

20 Building Minds Together

So, what have we learned? Where, exactly, does all of the philosophical and scientific research on the mind discussed in this book get us? It might be hoped that it will eventually get us to a place where we understand minds well enough to know how to make one. After all, if we have sophisticated theories of consciousness, perception, emotion, memory, belief, desire, agency, intentionality, inner speech, outer speech, intelligence, neural information processing, self-knowledge, sex and gender, empathy, neurodiversity, and madness, then we will have before us many of the ingredients that should go into a recipe for building minds.

In this chapter, Javier Gomez-Lavin suggests that the history of the philosophy of mind is well understood as a history of attempts to provide just such a recipe for building minds. But Gomez-Lavin worries that the recipes on offer haven't improved over time. On the contrary, he argues that they each—including current cognitive scientific recipes that feature ingredients like working memory—do little more than provide new jargon with which to express our ordinary, folk psychological ways of understanding thinking and other mental capacities. Gomez-Lavin concludes by suggesting that this lack of progress results from the fact that almost all recipes for building minds share a common flaw: they focus too much on the individual mind, and neglect the ways in which other people (and group dynamics) shape how minds work.

In their response, Tim Elmo Feiten and Anthony Chemero explore what it would mean for philosophers and cognitive scientists to embrace Gomez-Lavin's suggestion that cognition is a fundamentally social phenomenon.

HOW TO BUILD A MIND

Javier Gomez-Lavin

1. So you want to build a mind

The engineer-turned-philosopher-of-mind Fred Dretske taught that “if you can't make one, you don't know how it works” (1994, p. 468). One way of

DOI: 10.4324/9781003483632-25

understanding what philosophers of mind have been doing since antiquity is by following something like Dretske's maxim. If understanding a mind means knowing how to build one, then we can reframe the history of philosophy as a series of recipes, or attempts, to explain our mental lives by identifying characteristic mental achievements and then listing the parts and processes that should come together in the head to produce them. In this essay, I trace that tradition from Aristotle to contemporary **cognitive science**, showing how the same explanatory template reappears across centuries: first, define what minds do, and then posit inner **mechanisms** to make those **functions** possible. But I end up arguing that this approach, despite many advancements (particularly in related fields of psychology and neuroscience), makes a recurring mistake. The mistake is that all these recipes start with the same assumption: that a complete **model** of a single, individual mind is all that's needed to exhaustively explain our cognitive achievements. This misses out on the important role that our social and group contexts play in our cognitive lives, one that might require that we rethink our longstanding, individualistic approach.

Trying to determine what minds are, much less how to go about building one from scratch, is a difficult job. It may be best to start small, with a toy example. Suppose your friend gives you an old instant camera—the kind where you insert a cartridge of film, compose your picture, snap the shutter, and after some whirs out comes a plastic card onto which your image shortly appears. If you're curious you might try to figure out how the contraption works. It has certain inputs, most obviously the cartridge and light from the scene. One can interact with the camera by messing with some parameters using a few knobs and sliders that control the amount of light let through the lens, not to mention snapping the shutter. Finally, it has some outputs: first, it ejects a dark slide once a new cartridge of film is inserted, and then after each exposure it ejects a plastic card holding an image. But suppose you want to know *how* the image gets on to and appears on your film. You could take apart the camera, taking stock of all the bits and bobs of glass, plastic, and metal. You could (not recommended) open the cartridge and see what's going on during an exposure in broad daylight. Some educated guesses and background knowledge about the nature of light and optics, the history of cameras and photography, not to mention the magic of film chemistry, would certainly help. Eventually, you might figure it out: an INSTANT PHOTOGRAPH is created when LIGHT *focused* by a LENS *hits* sensitized layers of SILVER HALIDES and DYES contained in a PLASTIC SHEATH, which is then *developed* into a visible image by a *mechanical process* that releases DEVELOPING AND FIXING AGENTS that precipitate through the layers.

Apologies if you're not into analog photography. The point of this exercise is to get us to think about how we explain **agent**-like systems. Agent-like systems—from grasshoppers to multinational corporations—are systems that are located in time and space, and that seem to behave in regular ways: taking in inputs, adjusting parameters, doing their magic, and then issuing

outputs. Explaining these systems involves constructing an **ontology**, or categorization scheme, that makes sense of why the system behaves as it does. Specifically, constructing an ontology involves doing two things: identifying and categorizing the kinds of behaviors that an agent is involved in, and proposing a set of ENTITIES (along with their *functions*) that should, when in good working order, give rise to the behaviors we're interested in explaining. For our camera, then, we're identifying a behavior—producing an image—and we're proposing a set of entities—like the LENS and the DYES—along with the functions, including *focusing* and *development*, that, taken together, offer an explanation of the way this agent-like system, the instant camera, works. Obviously, we could go into more detail: we could drill down into the quantum nature of photons and how their interaction with the silver halide crystals frees electrons to form silver ions, which are revealed by the developer, and so on. Likewise, we could pull back and just offer a 'sketch' of an explanation: light goes into the box, photos come out.

