Explaining Mechanism-Task Fit in Neuroscience[1]

Aliya Rumana

Center for Philosophy of Science, University of Pittsburgh

David Marr famously argued that computational theory (i.e., analysis at the computational level) was required to explain both "what the device does and why". In a series of papers, Oron Shagrir and William Bechtel argue that computational theory explains how certain mechanisms are appropriate for certain tasks by showing that identity holds between the corresponding mechanisms and the tasks at an abstract, computational level of description. Call this the "computational identity account" of "mechanism-task fit" (or "M/T fit"). Inspired by their work, I propose an alternative account that grounds M/T fit in constraint satisfaction, where the mechanism is appropriate to the task because the mechanism's properties satisfy all the task-related constraints. I use retinal edge detection and sound localisation as two cases to demonstrate that constraint satisfaction may be a better way to ground M/T fit than identity. This account of M/T fit isn't confined to the computational level of description, so I describe it as "task-fitting explanation" rather than computational theory. I argue that task-fitting explanation is a species of constraint-based explanation: it is interested in which features of a mechanism *make possible* above-chance correct task performance for the mechanism. As such, it is "modally complementary" to mechanistic explanation, which, I argue, is interested in which activities done by a mechanism's parts *make actual* competent task performance for the mechanism.

Keywords: David Marr; computational level; task analysis; mechanism; constraint; how-possibly explanation

Between the 1950s and 1970s, innovations in single-unit microelectrode recording *in vivo* led to significant progress in mapping the flow of visual information from the retina through primary visual cortex and down the visual stream. Some were optimistic that single-unit recordings would suffice to completely reveal the mechanistic basis of vision (Barlow, 1972). In his influential book *Vision* (1982), though, David Marr argues that a complete description of these mechanisms would be insufficient for a complete explanation of vision. Marr takes a Barlow-style, cell-level conception of mechanistic explanation for granted, but confines it to an implementational level of analysis, and adds that computational and algorithmic levels of analysis are also necessary for genuinely complete explanations of vision.

Marr's arguments are difficult to interpret, and his untimely death meant that he couldn't clarify his arguments, so a cottage industry has formed around the exegesis and rational reconstruction of Marr's arguments since the 1980s. In a more recent series of papers, Oron Shagrir and William Bechtel (henceforth, S&B) propose a rational reconstruction of his argument: roughly,

[1] Penultimate draft of paper forthcoming in *The British Journal for Philosophy of Science*. Please cite the published paper, not this draft.

that a complete description of visual mechanisms is insufficient for a complete explanation of vision because the latter must not only describe the mechanism but also explain how the mechanism is *appropriate* for the visual tasks (Shagrir, 2010; Bechtel & Shagrir, 2015; Shagrir & Bechtel, 2017; Shagrir, 2022). I'll refer to this appropriateness relation as *mechanism-task fit* (henceforth, "M/T fit"). They argue that this appropriateness relation consists in identity at a computational level of description: a mechanism is appropriate to a task because a computational description of the mechanism's activity is identical to a computational description of the task. I'll refer to S&B's account of M/T fit as the *computational identity account*.

We'll see that Marr offers a more nuanced treatment of M/T fit for the case of retinal edge detection than the computational identity account can accommodate. For one, Marr distinguishes between tasks ("problems") and their optimal solutions. This suggests that M/T fit is better construed as some kind of identity between the mechanism and a solution to the task. However, Marr also distinguishes between multiple solutions that are optimal subject to different sets of constraints. This raises the question: which "boundedly optimal" solution must a computational description of a mechanism be identical to for M/T fit? Marr himself doesn't offer an answer to the latter question. We'll see that part of the problem is the case study itself: Marr partly infers the task for the retina from the retinal mechanism itself and he partly infers the retinal mechanism from the inferred retinal task. As a result, his characterisation of M/T fit is laden with his prior assumptions about M/T fit. We'll see that we can avoid these issues just by considering a case where complete, independent characterisations of the mechanism and task are available.

In this paper, I'll argue that mechanisms are appropriate to tasks in virtue of satisfying constraints on above-chance successful task performance, not in virtue of identity between computational descriptions of the mechanisms and tasks. Thus, I'll suggest, an explanation of M/T fit is more aptly called a *task-fitting explanation*, rather than a computational explanation. In §1, I'll review Marr's case study of retinal edge detection and argue that it diverges from S&B's identity account of M/T fit. I'll explain how retinal edge detection is a misleading case for M/T fit. In §2, I'll review my previous explanation of sound localisation in the barn owl (Rumana, 2024) and I'll argue that provides a better case study for M/T fit. In §3, I'll propose a reformulation of this explanation, as one that grounds M/T fit in constraint satisfaction. In §4, I'll conclude by arguing that a complete explanation of above-chance successful task performance must contain both mechanistic and task-fitting components—because there is an interesting sense in which they are "modal complements".

§1. Mechanism-Task Fit

Marr (1982) distinguishes between three levels of analysis: computational, algorithmic, and implementational. The arguments in this paper have implications for all three levels, but the focus of this paper is the computational level. Marr (1982: 22) claims that the computational level of analysis is "the level of *what* the device does and *why*". S&B argue that this distinguishes between a what-element and a why-element, which (I'll argue) they suggest are computational descriptions of the mechanism's input-output profile and the task, respectively. Moreover, they propose that the what- and why-elements stand in some kind of appropriateness relationship, where the what-element may (or may not be) appropriate to the why-element. In this section, I'll show that Marr thinks this appropriateness relationship is mediated by non-ideal

(constrained) solutions to the task. Then I'll show that S&B think there is a direct (unmediated) relationship between the what- and why-elements: they are identical at a computational level of description. I'll argue that S&B's computational identity account of M/T fit is unable to accommodate the case of retinal edge detection, before arguing that a better case study is needed to answer questions left open by Marr's account of M/T fit.

§1.1. Marr's Account

Marr (1982, chapter 2) aims to give a multi-level analysis of edge detection in the retina. A starting point for his account is empirical evidence that firing rates in retinal ganglion cells (RGCs) peak when the retina is presented with edges in specific orientations (e.g., Kuffler, 1953). To be clear, this evidence isn't sufficient to *entail* that RGCs perform edge detection. For that, we need an appropriate interpretive function, which reads the firing rates of RGCs and interprets them as instances of edge detection: e.g., peak firing rates are indication that there is an edge at the location in the retinal field enervated by RGC dendrites, and submaximal firing rates are indication that there isn't an edge at that location. Following Kuffler and others, Marr individuates the edge detection task partly from evidence about the actual activities of the RGC mechanisms. This is a *mechanism-to-task inference*.

This is the first respect in which retinal edge detection is a potentially misleading case of M/T fit. After all, our aim is to account for the relation that makes a mechanism appropriate to the task. If we partly individuate the task from evidence about the mechanism, then our mechanism-to-task inference may contaminate our account of M/T fit. More specifically, if we oversimplify the mechanism-to-task inference, that might lead us to oversimplify the relationship that makes a mechanism appropriate to the task. In fact, I'll argue in §1.2 that S&B succumb to this inferential risk. To avoid this, we'll consider a case in §2 and §3 where we can completely individuate the task independently from the mechanism.

Marr's analysis starts at the computational level, by characterising the edge detection task: taking a two-dimensional image, $I(x, y)$, that the environment projects onto the retinal surface and indicating positions in the image that have been projected from edges in the physical environment. Like any task, it has normative structure: it presents a stimulus that admits of *correct* responses (indicating all and only positions in the image that correspond to physical edges from the observer's point of view) and *incorrect* responses (indicating image positions that don't correspond to physical edges and failing to indicate image positions that do correspond to edges from their POV).

Next, Marr characterises a solution to this task. He argues that the image should be smoothed to remove noise and that this can be done efficiently by a Gaussian function, G, which produces a denoised image, $\mathrm{G} \times I(x, y)$. Marr argues that physical edges tend to produce sudden changes in luminance, which survive denoising. These sudden luminance changes can be detected in two ways: by identifying maximum rates of luminance change (first derivative) or by identifying points (called "zero-crossings") where the rate of luminance change turns from positive to negative, or vice versa (second derivative). There are many second derivative operators, but Marr settles on the second-order Laplacian, ∇^2, because it is an *efficient* way to calculate zero-crossings—it is directionally invariant (the derivative is the same no matter which angle it is

calculated from) whereas most other derivatives are directionally variant (derivatives must be calculated from each angle, which is much less computationally efficient). So, Marr proposes a non-ideal solution, known today as a *Marr-Hildreth edge detector*, which indicates edges wherever there are zero-crossings in $\nabla^2 G \times I(x, y)$ (Marr & Hildreth, 1980).

