) indicated that “knowledge” or “information” is defined broadly, including perceptual or sensory information. Importantly, unlike other major models of information-seeking, the framework features a “knowledge base,” which represents all the past experiences of the agent (the “kind of thing that I am” according to their terminology). The knowledge base serves as the basis for prediction, and by taking actions that reduce expected uncertainty, the agent tries to construct the optimal world model, that is, expand and improve the knowledge base. At the same time, one critique we offer of the predictive processing framework is its computational tractability (Gottlieb & Oudeyer, 2018; Ten et al., 2024) – given the tremendous amount of experiences we accumulate over development, how can we efficiently calculate the expected information gain at every moment? To understand mental computational processes underlying the concept of need for competence, perhaps we also need to find concrete heuristics people take to master their environment (Ten et al., 2024) or think seriously about how our knowledge is represented (Murayama, 2022a).

Bunzeck & Haesler argued that novelty is a key driver to explain exploration (and other motivated) behavior. Novelty and uncertainty are similar concepts but they can be distinguished via mental computational processes (Modirshanechi, Lin, Xu, Herzog, & Gerstner, 2023b; Poli, O'Reilly, Mars, & Hunnius, 2024). But novelty is not the only factor for exploration. In fact, one of the interesting aspects of information-seeking behavior is that people tend to become more and more interested as they acquire *more* knowledge (Alexander, Jetton, & Kulikowich, 1995; Fastrich & Murayama, 2020; Singh & Murayama, 2024; Witherby & Carpenter, 2022). This is because accumulated knowledge makes people aware the things they are not certain about (Murayama, 2022a; Murayama, FitzGibbon, & Sakaki, 2019). This means that the need for competence is likely to be governed by multiple processes such as uncertainty reduction, novelty seeking (Bunzeck & Duzel, 2006), and savoring (Kobayashi, Ravaioli, Baranès, Woodford, & Gottlieb, 2019). How we weigh these different processes and integrate them depending on the contexts is an important area for future research (Modirshanechi, Kondrakiewicz, Gerstner, & Haesler, 2023a; Poli et al., 2024).

Reber et al. proposed that subjective metacognitive feelings play a critical role in motivated behavior and should be actively incorporated in specifying mental computational processes. A similar point was made by **Del Giudice's** GAM model, in which affect serves a critical interface between the higher-order and lower-order goals. While our proposal put subjective experiences outside of the mental computational processes (Fig. 1 in the target article), as implied by the arrow from subjective experiences to mental computational processes, we did not preclude the possibility that subjective experiences modulate mental computational processes. In fact, in our work of metamotivation, we showed that people often have the wrong metacognition about how motivation functions and take actions that are not adaptive for motivation (Hatano, Ogulmus, Shigemasa, & Murayama, 2022; Kim, Sakaki, & Murayama, 2024; Kuratomi, Johnsen, Kitagami, Hatano, & Murayama, 2023; Murayama, Kitagami, Tanaka, & Raw, 2016). One critical question in this regard is, what are the mental computational processes that give rise to these metacognitive feelings? If these metacognitive feelings are calculated by relatively simple algorithms such as familiarity or fluency (Reber, Winkielman, & Schwarz, 1998), then it is possible that they serve as important heuristics for us to efficiently reduce

uncertainty in our knowledge (see also our response to **van Lieshout et al.**). In fact, Shenhav (2024) recently argued that the “goal” concept in decision-making literature may be an emergent property which is produced by individuals' affective associations.

Ainslie pushed the reward-learning framework further and discussed the nature of intrinsic or endogenous rewards (Ainslie, 2013), and how we manage them. Indeed, when unpacking the black box of motivational constructs according to the reward-learning framework, it is imperative to unpack the computational mechanisms that give rise to endogenous rewards. Ainslie also argued that for information gain (uncertainty reduction) to be regarded as rewards, they should (a) perform like rewards that have been studied in other contexts, (b) have a variable effect over a time course, and (c) depend on some kind of appetite. Temporal change in the rewarding value of information gain has been studied in empirical studies (Hsiung, Poh, Huettel, & Adcock, 2023; Noordewier & van Dijk, 2015) but for other aspects, we still know little. We agree that this is an important area for research in the future.

Acknowledgement. This research was supported by the Alexander von Humboldt Foundation (the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research; to Kou Murayama).

Notes

1. We acknowledge that the name “mental computational processes” was confusing as it includes the term computational, which Marr used for the first level. But we used the term computational to be related to “computational modeling” in cognitive science/neuroscience.
2. In fact, we feel this assumption is tenable for more specific types of goals which are made salient by the environment. Please be reminded that the target article criticizes motivation concepts that explain broad range of behaviors (“high-level motivation constructs”).
3. We would like to add Campbell and Frey (1970) as another great example in education.

References

- Ainslie, G. (2013). Grasping the impalpable: The role of endogenous reward in choices, including process addictions. *Inquiry*, 56(5), 446–469. <https://doi.org/10.1080/0020174X.2013.806129>
- Alexander, P. A., Jetton, T. L., & Kulikowich, J. M. (1995). Interrelationship of knowledge, interest, and recall: Assessing a model of domain learning. *Journal of Educational Psychology*, 87(4), 559–575. <https://doi.org/10.1037/0022-0663.87.4.559>
- Allen, K., Brändle, F., Botvinick, M., Fan, J. E., Gershman, S. J., Gopnik, A., ... Schulz, E. (2024). Using games to understand the mind. *Nature Human Behaviour*, 8(6), 1035–1043. <https://doi.org/10.1038/s41562-024-01878-9>
- Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review*, 64(6, Pt.1), 359–372. <https://doi.org/10.1037/h0043445>
- Bailey, D. H., Jung, A. J., Beltz, A. M., Eronen, M. I., Gische, C., Hamaker, E. L., ... Murayama, K. (2024). Causal inference on human behaviour. *Nature Human Behaviour*, 8(8), 1448–1459. <https://doi.org/10.1038/s41562-024-01939-z>
- Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, 74(3), 183–200. <https://doi.org/10.1037/h0024835>
- Berridge, K. C. (2023). Separating desire from prediction of outcome value. *Trends in Cognitive Sciences*, 27(10), 932–946. <https://doi.org/10.1016/j.tics.2023.07.007>
- Bindra, D. (1974). A motivational view of learning, performance, and behavior modification. *Psychological Review*, 81(3), 199–213. <https://doi.org/10.1037/h0036330>
- Bloom, B. S. (1976). *Human characteristics and school learning*. (pp. xii, 284). McGraw-Hill.
- Bolles, R. C. (1972). Reinforcement, expectancy, and learning. *Psychological Review*, 79(5), 394–409. <https://doi.org/10.1037/h0033120>
- Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, 22(6), 956–962. <https://doi.org/10.1016/j.conb.2012.05.008>