Marr's solution to the edge detection task isn't perfectly ideal: it does take minimal computational constraints into account (Griffiths et al., 2015, 2024; Lieder & Griffiths, 2020). There are arbitrarily many ways to denoise image data and there are many derivative operators, but a Gaussian function and a Laplacian operator are *efficient* ways to calculate these respective steps. Efficiency considerations are irrelevant to an ideal performer with unlimited resources. They are only relevant to non-ideal performers who face some kind of computational constraints. For Marr, then, fit between the RGC mechanism and the edge detection task is partly mediated by non-ideal solutions to the task that are indexed to specific assumptions about computational constraints. This is a critical feature of Marr's view, which I'll argue in §1.2 that S&B's account of M/T fit fails to accommodate. An important gap in Marr's own account is that mechanisms are subject to an enormous set of constraints, but Marr doesn't specify which of these many constraints are relevant to M/T fit. Hence, it is unclear which non-ideal solution mediates M/T fit, whether multiple do, and what exactly the nature of this mediation is. I'll develop answers to these questions in §3.

Marr argues that it is physiologically implausible that a neuron could *directly* calculate the zero-crossings of $\nabla^2 G \times I(x, y)$. So, he devises a hypothetical circuit of RGCs consistent with their known properties that would allow them to reasonably approximate the zero-crossings of $\nabla^2 G \times I(x, y)$. However, Marr had no direct evidence that RGCs were arranged and wired together in just such a way to approximate a Marr-Hildreth edge detector. If RGCs do calculate the zero-crossings of $\nabla^2 G \times I(x, y)$, then this circuit description provides a possible explanation for *how* RGCs do so. This is a clear case of "reverse engineering", a strategy often touted in modern approaches inspired by Marr (Miłkowski, 2015; Griffiths et al., 2015, 2024; Lieder & Griffiths, 2020). Thus, Marr characterises the RGC mechanism partly by appealing to his non-ideal solution to the edge detection task. This is a *task-to-mechanism inference*.

This is the second respect in which retinal edge detection is a potentially misleading case of M/T fit. After all, the actual organisation of RGC circuits was unknown to Marr and 1980s visual neuroscientists. However, if we partly individuate the mechanism from the task we attribute to the mechanism, then our task-to-mechanism inference may contaminate our account of M/T fit. More specifically: we run the risk of oversimplifying the task-to-mechanism inference, which would lead us to oversimplify the relationship that makes a mechanism appropriate to the task. We'll see below that S&B also succumb to this inferential risk, and we'll avoid this by considering a case in §2 and §3 where we can completely individuate the mechanism independently from the task that the mechanism performs.

§1.2. Shagrir & Bechtel's Account

S&B take as their starting point Marr's (1982: 22) quote that the computational level of analysis is "the level of *what* the device does and *why*". First, they characterise the *what-question* as: "what mathematical function characterises the relationship from the RGC inputs to their

outputs?" For edge detection, Shagrir (2022: §9.5.2) says, "The *what* aspect specifies that early visual processes compute the zero-crossings of $\nabla^2 G \times I$." More generally, that is: the what-aspect is a mathematical description of the mechanism's input-output profile (the task *qua* solution). Elsewhere, though, S&B (2017, p. 21) say that "the what-aspect provides a description of the mathematical function that is being computed". Strictly speaking, though, this is inconsistent: the what-aspect is either a mathematical description of a mechanism's input-output profile (as a mathematical function) or a description of a mathematical function, not both. It seems to me that the former claim is more consistent with the rest of S&B's account (especially with Shagrir, 2010), so I'd recommend that the reader take the latter claim less literally.[2]

Second, they note that this what-element isn't sufficient to answer the *why-question*: "why do RGCs compute $\nabla^2 G \times I(x, y)$?" Shagrir elaborates the question as follows:

> "Why does the neural process that starts with a representation of light intensities, that is, the retinal image I, and computes the zero crossings of $\nabla^2 G \times I(x, y)$ end up with a representation of physical edges such as object boundaries and not, say, with a representation of the object's colours?" (Shagrir, 2010: 12)

For Marr and Shagrir, the answer to this why-question describes correspondence between a computational description of RGC activity and contingent features of the environment:

> "We happen to live in a world in which sudden changes in the retinal image (which the zero values of $\nabla^2 G \times I$ measure) strongly correlate with object boundaries. We could have lived in a world that consists of surfaces that sharply change reflectance across their solid faces. In such a world, the zero crossings of $\nabla^2 G \times I$ would be a poor way of detecting boundaries of objects." (Shagrir, 2010: 13)

Shagrir describes this explanation as a contrast between the actual world and counterfactual worlds, but the modal scope isn't nearly that large. In fact, the zero-crossings of $\nabla^2 G \times I(x, y)$ are weakly correlated with physical edges in the actual world: they fail to detect round edges (where luminance changes are smooth) and tend to be triggered by glare on object surfaces that aren't edges (where there are sudden luminance changes). To avoid this, modern machine vision tends to use algorithms like Canny edge detection (Canny, 1986).

Next, S&B argue that answers to these what- and why-questions are necessary but insufficient for computational theory. What's missing is an account of how the mechanism described by the what-element is *appropriate* to the task described by the why-element:

> "A computational level also has to explain why computing this function is appropriate for the information-processing task." (Shagrir, 2010: 9)

In other words, a mathematical, input-output description of the mechanism and a description of the task (*qua* problem) must be integrated into a description of M/T fit. The job of computational

[2] Some readers may prefer a different interpretation of S&B. What matters for my argument is just that the computational identity account (which I reject in this paper) is entailed by *some* of what S&B say—especially in Shagrir's solo work (2010, 2022)—even if it's inconsistent with other things they say—especially in their joint work. After all, part of my point here is that S&B's account isn't fully consistent.

theory, according to S&B, is to *explain* M/T fit by identifying the relationship between mechanism and task that *makes* the mechanism appropriate to the task:

> "We thus should wonder what is it about the causal relation between [the input brain state] and [the output brain state] that guarantees that the neural representation of [the input brain state's content] will lead to a neural representation of [the output brain state's content]." (Shagrir, 2010: 9)

This is an insightful observation, only implicit in Marr's account, which motivates my account here. Moreover, this quote is further textual evidence that the what-element is a mathematical description of the mechanism's input-output profile (the causal relation considered at a very coarse grain of description), *not* a mathematical function per se.

By my count, S&B consider three different suggestions for what kind of relation could ground M/T fit. Their weakest suggestion is some kind of *morphism*:

> "The detection of visual edges mirrors a pertinent relation in the visual field in the sense that there is an isomorphism (or some other morphism) between this visual process and the visual field. This morphism is exemplified by the (alleged) fact that the visual system and the visual field have a shared mathematical description (or structure)." (Shagrir & Bechtel, 2017: 20)

However, this suggestion is too weak. Taken literally, morphisms are only sensitive to first-order logical relations in a mapping's extension, like whether all or some elements in a mapping's domain are mapped onto unique or common elements in a mapping's range. They aren't sensitive to the mathematical identities of a mapping's intension, like whether each element in a mapping's range is equal to $\nabla^2 G \times I(x, y)$ for x and y in the mapping's domain. Since S&B emphasise that M/T fit in the retinal case is explained by appeal to an intensional mathematical relation like $\nabla^2 G \times I(x, y)$, I doubt that they think we should take literally their appeal to a weak extensional relation like any kind of morphism (c.f., Shagrir, 2022).

A stronger suggestion is *adequacy*:

> "The role of the Why element is to demonstrate the appropriateness and *adequacy* of what is being computed to the information-processing task." (Shagrir, 2010: 8, emphasis added)

I take adequacy to be some kind of modal relation, which involves the features of a mechanism's input-output profile making possible above-chance successful task performance, presumably by satisfying certain constraints on competent task performance. S&B seem to agree:

> "This match between the task and the mechanism shows why the mechanism succeeds." (Shagrir & Bechtel, 2017: 21)

Unfortunately, S&B don't say anything more about this proposal, and we'll see that it doesn't call for any kind of computational (mathematical input/output) description (which they both emphasise, especially in Shagrir's work), so it's unclear how seriously they take this view. I think this is their most plausible proposal, though, so I'll defend a version of it in §3.

Their strictest suggestion, developed at length by Shagrir (2010), is *identity at a computational level of description*:

> "The role of the What element is to specify the mathematical function that is being computed, namely, the *f*-relation between [the input brain state] and [the output brain state]. The role of the Why element is to demonstrate that the [relation between the representational contents of the input and output brain states] is also an *f*-relation." (Shagrir, 2010: 16)

According to this proposal, the activity of RGCs performs edge detection because 'detects zero-crossings of $\nabla^2 G \times I(x, y)$' is an accurate computational description of both (a) the input-output profile of RGC activity (the what-element) and (b) luminance changes over object boundaries (the why-element). This is the proposal repeated most often across S&B's work, and the only proposal evaluated by Shagrir (2022), so it seems to me that this is the view they take most seriously. However, regardless of whether this is in fact their preferred conception of M/T fit, I think it is an intuitive view that is worth engaging with and ultimately, I'll argue, rejecting.

In my view, this computational identity account makes three key mistakes. First, it doesn't draw Marr's distinction between a task (problem) and its solution (c.f., Kay et al., 2023). After all, 'detects zero-crossings of $\nabla^2 G \times I(x, y)$' is a plausible description of *one* solution to the task but it isn't a plausible description of the task itself. A plausible description of the task ought to be formulated in terms that make it possible for us to recognise all solutions as solutions of that task—even if they don't detect the zero-crossings of $\nabla^2 G \times I(x, y)$ (as in Canny edge detection). Perhaps, though, S&B could say that M/T fit consists in identity between computational descriptions of the mechanism and the task's solution. There are other problems, though.

Second, it doesn't draw Marr's distinctions between multiple solutions to the task, which vary to the extent that they make correct performance on the task possible for performers subject to different constraints. As I argued in §1.1, 'detects zero-crossings of $\nabla^2 G \times I(x, y)$' isn't a plausible description of the ideal solution to the edge detection task, but it is a plausible description of one non-ideal solution to it (among very many). Thus, M/T fit is mediated by a non-ideal solution (at the algorithmic level, on Marr's view). Perhaps, though, S&B could say that M/T fit consists in identity between computational descriptions of the mechanism and one of the task's non-ideal solutions. Of course, this raises the question that I posed to Marr: which non-ideal solution? Like Marr's account before it, S&B's computational identity account of M/T fit lacks the resources to answer this question.

Third, it assumes that there will always be a non-arbitrary mathematical description of the mechanism. This may be true for some cases, but we'll consider a mechanism in §3 for which there is no non-arbitrary mathematical description of its activities. There may not even be a non-arbitrary mathematical description of RGC activities. After all, Marr infers a putative model of RGCs from his computational description of the task solution that he tentatively attributes to RGCs. It's unclear whether this reverse engineering strategy is a *reification* of the computational description, or whether the *actual* RGC organisation makes this computational description true. A better case should use a mechanism for which we have a model that is complete and accurate, and so is less likely to be biased by our prior assumptions about M/T fit.

Since our notion of M/T fit is no longer confined to a computational (mathematical input-output) level of description, I suggest that computational theory is more appropriately described as the explanation of M/T fit, or what I propose we call "*task-fitting explanation"*.

§2. Case Study: Sound Localisation

Retinal edge detection is a misleading case study for M/T fit because the task is partly characterised by evidence about the mechanism and the mechanism is partly characterised by Marr's non-ideal solution to the inferred task (among many other possible non-ideal solutions). My objection isn't that these inferences are viciously circular, but that co-characterising mechanism and task might lead us to characterise the task in ways that make the mechanism easier to "fit" to the task and vice versa. Hence, we might underestimate its complexity. In §1, I argued that this might have led S&B's account astray. To avoid the same fate, we need a case where we have complete, mutually independent characterisations of the mechanism and the task.

One way to characterise the task independently of the mechanism is for an experimenter to explicitly design a task and administer it to an organism within an experimental paradigm. Kay et al. (2023) refer to such tasks as "experimenter-defined tasks", which are just experimental situations that make possible a range of behavioural responses for the organism, some of which are categorised as correct and others as incorrect. They argue that a critical caveat of 1970s visual neuroscience was that it lacked experimenter-defined tasks. After all, technical limitations only allowed for *in vivo* electrode recording in animals who were lightly anaesthetised and so incapable of producing behavioural responses. These technical limitations were mostly overcome in the 1980s, so Kay et al. urge modern visual neuroscience to incorporate more experimenter-defined tasks into vision paradigms.

Unfortunately, the sensorimotor mechanisms involved in vision are complicated, such that our understanding of their performance on experimenter-defined tasks remains highly incomplete. For a simple case study that avoids the pitfalls of retinal edge detection, we'd do well to look outside of visual neuroscience. In recent previous work, I developed a case study that meets our desiderata (Rumana, 2024). It reviews the mechanism for sound localisation in the barn owl (*Tyto alba*). Task-to-mechanism inferences played an important role in mechanism discovery: an idealised solution to the sound localisation task by Jeffress (1948) was indispensable to discovering the barn owl's sound localisation mechanism (Konishi, 1993). This might seem to raise the same problem of contamination for M/T fit that I raised in §1. Fortunately, though, we've kicked away the ladder by now: between the 1980s and 1990s, the sound localisation mechanism was completely mapped from stimuli to behavioural responses. Now it can be individuated by task-neutral criteria like Craver et al.'s (2021) causal betweenness criterion (Rumana, 2024). I previously took this explanation to be mechanistic, so I developed and motivated it as such. I now believe it is better formulated as a non-mechanistic explanation of M/T fit. I'll review the explanation in its previous formulation (Rumana, 2024) in this section and I'll propose a reformulation of it as an explanation of M/T fit in §3.

In the studies I previously reviewed (Rumana, 2024), the barn owls are given a specific, experimenter-designed behavioural task to perform. A barn owl starts perched in complete darkness. A zeroing speaker plays directly ahead of the owl, who automatically saccades

forward, and then a target speaker plays from some point in their frontal azimuth (the left-to-right field) (Knudsen et al., 1979). The sound from the target speaker will reach the owl's two ears at different times, depending on where the sound is in the owl's frontal azimuth: e.g., a sound emanating from 30 degrees to the right will reach the right ear 66 microseconds before it reaches the left ear. Thus, task information is inputted to the barn owl at the level of the whole system: in the spatial position of the sound relative to the spatial positions of the owl's two ears. Barn owls are exceptionally competent at performing this task: for most sounds in normal conditions, barn owls can move their line of gaze to within 1 or 2 degrees of the sound source. Thus, task information is also outputted from the barn owl at the level of the whole system: in the position of their line of gaze relative to their initial position on the perch and the true location of the sound source.

This specific task precisely individuates a mechanism that I call the frontal azimuthal sound localisation (FASL) mechanism. As far as we know, the entire FASL mechanism from stimulus inputs to behavioural outputs was mapped by Knudsen, Konishi, and others in the 1970s through the 1990s. Thus, they achieved the dream of 1970s visual neuroscience, at least as articulated in Barlow's (1972) neuron doctrine: a complete mechanistic explanation achieved through *in vivo* single-unit recording. Arguably, they also achieved completeness by New Mechanist standards: I counted 10 working parts that are constitutively relevant to auditory localisation and form a single-stream, feedforward, Markov chain from stimulus to response.[3] In Marrian form, I argued that this explanation is incomplete, because it leaves open an important why-question: why does the barn owl require 10 working parts to solve the task rather than more or fewer working parts?

In response, I offered an empirically informed, but speculative, answer. The barn owl isn't just a FASL task performer. On the contrary, they have to manage a continuous stream of demands from the environment. To do this, they represent these demands together in their premotor system and decide which demands get control of the motor system at any given time. To represent all of these demands together without interference, the premotor system requires these demands to be represented in a very sparse format—i.e., in the firing rates of small sub-populations (sometimes call *loci*) of neurons (e.g., Billings et al., 2014; Cayco-Gajic et al., 2017; Cayco-Gajic & Silver, 2019; Litwin-Kumar et al., 2017). Thus, the barn owl's FASL mechanism is subject to a key constraint: to route all task-relevant information between the external nucleus of the inferior colliculus (ICX) and the midbrain tegmentum (which includes a relay in the optic tectum), FASL task information must be encoded in a state of the unilateral nuclear locus (a local cluster of neurons in a region of one side of the brain). I'll refer to this as the *locus constraint*.