- Brehm, J. W., & Self, E. A. (1989). The intensity of motivation. *Annual Review of Psychology*, 40, 109–131. <https://doi.org/10.1146/annurev.ps.40.020189.000545>
- Bunzeck, N., & Duzel, E. (2006). Absolute coding of stimulus novelty in the human substantia nigra/VTA. *Neuron*, 51(3), 369–379. <https://doi.org/10.1016/j.neuron.2006.06.021>
- Campbell, D. W., & Frey, P. W. (1970). The implications of learning theory for the fade-out of gains from compensatory education. In J. Hellmuth (Ed.), *Compensatory education: A national debate: Vol. 3. Disadvantaged child* (pp. 455–463). Brunner/Mazel.
- Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, 64(8), 723–733. <https://doi.org/10.1177/016146816306400801>
- Crane, T., & Farkas, K. (2022). Mental fact and mental fiction. In T. Demeter, T. Parent, & A. Toon (Eds.), *Mental fictionalism: Philosophical Explorations* (pp. 303–319). Routledge.
- Cummins, R. (2000). “How does it work?” versus “what are the laws?”: Two conceptions of psychological explanation. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 117–144). The MIT Press.
- Dayan, P., & Balleine, B. W. (2002). Reward, motivation, and reinforcement learning. *Neuron*, 36(2), 285–298. [https://doi.org/10.1016/S0896-6273\(02\)00963-7](https://doi.org/10.1016/S0896-6273(02)00963-7)
- Den Hartigh, R., Cox, R., & van Geert, P. (2017). Complex versus complicated models of cognition. In L. Magnani & T. Bertolotti (Eds.), *Springer handbook of model-based science* (pp. 657–669). Springer. <https://doi.org/10.1007/978-3-319-30526-4>
- Dubey, R., & Griffiths, T. L. (2020). Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review*, 127(3), 455–476. <https://doi.org/10.1037/rev0000175>
- Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, 61, 101859. <https://doi.org/10.1016/j.cedpsych.2020.101859>
- Eronen, M. I. (2020). Causal discovery and the problem of psychological interventions. *New Ideas in Psychology*, 59, 100785. <https://doi.org/10.1016/j.newideapsych.2020.100785>
- Fastrich, G. M., & Murayama, K. (2020). Development of interest and role of choice during sequential knowledge acquisition. *AERA Open*, 6(2), 2332858420929981. <https://doi.org/10.1177/2332858420929981>
- Fitzgibbon, L., & Murayama, K. (2022). Counterfactual curiosity: Motivated thinking about what might have been. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1866), 20210340. <https://doi.org/10.1098/rstb.2021.0340>
- Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, 56, 82–92. <https://doi.org/10.1016/j.jrp.2014.10.009>
- Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521), 1211–1221. <https://doi.org/10.1098/rstb.2008.0300>
- Gernigon, C., Vallacher, R., Nowak, A., & Conroy, D. (2015). Rethinking approach and avoidance in achievement contexts: The perspective of dynamical systems. *Review of General Psychology*, 19, 443–457.
- Gernigon, C., Den Hartigh, R. J. R., Vallacher, R. R., & van Geert, P. L. C. (2024). How the complexity of psychological processes reframes the issue of reproducibility in psychological science. *Perspectives on Psychological Science*, 19(6), 952–977. <https://doi.org/10.1177/17456916231187324>
- Giron, A. P., Ciranka, S., Schulz, E., van den Bos, W., Ruggeri, A., Meder, B., & Wu, C. M. (2023). Developmental changes in exploration resemble stochastic optimization. *Nature Human Behaviour*, 7(11), 1955–1967. <https://doi.org/10.1038/s41562-023-01662-1>
- Golman, R., Gurney, N., & Loewenstein, G. (2021). Information gaps for risk and ambiguity. *Psychological Review*, 128(1), 86–103. <https://doi.org/10.1037/rev0000252>
- Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, 19(12), Article 12. <https://doi.org/10.1038/s41583-018-0078-0>
- Gruaz, L., Modirshanechi, A., & Brea, J. (2024). *Merits of curiosity: A simulation study*. OSF. <https://doi.org/10.31234/osf.io/evm9n>
- Harnischfeger, A., & Wiley, D. E. (1978). Conceptual issues in models of school learning. *Journal of Curriculum Studies*, 10(3), 215–231. <https://doi.org/10.1080/0022027780100304>
- Hatano, A., Ogulmus, C., Shigemasa, H., & Murayama, K. (2022). Thinking about thinking: People underestimate how enjoyable and engaging just waiting is. *Journal of Experimental Psychology: General*, 151(2), 3213–3229. <https://doi.org/10.1037/xge0001255>
- Horstmann, K. T., Rauthmann, J. F., Sherman, R. A., & Ziegler, M. (2020). Unveiling an exclusive link: Predicting behavior with personality, situation perception, and affect in a preregistered experience sampling study. *Journal of Personality and Social Psychology*, 120(5), 1317–1343. <https://doi.org/10.1037/pspp0000357>
- Hsiung, A., Poh, J.-H., Huettel, S. A., & Adcock, R. A. (2023). Curiosity evolves as information unfolds. *Proceedings of the National Academy of Sciences*, 120(43), e2301974120. <https://doi.org/10.1073/pnas.2301974120>
- Kaplan, A., & Garner, J. K. (2017). A complex dynamic systems perspective on identity and its development: The dynamic systems model of role identity. *Developmental Psychology*, 53(11), 2036–2051. <https://doi.org/10.1037/dev0000339>