The locus constraint creates a serious challenge for the auditory localisation system. Recall that task-relevant information enters and exits the barn owl in a very *distributed* code—i.e., task-relevant information is encoded in states of the whole system. Somewhere between input and output, then, it has to encode task information in a very *sparse* code—i.e., encoded in states of a unilateral nuclear locus—so it can be routed through the premotor system. I argued that the sparse-distributed continuum can be coherently operationalised in terms of levels of anatomical composition: codes are "more distributed" when they are states at higher levels of anatomical composition, and they are "sparser" when they are states at lower levels of composition. I made

[3] This just means that the activity of each working part automatically triggers the activity of the next working part without any kind of attention, decision-making, or memory.

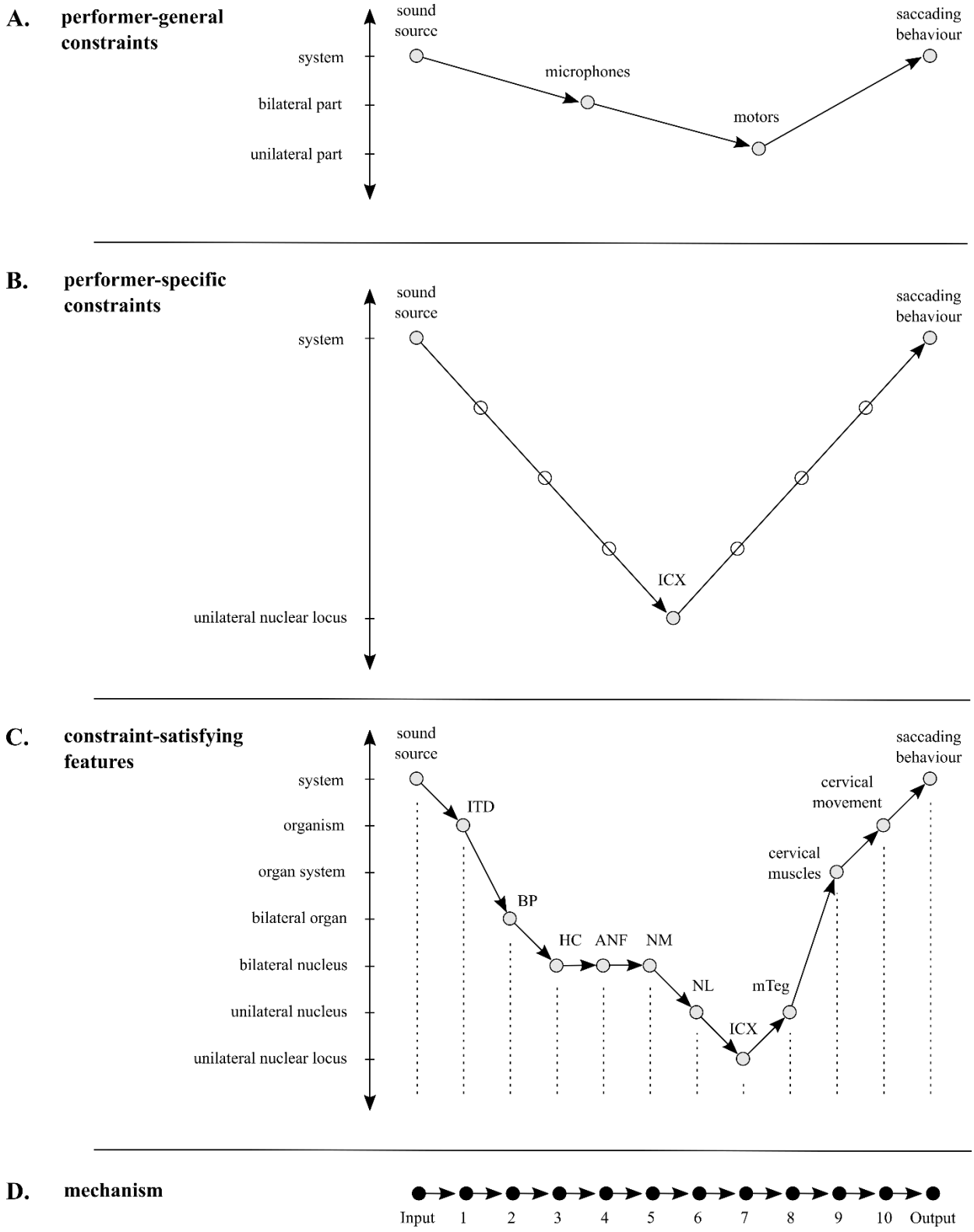

Figure 1. (a) Working parts of minimal performer plotted against three levels of composition (see §3.2). (b) A rough illustration of the locus and transport constraints as if imposed on the minimal performer. (c) Working parts of the FASL mechanism plotted against seven levels of anatomical composition. (d) Causal network diagram of the ten constitutively relevant working parts of the FASL mechanism itself. Modified from Rumana (2024).

an important concession here: anatomical levels may not be coherent in general (c.f., Craver,

2007; Craver & Bechtel, 2007; Potochnik, 2021), but they are coherent for the working parts of single-stream, feedforward, Markovian networks like the FASL mechanism.

For example, take the nucleus magnocellularis (NM), which synapses onto the nucleus laminaris (NL). The frontal azimuthal location of the sound source is encoded in the interaural (between-ear) time delay (ITD) between action potentials in the left NM (at one point in space and time) and those in the right NM (at another point in space and time), but not in the separate occurrences at any one place and time (in either the left or right NM). Thus, ITD information is encoded at the level of the bilateral nucleus (the left and right NM), but not at the level of the unilateral nucleus (either the left or right NM). By comparison, the synaptic junction from NM to NL is a complicated circuit that translates the ITD into a place code, such that ITD is encoded in the position of a few neurons in either the left or right NL. Thus, the NL re-encodes ITD information at the level of the unilateral nucleus (either the left or the right NL). Since unilateral nuclei are proper parts of bilateral nuclei, the NM-NL junction can be said to re-encode ITD at a lower level. In this way, I suggested that the FASL mechanism has to progressively reencode task information in states at lower and then higher levels of anatomical composition.

Finally, recall my previous question: why does the barn owl require 10 working parts to solve the task rather than more or fewer working parts? I proposed that 10 working parts are required to solve the task because the FASL mechanism is only able to *gradually* increase or decrease the sparseness of the code. This is an appeal to what we might call a *transport constraint*: task-relevant information can only be transported across a few levels of anatomical composition at a time. In particular, the FASL mechanism requires 6 working parts to move task-relevant information down to a very sparse code at the level of the unilateral nuclear locus, so that it can be routed securely through the premotor system, and then it requires 4 working parts to move task-relevant information back up to a very distributed code at the level of the whole system (Figure 1c). Inspired by Figure 1c, we could refer to this as a "V-path" for task information—insofar as task information travels along a V-shaped path across the 7 levels of anatomical composition. It is a description of this V-path that answers my previous question.

From this, I concluded that the completeness conditions for mechanistic explanation include more than just the set of all working parts constitutively relevant to the mechanism's behaviour and the set of all causal links among these working parts. *Inter alia*, they must also include the compositional relation between the anatomical types of these working parts. This conclusion assumes that a mechanistic explanation is responsible for the answer to the organisational question (Rumana, 2024). We aren't committed to this conclusion. In fact, I reject it now. What we need for this paper is just that the V-path is an anatomical property of the whole FASL mechanism, and it is meant to explain why the FASL mechanism is appropriate for FASL task performance. Whether this explanation is true is beside the point here. What matters for our purposes is whether it can provide a rich, coherent example of M/T fit grounded in constraint satisfaction, rather than identity.

§3. Task-Fitting Explanation

In this section, I'll argue that the explanation I previously described (Rumana, 2024) is better analysed as an explanation of M/T fit. In particular, I propose that what I described as an

"organisational question" is a why-question about *appropriateness*: "why is 10 an appropriate number of layers for FASL task performance?" Then I'll argue that my explanation is better understood as the following answer to that why-question: "10 layers is an INUS condition (to be defined below) for the holistic anatomical organisation that *satisfies* not only the (generic) *constraints* on competent task performance but also the (specific) locus and transport *constraints*." Hence, this illustrates a task-fitting explanation, which appeals to constraint satisfaction but not identity. At the end of this section, I'll argue that this reformulated explanation offers a useful corrective to Marr's case of retinal edge detection for a more accurate, representative account of M/T fit.

§3.1. Experimenter-Defined Task

The first component of a task-fitting explanation is a description of the experimenter-defined task—the experimental situation that (a) presents an organism with stimuli, (b) affords them various behavioural responses to those stimuli, and (c) makes it such that certain responses count as correct and others as incorrect. Let's discuss each of these three components in turn.

First is a *stimulus description*, which specifies which stimuli presented by the experiment are relevant to the experimenter-defined task. This description is incomplete insofar as it leaves out irrelevant details and inaccurate insofar as it represents what the stimuli ought to be (rather than what they actually are) when they slightly diverge from the intended experimental design. For example, here's an input description for the FASL task: a performer rests on a perch with a line-of-gaze fixed at 0° (i.e., facing ahead) and a sound playing from a target speaker at a different position in the frontal azimuth of the performer. Experimenters do their best to minimise incompleteness and inaccuracy in the input description by eliminating extraneous features in the experimental setup, e.g., by sound proofing the rooms and keeping the lights off.

The second component of the task description is the *behavioural description*, which specify which aspects of the behavioural responses from the performer are relevant to the task. For example, here's an output description for the FASL task: the performer's line-of-gaze intersects the target speaker. As before, this description will be incomplete insofar as it leaves out irrelevant details. For example, the performer undergoes various positional changes during the task, but only their line-of-gaze is relevant. In fact, Knudsen et al. (1979) note that the most challenging part of training is to minimise the barn owl's extraneous movement during performance, which aims to limit the amount of abstraction required for the output description. Likewise, this description will be inaccurate insofar as edge cases where the line-of-gaze misses the target speaker by a degree or two may be regarded as intersecting the target speaker. These inaccuracies are unimportant because they are rare—barn owl performance is highly accurate and precise.

The third component of the task description is the *task norm*, which categorises each possible stimulus-response pair as either correct or incorrect. For example, performance is correct iff the performer's line-of-gaze intersects the target speaker (the behavioural condition) by the end of the first head saccade after the target speaker emits a sound (after the stimulus condition)—given a certain margin of error (i.e., a certain tolerance for inaccuracy in the behavioural description). This task norm may be communicated to the performer via explicit instructions (if the performer and the experimenter use the same language) or a training period (as in the barn owl studies).

Matters get more complicated if participants don't follow the instructions (Rumana, 2022), but that isn't an issue here since barn owls follow the task after brief training (Knudsen et al., 1979).

Two points about this task description are worth mentioning. The first is that a computational description is available for this task, consistent with S&B's suggestion that tasks admit of computational descriptions. Moiseff & Konishi (1981) note that the geometry of the barn owl's head is such that the FASL task involves measuring the interaural time delay (ITD) between the sound reaching each of the barn owl's ears (in microseconds), solving the equation "frontal azimuth = 0.38 × ITD – 0.54" for frontal azimuth (in degrees), and then conditioning the physiological response on the solved value for frontal azimuth. Note that this is a mathematical relationship between each stimulus and the correct behavioural response to that stimulus. No such description was available for retinal edge detection, because the experimental situations didn't afford (correct or incorrect) behavioural responses (animals were anaesthetised).

The second point is that the task is defined with respect to the external features of a generic performer—how the stimulus must enter the performer and how the response must exit from the performer. In this sense, a task description is *performer-relative*, *performer-general*, and *externalist*. On the input side, for example, the FASL task must describe the position of the sound source relative to the position of the barn owl's left and right ears. Likewise, what part of an object counts as an edge is only definable by its relative position to an organism's left and right eyes. Shagrir (2010) follows Sterelny (1990) in describing the task description as "ecological", but this is misleading—it suggests the task description is performer-neutral, when it is performer-relative but performer-general. We'll see that this issue has repercussions in the next section.

§3.2. Generic Constraints

Although a task description appeals to the external (observable) features of the performer, it doesn't purport to describe the "inner workings" of any performer. Nevertheless, any non-trivial task will make certain demands that the "inner workings" of a performer must satisfy in order to perform the task correctly at rates above chance. Consider Kahneman & Frederick's (2005) bat-and-ball task, for example: "A bat and a ball are $1.10. The bat is $1.00 more than the ball. How much is the ball?" It's impossible to perform this task correctly above-chance without undergoing processes that more or less closely track the rules of elementary algebra. Of course, this is consistent with a lot of diversity: a performer could solve the bat-and-ball task using the methods of elimination, substitution, graphing, etc. but all these methods track the same rules of elementary algebra in the relevant sense. I'll describe these generic constraints as *task demands*.

Thus, the second component of a task-fitting explanation answers this question: what "inner workings" are *required* for any performer to transform the inputs into the correct outputs, given the nature of the task? The answer to this question is a *task analysis*, which describes the generic constraints that any performer must satisfy in order to select correct responses to the relevant stimuli at a given level of performance above chance. Of course, higher rates of correct performance are more difficult than lower rates of correct performance, so they impose more (and/or more restrictive) task demands. Thus, a task analysis isn't well-defined unless it is indexed to a particular rate of correct performance. Since barn owls are capable of correct

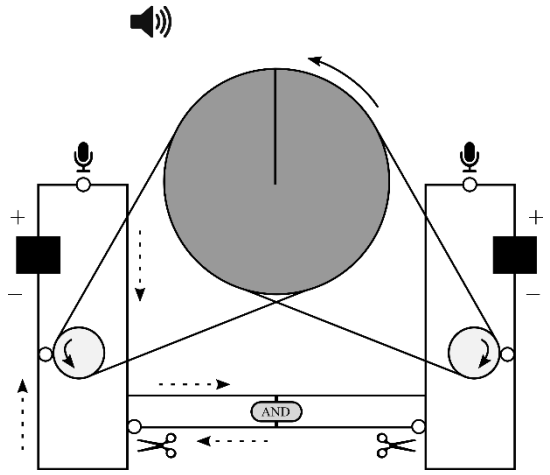

Figure 2. A mechanism sketch of a minimal FASL performer. Two microphones are divided by an insulating object, ensuring that each position of the source corresponds to a unique inter-recorder delay. The first microphone to receive sound powers a motor that quickly rotates the insulating object in the direction of the sound source. If the speed is appropriately calibrated, the object's 'line-of-gaze' will be pointing towards the sound source by the time the sound reaches the second microphone. This microphone will close the second circuit. Since the first circuit is already closed, this will trip the AND gate, which will cut both circuits. If the motors have an appropriate braking mechanism, the object will instantly stop rotating with its line-of-gaze intersecting the position of the sound source.

performance on nearly all trials, I'll develop a tentative analysis of the FASL task for near-optimal performance.

Formal tasks, like Kahneman & Frederick's (2005) bat-and-ball task, may admit of axiomatic task analysis, which just has to derive the solution within the rules of the formal system. Clearly, that won't be true for informal tasks like the FASL task. Instead, I propose that informal task analysis involves engineering a *minimal performer*—i.e., a performer with the simplest internal structure required to competently transform the input conditions to the appropriate output conditions (at a given competency, or rate of success). Since any task performer will have to instantiate the same structure as the minimal performer in order to satisfy the generic task demands, I'll sometimes describe the minimal performer as a *generic performer*. Since the minimal performer isn't subject to any other constraints besides the constraints imposed by the task itself, I'll sometimes also describe the minimal performer as an *ideal performer*.

Engineering a minimal performer for a simple task like the FASL task is straightforward. Figure 2 offers a crude representation of a minimal FASL performer. Two remarks about this minimal performer are relevant for our discussion here. The first remark is that the input and output conditions for the minimal performer must be carefully matched to the input and output conditions for the target performer. This is necessary to ensure that the task description is identical for the minimal and target performers. Since we're interested in the movement of task-relevant information across levels of composition, the relevant point here is that the minimal performer must be designed such that task-relevant information is encoded in states of the whole system in the input and output conditions. In particular, line-of-gaze must be defined as a holistic property of the system, rather than any proper part of the system.

The second remark is that our minimal FASL performer is so simple that it has just two *working parts*—at least if we're using the diagnostic criteria I offered to count constitutively relevant

working parts: (a) two recording devices to translate mechanical work into electrical work and (b) two motors to translate electrical work back into mechanical work (Rumana, 2024). Task-relevant information is encoded in states of each constitutively relevant working part: (a) the difference between the arrival of sound to the left and right microphones and (b) the duration of time that the left or right motor is allowed to rotate the system before the AND gate breaks the circuits. I'll describe this compositional organisation as the *minimal V-path* (Figure 1b).