- Kim, S., Sakaki, M., & Murayama, K. (2024). Metacognition of curiosity: People underestimate the seductive lure of non-instrumental information. *Psychonomic Bulletin & Review*, 31(3), 1–12. <https://doi.org/10.3758/s13423-023-02404-0>
- Kobayashi, K., Ravaioi, S., Baranès, A., Woodford, M., & Gottlieb, J. (2019). Diverse motives for human curiosity. *Nature Human Behaviour*, 3(6), 587–595. <https://doi.org/10.1038/s41562-019-0589-3>
- Kruglanski, A. W., Bélanger, J. J., Chen, X., Köpetz, C., Pierro, A., & Mannetti, L. (2012). The energetics of motivated cognition: A force-field analysis. *Psychological Review*, 119(1), 1–20. <https://doi.org/10.1037/a0025488>
- Kuper, N., Breil, S. M., Horstmann, K. T., Roemer, L., Lischetzke, T., Sherman, R. A., ... Rauthmann, J. F. (2022). Individual differences in contingencies between situation characteristics and personality states. *Journal of Personality and Social Psychology*, 123(5), 1166–1198. <https://doi.org/10.1037/pspp0000435>
- Kuratomi, K., Johnsen, L., Kitagami, S., Hatano, A., & Murayama, K. (2023). People underestimate their capability to motivate themselves without performance-based extrinsic incentives. *Motivation and Emotion*, 47, 509–523. <https://doi.org/10.1007/s11031-022-09996-5>
- Laskar, P., & van der Maas, H. L. (2024). A reciprocal-practice-success (RPS) model of free practice. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46, 2818–2824. <https://escholarship.org/uc/item/98s6q1bc>
- Lewin, K. (1942). Field theory and learning. *Teachers College Record*, 43(10), 215–242. <https://doi.org/10.1177/016146814204301006>
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman and Company.
- Meliss, S., Tsuchiyagaito, A., Byrne, P., van Reekum, C., & Murayama, K. (2024). Broad brain networks support curiosity-motivated incidental learning of naturalistic dynamic stimuli with and without monetary incentives. *Imaging Neuroscience*, 2, 1–27. https://doi.org/10.1162/imag_a_00134
- Modirshanechi, A., Kondrakiewicz, K., Gerstner, W., & Haesler, S. (2023a). Curiosity-driven exploration: Foundations in neuroscience and computational modeling. *Trends in Neurosciences*, 46(12), 1054–1066. <https://doi.org/10.1016/j.tins.2023.10.002>
- Modirshanechi, A., Lin, W.-H., Xu, H. A., Herzog, M. H., & Gerstner, W. (2023b). *The curse of optimism: A persistent distraction by novelty* (p. 2022.07.05.498835). bioRxiv. <https://doi.org/10.1101/2022.07.05.498835>
- Murayama, K. (2022a). A reward-learning framework of knowledge acquisition: An integrated account of curiosity, interest, and intrinsic–extrinsic rewards. *Psychological Review*, 129(1), 175–198. <https://doi.org/10.1037/rev0000349>
- Murayama, K. (2022b). Are cognition, motivation, and emotion the same or different?: Let's abandon that thinking. In M. Bong, S. Kim, & J. Reeve (Eds.), *Motivation science: Controversies and insights* (pp. 243–245). Oxford University Press.
- Murayama, K., & Jach, H. (2024). A critique of motivation constructs to explain higher-order behavior: We should unpack the black box. *Behavioral and Brain Sciences*, 1–53. <https://doi.org/10.1017/S0140525X24000025>
- Murayama, K., Kitagami, S., Tanaka, A., & Raw, J. A. L. (2016). People's naiveté about how extrinsic rewards influence intrinsic motivation. *Motivation Science*, 2(3), 138–142. <https://doi.org/10.1037/mot0000040>
- Murayama, K., FitzGibbon, L., & Sakaki, M. (2019). Process account of curiosity and interest: A reward-learning perspective. *Educational Psychology Review*, 31(4), 875–895. <https://doi.org/10.1007/s10648-019-09499-9>
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259. <https://doi.org/10.1037/0033-295X.84.3.231>
- Noordewier, M. K., & van Dijk, E. (2015). Curiosity and time: From not knowing to almost knowing. *Cognition and Emotion*, 31(3), 411–421. <https://doi.org/10.1080/02699931.2015.1122577>
- Pessoa, L. (2017). A network model of the emotional brain. *Trends in Cognitive Sciences*, 21(5), 357–371. <https://doi.org/10.1016/j.tics.2017.03.002>
- Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications* (2nd ed.). Merrill-Prentice Hall.
- Poli, F., O'Reilly, J. X., Mars, R. B., & Hunnius, S. (2024). Curiosity and the dynamics of optimal exploration. *Trends in Cognitive Sciences*, 28(5), 441–453. <https://doi.org/10.1016/j.tics.2024.02.001>
- Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science*, 9(1), 45–48. <https://doi.org/10.1111/1467-9280.00008>
- Richter, M., Gendolla, G. H. E., & Wright, R. A. (2016). Three decades of research on motivational intensity theory: What we have learned about effort and what we still don't know. *Advances in Motivation Science*, 3, 149–186.
- Roemer, L., Horstmann, K. T., & Ziegler, M. (2021). Sometimes hot, sometimes not: The relations between selected situational vocational interests and situation perception. *European Journal of Personality*, 35(2), 212–233. <https://doi.org/10.1002/per.2287>
- Sawyer, R. K. (2002). Emergence in psychology: Lessons from the history of non-reductionist science. *Human Development*, 45(1), 2–28. <https://doi.org/10.1159/000048148>
- Shenhav, A. (2024). The affective gradient hypothesis: An affect-centered account of motivated behavior. *Trends in Cognitive Sciences*, 28(12), 1089–1104. <https://doi.org/10.1016/j.tics.2024.08.003>