That the minimal performer has a minimal V-path is defeasible evidence that a minimal V-path is a generic constraint for competent task performance—i.e., that it is a *task demand*. Let's call this the *subsystem constraint*: the inner workings of a competent performer must encode task-relevant information in at least two proper parts of the whole system. Obviously, this is consistent with, but much weaker than, the locus constraint that we attributed earlier to the barn owl. However, this evidence is defeasible just because I haven't ruled out the possibility that there is a competent performer without a minimal V-path. For all we know, there might be a competent performer that keeps task-relevant information encoded at the level of the whole system and never has to encode task-relevant information in any of its proper parts. This might seem inconceivable (at least, it falls outside the boundaries of my imagination), but inconceivability is *sometimes* consistent with possibility.[4, 5]

A task analysis like this is different from a task description in at least two important respects. First, it is concerned with the generic internal architecture required for competent task performance by any system. In this sense, a task analysis is *performer-relative*, *performer-general*, but *internalist*. Second, it isn't amenable to mathematical description. This is clearest for the FASL mechanism: the minimal V-path is the compositional organisation of the whole performer, which isn't mathematical in any interesting sense. But this is also true for Marr's example of edge detection. A minimal edge detector will presumably require (a) a surface that can be photo-stimulated, (b) a part that can calculate differences across the photo-stimulated surface, and (c) a part that can apply some kind of criterion to classify certain differences as edges. These requirements are substantive, not formal or mathematical.

As a result, M/T fit is essentially mediated by generic constraints, as revealed by the minimal performer. Thus, it is misleading to say that M/T fit is grounded in identity between certain descriptions of the mechanism and the task. It would be better for S&B to say that M/T fit is grounded in identity between the mechanism and the minimal performer at a coarse-grained level of substantive description (however, see §3.4). Let's call this the *ideal counterpart conception* of M/T fit. For example, we could say that the FASL mechanism fits the FASL task *partly because* the FASL mechanism and a minimal FASL performer both have V-paths. This coarse grain of substantive description abstracts away from the finer-grained differences between the FASL mechanism and our minimal FASL performer: e.g., that the FASL mechanism's V-path reaches down seven levels of composition whereas our minimal performer's V-path only reaches down one level of composition.

[4] We might wish that we could appeal to some kind of analytic proof, but this is a matter of engineering and engineering isn't really amenable to analytic proof.

[5] The subsystem constraint is only one of the task demands suggested by our minimal performer. Another has to do with the translation between electrical and mechanical work, but I don't have the space to discuss that here.

§3.3. Specific Constraints

Of course, generic constraints aren't the only constraints that can make competent task performance impossible. Organism constraints are equally capable of making it impossible for an organism to competently perform a task even if they have satisfied all the task demands. For example, I suggested in §2 that that the locus and transport constraints make it impossible for the barn owl to competently perform the FASL task—*unless* the FASL mechanism uses 6 working parts to transport task-relevant information down to the level of the unilateral nuclear locus, passes that sparse code for task-relevant information through the premotor system, and then uses 4 more working parts to transport task-relevant information back up to the whole system level. Thus, organism constraints are an essential part of the "match between the task and the mechanism [that] shows why the mechanism succeeds" (Shagrir & Bechtel, 2017: 21).

Thus, the third component of a task-fitting explanation answers this question: what "inner workings" are *required* to transform the inputs into the correct outputs, given the constraints on the organism themself? Answering this question is difficult due to the well-known problem that constraints can be easily proliferated—just by holding more and more features of a system fixed. After all, there is a sense in which the basic circuitry of the barn owl's brain is fixed at this moment in their evolutionary history, such that it is impossible for the barn owl to "re-route" task-relevant information around the actual working parts of the FASL mechanism (in the way assumed by my explanation from §2). However, this would trivialise our task-fitting explanation: the barn owl routes task-relevant information through each working part of the FASL mechanism just because it has already evolved to do so, and it is impossible for it to do anything else at this moment in their evolutionary history. In general, nothing but the actual state of the system becomes possible for the system, so a task-fitting explanation is just left to a description of the system's actual features.

To avoid this, task-fitting explanation requires a more permissive conception of possibility, which holds fewer features of the system fixed. For example, this conception of possibility should be relaxed enough to allow for the possibility that the barn owl could re-route task-relevant information through different parts of its brain. At the same time, though, it should remain strict enough to require that the barn owl could not route task-relevant information around the premotor system. It should strike just the right balance, such that it can justify singling out the locus and transport constraints without describing every part of the barn owl's V-path as a constraint (see Figure 1b). In a sense, we need to take a "design stance" (Dennett, 1987), where we pretend that we have a counterfactually-powerful but still-limited capacity to re-design the mechanism to improve its fit with the task. I don't have the space to develop a principled conception of the relevant kind of possibility here, but this is a matter of some urgency for anyone interested in task-fitting explanation.

Incorporating organism constraints into the explanans of task-fitting explanation again complicates our picture of M/T fit. After all, an essential commitment of the minimal counterpart account of M/T fit is that the appropriateness of a target performer's features is grounded in performer-general considerations, as represented by a minimal performer. I've suggested that this is false: the appropriateness of a target performer's features is *also* grounded in performer-specific considerations, which *cannot* be represented by a minimal performer. One option here

would be to individuate an ideal counterpart of the target performer with specific features of the target performer. Then we could ground M/T fit in identity between certain descriptions of the target performer and their semi-idealised counterpart. Call this the *semi-idealised counterpart* account of M/T fit. In fact, the Marr-Hildreth edge detector is a specific, semi-idealised counterpart of RGCs in this sense: it is subject to relatively strict computational constraints (stricter than an ordinary digital computer but which Marr thinks reflects the constraints of RGCs) that make edge detection impossible with directional variants and so requires isotropic derivatives like second-order Laplacians.

§3.4. Constraint Satisfaction

It seems to me that talk of counterparts is just an elliptical way of talking about constraints and constraint satisfaction: we could only construct the semi-idealised counterparts to the target performer by constructing the simplest performer that satisfies all the relevant generic and specific constraints. Hence, I propose, a more direct, parsimonious account should directly appeal to constraint satisfaction: M/T fit consists in the satisfaction of generic and specific constraints on competent task performance for an organism by various features of a mechanism. I take this to be a plausible way of cashing out Shagrir's (2010) passing suggestion that M/T fit might be grounded in a modal notion like *adequacy*.

Overall, then, here is my final proposal for the relation that makes a mechanism appropriate to a task. The task is defined by the experimenter independently of the mechanism. The mechanism is individuated by the stimulus inputs and behavioural outputs that are given by the task, but independently from the task norm (which specifies what counts as correct performance on the task). For Craver et al. (2021), e.g., the mechanism is the organised set of activities *causally between* the (task-relevant) stimulus inputs and behavioural outputs. The task norm is needed to individuate M/T fit, which is the special case where the mechanism produces activity that counts as correct vis-à-vis the task norm at rates above chance. Above-chance correct performance is *competent*: it reflects some kind of competence in the mechanism.

The task itself is sufficient to individuate generic constraints on competent task performance, which apply to all performers. Certain features of the mechanism satisfy these generic constraints and count *towards* the mechanism's competence (e.g., the barn owl's V-path). Call these *first-order contributing features*. If mechanisms only had first-order contributing features, satisfying generic constraints would be sufficient for competent performance. However, mechanisms also have features that count *against* the mechanism's competence. Call these *undermining features*. Undermining features make competent task performance impossible for the mechanism *unless* they are "counteracted" by further features of the mechanism (e.g., the barn owl's 10-step V-path counteracts the locus and transport constraints). Call those *second-order contributing features*. In other words, undermining features create performer-specific constraints, which are satisfied by second-order contributing features and thereby make possible competent task performance for the mechanism. Thus, specific constraints are jointly co-individuated (in a non-circular way) by the mechanism and the task. Overall, M/T fit is grounded in relations of constraint and constraint satisfaction between the task and three basic kinds of mechanism features.

Next, I propose that task-fitting explanation takes M/T fit as its explanandum and return various mechanism-task relationships of constraint satisfaction as its explanans:

> **Task-fitting explanation:** any explanation of the form:
> *Explanandum*: it is possible for a performer *P* to perform an experimenter-defined task *T* at a certain rate of success *r* whenever a mechanism *M* in *P* does Φ.
> *Explanans*: one or more features of *M*-doing-Φ (the working mechanism) satisfy generic constraints on any performer successfully performing *T* at *r* and specific constraints on *P* successfully performing *T* at *r*.

In particular, it is the satisfaction of both generic and specific constraints that *makes* it possible for a performer *P* to perform a task *T* at a certain rate of success *r* whenever a mechanism *M* in *P* does Φ. This making-possible relation gives task-fitting explanation its explanatory force.

Moreover, it comes in two strengths. The higher-strength relation is the satisfaction of generic constraints, which is a necessary condition for competent task performance for all performers. If there are multiple generic constraints, the satisfaction of any single generic constraint will be insufficient to make possible competent task performance. In general, satisfying a generic constraint is a necessary but insufficient (IN) condition for competent task performance. The lower-strength relation is the satisfaction of specific constraints, which are necessary conditions for competent task performance in some, but not all, performers. If there are multiple specific constraints, the satisfaction of any single specific constraint will be insufficient for competent task performance. In general, satisfying a specific constraint is an insufficient but necessary part of an unnecessary but sufficient (INUS) condition (*a la* Mackie, 1965) to make possible competent task performance.

§3.5. Modality

While necessity (in both strengths) features in the explanans of our task-fitting explanation, sufficiency does not. After all, task-fitting explanations describe features of mechanisms that make competent task performance *possible*, not *actual*. For this, the relevant kind of modality is necessity (both in the stronger IN sense and the weaker INUS sense), not sufficiency. By comparison, I would say that sufficiency is the job of mechanistic explanation. After all, it is the mechanism (given that it satisfies all the constraints on competent task performance) doing the actual work that turns the possibility of competent task performance into an actuality. Unless the mechanism does the work, though, its constraint-satisfying features will never be sufficient for an actual episode of competent task performance. I propose this is the fundamental difference between task-fitting explanation and mechanistic explanation—the former is concerned with the satisfaction of necessary conditions and the latter is concerned with the satisfaction of sufficient conditions (given that the necessary conditions are already satisfied).

My suggestion is consistent with the emphasis on sufficiency within the New Mechanist project. For example, take the project of specifying which working parts are constitutively relevant to mechanisms: very roughly, these are interested in which ideal interventions are *sufficient* (but unnecessary) for difference-making to particular phenomena (Craver et al., 2021). Likewise, their ontology has always stressed the active, working nature of mechanisms and their parts

(Machamer et al., 2000; Machamer, 2004; Darden, 2007; Craver & Darden, 2013; Kaiser, 2017; Kaiser & Krickel, 2017). If mechanistic explanation essentially traffics in sufficiency, then this emphasis is critical for realism about mechanistic explanation: it is the work that mechanisms do that *makes actual* the possibility of competent task performance (as given by the constraint-satisfying features). If this is right, New Mechanists should be happy to grant that mechanistic explanation finds its "modal complement" in task-fitting explanation.

This important point is foreshadowed, but misrepresented, by the suggestion that mechanistic explanation is *how-actually* explanation, whereas other forms of explanation like functional analysis (see §4) are *how-possibly* explanations (e.g., Machamer et al., 2000). This suggestion is misleading because it is generally assumed that this difference in modality is located in the explanans: e.g., that functional analysis identifies a possible mechanism, which could explain a phenomenon, whereas mechanistic explanation identifies the actual mechanism, which does explain the phenomenon (*a la* Piccinini & Craver, 2011). It seems to me that this difference is more naturally described as "possibly-how" versus "actually-how". By comparison, I think this difference in modality should be located in the explanandum: e.g., the possibility of competent task performance is the explanandum of task-fitting explanation whereas the actuality of competent task performance is the explanandum of mechanistic explanation. I think this is the more natural interpretation of the terms "how-actually" and "how-possibly". Under this latter interpretation, I would feel comfortable saying that task-fitting explanation is how-possibly explanation whereas mechanistic explanation is how-actually explanation.

Consistent with this emphasis on necessity and possibility (vs. sufficiency and actuality), I propose that task-fitting explanation is a species of *constraint-based explanation* (Ross, 2023; c.f., Lange, 2013, 2023). In general, constraint-based explanations explain why certain states are possible and others impossible for a system by describing the features of the system, which *make* those states possible and impossible, respectively. In other words, these possibility-making features are *necessary* for the system to occupy a given state. Task-fitting explanations are just a specific version of this: they explain why competent task performance is possible by describing features of the mechanism, which *make* competent task performance possible. In other words, the distinguishing feature of task-fitting explanations (*qua* species of constraint-based explanation) is that they are interested in the possibility vs. impossibility of competent task performance.

This emphasis on necessity and possibility also accounts for why task-fitting explanation should be distinguished from what we might call *advantage-based explanations*. For example, Chirimuuta (2014, p. 146) asks: "why do V1 simple cells have elongated receptive fields of the sort that can be fit by [the Gabor model]?" And her answer is an *efficient coding explanation*: "because this minimises Gabor–Heisenberg–Weyl uncertainty, and thus is an efficient way of coding visual information" (ibid.). This purports to explain a feature's instantiation by describing its advantages. Critically, though, this explanation (if it is one, and I am sceptical that it is) is non-modal: minimising Gabor–Heisenberg–Weyl uncertainty can be advantageous for coding visual information without making it possible or actual. Thus, it's distinct from mechanistic and task-fitting explanations insofar as it doesn't appeal to either actuality or possibility, respectively. See Shagrir (2022, §9.5.3) for a complementary discussion on the distinction between optimality explanations and computational (or task-fitting) explanations.

§4. Conclusion

A key insight from S&B is that a computational theory is meant to explain how a mechanism like the RGC circuit is *appropriate* to a task like edge detection (the explanandum) by describing the relationship that *makes* the mechanism appropriate to the task (the explanans). In this paper, I argued that they erred in claiming that identity between computational descriptions of the mechanism and task is the relationship that makes the mechanism appropriate to the task. I also considered alternative accounts where identity holds between the target performer and an ideal, generic, minimal performer (§3.2) and a semi-idealised counterpart of the target performer (§3.3). Next, I argued that there is a better alternative to each of these identity accounts, which grounds M/T fit in *constraint satisfaction*. On this alternative account, (a) the explanandum of task-fitting explanation is that a certain rate (above chance) of successful task performance is possible for a mechanism and (b) the explanans describes the mechanism's features that make that rate of successful task performance possible by satisfying performer-general constraints (IN conditions) and performer-specific constraints (INUS conditions).

An emphasis on the modal nature of M/T fit suggests that task-fitting explanation is not only different from but also *complementary* to mechanistic explanation. On the one hand, I've argued that task-fitting explanation is interested in the conditions that are *necessary for making possible* competent task performance—by describing the constraints that the properties of the mechanism must satisfy. But *making possible* competent task performance is insufficient for *making actual* competent task performance. For that, we need mechanistic explanation, which is interested in the conditions that are *sufficient for making actual* competent performance on a task—by describing the work that a mechanism actually does. But a mechanism's work *making actual* competent task performance takes for granted that the mechanism has all the properties necessary for *making possible* competent task performance. In this way, task-fitting explanation and mechanistic explanation are "modal complements". Thus, a *modally-complete* explanation of competent task performance must involve both mechanistic and task-fitting explanations.

This complementary relationship is helpful for recognising when the theoretical work that has been assigned to mechanistic explanation should have been assigned to its modal complement, task-fitting explanation. For example, I think task-fitting explanation is much better situated than mechanistic explanation to provide the accuracy and completeness conditions for functional analysis (*contra* Piccinini & Craver, 2011). After all, functional analysts generally start with task analysis, which aims to characterise generic constraints on competent task performance. Then they use behavioural evidence to draw inferences about performer-specific constraints, like limitations on working memory (Cummins, 1983; Weiskopf, 2011; Barrett, 2014). Finally, functional analysis comprises an incomplete ("abstract") and potentially inaccurate ("idealised") characterisation of a cognitive architecture that *might* satisfy the constraints—without identifying the features of the mechanism that actually satisfy the constraints.