- Silvestrini, N., Musslick, S., Berry, A. S., & Vassena, E. (2023). An integrative effort: Bridging motivational intensity theory and recent neurocomputational and neuronal models of effort and control allocation. *Psychological Review*, 130(4), 1081–1103. <https://doi.org/10.1037/rev0000372>
- Singh, A., & Murayama, K. (2024). Creativity is motivated by novelty. Curiosity is triggered by uncertainty. *Behavioral and Brain Sciences*, 47, e115. <https://doi.org/10.1017/S0140525X23003291>
- Sizemore, A. E., Karuza, E. A., Giusti, C., & Bassett, D. S. (2018). Knowledge gaps in the early growth of semantic feature networks. *Nature Human Behaviour*, 2(9), 682–692. <https://doi.org/10.1038/s41562-018-0422-4>
- Ten, A., Oudeyer, P.-Y., Sakaki, M., & Murayama, K. (2024). *The curious U: Integrating theories linking knowledge and information-seeking behavior*. OSF. <https://doi.org/10.31234/osf.io/s8mkj>
- Toates, F. M. (1986). *Motivational systems*. Cambridge University Press.
- Toon, A., & Toon, A. (2023). *Mind as metaphor: A defence of mental fictionalism*. Oxford University Press.
- van der Maas, H. L. J. (2024). *Complex-systems research in psychology*. SFI Press.
- Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, 132(3), 331–350. <https://doi.org/10.1037/0096-3445.132.3.331>
- 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>
- Vroom, V. H. (1964). *Work and motivation*. Wiley.
- Weiner, B. (1991). Metaphors in motivation and attribution. *American Psychologist*, 46(9), 921–930.
- Witherby, A. E., & Carpenter, S. K. (2022). The rich-get-richer effect: Prior knowledge predicts new learning of domain-relevant information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(4), 483–498. <https://doi.org/10.1037/xlm0000996>
- Yan, V. X., Sana, F., & Carvalho, P. F. (2024). No simple solutions to Complex problems: Cognitive science principles can guide but not prescribe educational decisions. *Policy Insights from the Behavioral and Brain Sciences*, 11(1), 59–66. <https://doi.org/10.1177/23727322231218906>