If that's right, then any accurate functional analysis is just a "sketch" of a more complete (and accurate) task-fitting explanation. A common complaint with this sort of argument is that it neglects the value of abstraction and idealisation (e.g., Batterman & Rice, 2014; Chirimuuta, 2014; Levy, 2014). But it's important to stress that we can take the completeness and accuracy conditions of explanatory models to be worth characterising (as I've started to do here) without

taking any stand on (a) whether and when completeness and accuracy are good-making features of models, (b) whether they ground the explanatory status or power of explanatory models, or (c) when they override other good-making features of models like simplicity and intelligibility (Craver & Kaplan, 2020). On my proposal, functional analysis may provide a sketch of M/T fit, but the verdict is out on whether and when the accuracy and completeness that comes with the task-fitting explanation is important. Of course, future work is needed for a full defence and evaluation of this proposal. My point is just to illustrate that task-fitting explanation may do some of the work previously ascribed to mechanistic explanation.

Acknowledgements: I thank Carl Craver, Eric Hochstein, Edouard Machery, Bob Batterman, Oron Shagrir, Richard Samuels, and Sarah Robins for conversations that significantly improved this manuscript. I also thank the two anonymous reviewers and an editor at BJPS for their helpful feedback. Finally, I thank audiences at the Center for Philosophy of Science at the University of Pittsburgh, the 2024 meeting of the International Society for the Philosophy of the Sciences of the Mind (ISPSM), the 2024 meeting of the Deep South Philosophy of Neuroscience Workgroup, and the 2024 conference “Concepts for Understanding Brain Organisation” at City University Paris.

References

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception, 1*(4), 371-394. https://doi.org/10.1068/p010371

Barrett, D. (2014). Functional analysis and mechanistic explanation. *Synthese*, *191*(12), 2695–2714. https://doi.org/10.1007/s11229-014-0410-9

Bechtel, W., & Shagrir, O. (2015). The non-redundant contributions of Marr's three levels of analysis for explaining information-processing mechanisms. *Topics in Cognitive Science*, *7*(2), 312–322. https://doi.org/10.1111/tops.12141

Billings, G., Piasini, E., Lorincz, A., Nusser, Z., Silver, R. A., L˝orincz, A., Nusser, Z., & Silver, R. A. (2014). Network structure within the cerebellar input layer enables lossless sparse encoding. *Neuron, 83*, 960–974.

Canny, J. (1986). A computational approach to edge detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 8*(6), 679–98, https://doi.org/10.1109/TPAMI.1986.4767851

Cayco-Gajic, N. A., Clopath, C., & Silver, R. A. (2017). Sparse synaptic connectivity is required for decorrelation and pattern separation in feedforward networks. *Nature Communications, 8*, 1116. https://doi.org/10.1038/s41467-017-01109-y

Cayco-Gajic, N. A., & Silver, R. A. (2019). Re-evaluating circuit mechanisms underlying pattern separation. *Neuron, 101*(4), 584–602. https://doi.org/10.1016/j.neuron.2019.01.044

Chirimuuta, M. (2014). Minimal models and canonical neural computations: The distinctness of computational explanation in neuroscience. *Synthese*, *191*(2), 127–153. https://doi.org/10.1007/s11229-013-0369-y

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

Craver, C. F., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology & Philosophy*, *22*(4), 547–563. https://doi.org/10.1007/s10539-006-9028-8

Craver, C. F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. University of Chicago Press.

Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science*, *71*(1), 287–319. https://doi.org/10.1093/bjps/axy015

Craver, C. F., Glennan, S., & Povich, M. (2021). Constitutive relevance & mutual manipulability revisited. *Synthese*, *199*(3), 8807–8828. https://doi.org/10.1007/s11229-021-03183-8

Cummins, R. (1983). *The nature of psychological explanation*. MIT Press.

Darden, L. (2008). Thinking again about biological mechanisms. *Philosophy of Science*, *75*(5), 958–969. https://doi.org/10.1086/594538

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

Griffiths, T.L., Chater, N., & Tenenbaum, J. B. (eds.) (2024). *Bayesian models of cognition: Reverse engineering the mind*. The MIT Press.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. https://doi.org/10.1111/tops.12142

Jeffress, L. A. (1948). A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, *41*(1), 35–39. https://doi.org/10.1037/h0061495

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In *The Cambridge Handbook of Thinking and Reasoning* (pp. 267–293). Cambridge University Press.

Kaiser, M. I. (2017). The components and boundaries of mechanisms. In S. Glennan & P. Illari (Eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. Routledge. https://philarchive.org/rec/KAITCA

Kaiser, M. I., & Krickel, B. (2017). The metaphysics of constitutive mechanistic phenomena. *The British Journal for the Philosophy of Science*, *68*(3), 745–779. https://doi.org/10.1093/bjps/axv058
Kay, K., Bonnen, K., Denison, R. N., Arcaro, M. J., & Barack, D. L. (2023). Tasks and their role in visual neuroscience. *Neuron*, *111*(11), 1697–1713. https://doi.org/10.1016/j.neuron.2023.03.022
Konishi, M. (1993). Neuroethology of sound localization in the owl. *Journal of Comparative Physiology A*, *173*(1), 3–7. https://doi.org/10.1007/BF00209613
Knudsen, E. I., Blasdel, G. G., & Konishi, M. (1979). Sound localization by the barn owl (Tyto alba) measured with the search coil technique. *Journal of Comparative Physiology*, *133*(1), 1–11. https://doi.org/10.1007/BF00663105
Kuffler, S.W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16(1): 37–68. https://doi.org/10.1152/jn.1953.16.1.37.
Lange, M. (2013). What makes a scientific explanation distinctively mathematical? *British Journal for the Philosophy of Science, 64*(3): 485–511.
Lange, M. (2023). Explanations by constraint: Not just in physics. *International Studies in the Philosophy of Science*, *36*(4), 265–277. https://doi.org/10.1080/02698595.2023.2298085
Levy, A. (2014). What was Hodgkin and Huxley's achievement? *British Journal for the Philosophy of Science, 65*, 469–92. https://doi.org/10.1093/bjps/axs043
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*. https://doi.org/10.1017/S0140525X1900061X
Litwin-Kumar, A., Harris, K. D., Axel, R., Sompolinsky, H., & Abbott, L. F. (2017). Optimal degrees of synaptic connectivity. *Neuron, 93*, 1153–1164.
Machamer, P. (2004). Activities and causation: The metaphysics and epistemology of mechanisms. *International Studies in the Philosophy of Science, 18*, 27–39.
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25. https://doi.org/10.1086/392759
Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly, 2*(4): 245–64.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.
Marr, D. & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *207*(1167): 187–217. https://doi.org/10.1098/rspb.1980.0020
Miłkowski, M. (2013). Reverse-engineering in cognitive-science. In M. Miłkowski & K. Talmont-Kaminski (Eds.), *Regarding Mind, Naturally* (pp. 12–29). Cambridge Scholars Press.
Moiseff, A., & Konishi, M. (1981). Neuronal and behavioral sensitivity to binaural time differences in the owl. *Journal of Neuroscience*, *1*(1), 40–48. https://doi.org/10.1523/JNEUROSCI.01-01-00040.1981
Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311. https://doi.org/10.1007/s11229-011-9898-4
Potochnik, A. (2021). Our world isn't organized into levels. In D. Brooks, J. DiFrisco, & W. C. Wimsatt (Eds.), *Levels of organization in biology*. MIT Press.
Ross, L. N. (2023). The explanatory nature of constraints: Law-based, mathematical, and causal. *Synthese*, *202*(2), 56. https://doi.org/10.1007/s11229-023-04281-5
Rumana, A. (2022). Arbitrating norms for reasoning tasks. *Synthese*, *200*(6), 502. https://doi.org/10.1007/s11229-022-03981-8
Rumana, A. (2024). Anatomy's role in mechanistic explanations of organism behaviour. *Synthese*, *203*(5), 137. https://doi.org/10.1007/s11229-024-04534-x
Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of Science*, *77*(4), 477–500. https://doi.org/10.1086/656005
Shagrir, O. (2022). *The nature of physical computation.* Oxford University Press.
Shagrir, O., & Bechtel, W. (2017). Marr's computational level and delineating phenomena. In *Explanation and integration in mind and brain science* (pp. 190–214). Oxford University Press.
Sterelny, K. (1990). *The representational theory of mind: An introduction*. Cambridge, MA: Blackwell.

Weiskopf, D. A. (2011). Models and mechanisms in psychological explanation. *Synthese*, *183*(3), 313–338. https://doi.org/10.1007/s11229-011-9958-9