What constitutes a *good* explanation is another deep topic of discussion, but we can start with Dretske's maxim that frames this chapter: *if you can't make one, you don't know how it works*. More specifically, the idea is that knowing how to build a thing—like a camera—is a necessary condition for understanding how it works. It's not quite like one must be physically able to build the thing in question to know how it works; despite our best efforts no one is going to build a star anytime soon, despite having a good guess as to how one would go about it (hint: get a bunch of hydrogen).

I also think most people would agree to a Goldilocks heuristic (or rule of thumb) here. A Goldilocks heuristic says to seek the explanation that's *just right*. The sketch of how the camera works (light goes in, photos come out) isn't much of an explanation; it just restates the problem. But the quantum-level explanation is too complicated; it risks multiplying the entities and functions in our description without giving us a better handle on the agent's behavior. An explanation that tells us something we don't already know, pitched at a level of detail that we can understand, seems to be just right: it can appreciably add to our knowledge of how the agent works and maybe even allow us to begin to build our own version of it.

Pulling these threads together, then, we can start to sketch an approach to the daunting challenge that we staked out at the beginning: how to build a mind? Minds are associated with agents—creatures that exist in localized times and spaces. We can start by categorizing the kinds of behaviors we think are characteristic of minded agents, and then propose the ENTITIES (like thoughts and feelings) and *functions* whose interaction should give us a story about how those agents manage the behaviors we're interested in. Throughout this process, we can evaluate the model we're building with the philosophical tools at our disposal: are we giving a recipe for how to build minds? Are we seeking the explanation that's *just right*, constructing the model and its parts with enough (but not too much) detail? The hope is that if we do this carefully enough, we'll have just what we're after: an explanatory model of the mind.

2. A history of cognitive ontology

Of course, we're not the first to attempt to give a story about how minds work. In fact, we can situate what we're doing within a long lineage of accounts that in the Western tradition can be traced back at least to the ancient Greek philosopher Aristotle and his work *De Anima*.

In *De Anima*, Aristotle models the soul—the rough equivalent of what we'd call 'mind'—in terms of **faculties**. There are some faculties we share with all creatures, like our nutritive faculty (for eating, growing, and reproducing), whereas others we share with smaller sets of animals, like the sensitive faculty, which comprises our five senses along with our 'common' sense that integrates multimodal inputs (not to be confused with 'street smarts'). The faculty of intellect, though, is uniquely human, and related to this faculty are its aspects of imagination, thought, and what we might term motivation and **desire**. We can't get into the weeds here, but a general sketch of the view goes as follows: **stimuli** from the world leave their mark on our senses—or if you want to be literal about it, our senses take on aspects of those stimuli in the world. Imagination helps retain those marks so we can keep them in mind while we think about them. In turn, these ideas can lead to desires that motivate us to act in the world. It's a simple view, but it's remarkable how many similarities we can spot between this and more recent accounts.

For instance, take the Persian philosopher Ibn Sina's ventricular model of the mind, which he adapted from the works of Galen in the 10th century AD. This model locates many of the faculties identified by Aristotle into specific parts of the brain, anticipating our more recent drive towards **neural localization** in cognitive science. More accurately, it's not the brain that Ibn Sina is interested in, but the fluid-filled cavities at its center that help prop it up. In his model, information flows—quite like a fluid—from the anterior to posterior ventricles, with each one housing a different faculty related to **information processing**. First, sensory traces enter from the sensory nerves to the common sense at the frontmost ventricle, and some of those traces persist in the imagination, which was co-located with common sense. A small organ, the *vermis* or worm, would direct the flow of fluid/information to the 'higher' faculties of 'cogitation,' which enable us to reason with the traces preserved by the imagination. If we needed to retain information for a long time, it would flow to the rearmost ventricle which housed our long-term **memories**. Interestingly, the rationale given for the progression of information was that the fluid in the rearmost ventricles (memory) was 'drier' than the frontal ventricles, and this was thought to explain why we can retain memories far longer than most sensory traces.

While popular during the Middle Ages, the rise of anatomy as a science quickly doomed the ventricular model, with Vesalius's dissections showing that there was no such 'worm' organ that could adequately serve the mechanistic role that ventricular theorists gave it. Worse yet for the ventricular model, he showed that other animals have a similar ventricular morphology,

but (were thought to) lack cogitation or other ‘higher’ aspects of minds; hence, the ventricular model couldn’t sufficiently account for our mental abilities.

With the rise of enlightenment thinkers in Europe there came another wave of models, less focused on mapping functions to hypothetical entities or parts of our brain, but which instead aimed to chart the laws and structures that govern the mind’s operation. The German philosopher Immanuel Kant describes a systematic architecture of thought with aspects corresponding to the inputs (sensibility), integration (imagination), organization (understanding), systematization (reason), and unification (apperception) of thought. In contrast, the Scottish philosopher David Hume attempts a ‘mental geography’ that outlines how a few basic elements, ideas, and impressions are organized by laws of association and the abilities of memory and imagination give rise to our mental lives. By the 19th century, these two schools of thought generated competing theories that had begun to be tested in the earliest psychology laboratories, which attempted to identify the operation and boundaries of many of these concepts in real life with real participants. It’s also at this point that we see the reemergence of a **reductionist** impulse to find the real neural entities responsible for these abstract mental terms, like imagination and memory. The American philosopher and psychologist William James is particularly clear on this aim: “all that psychology can do is to seek to determine *what* the several brain-processes [that occasion knowledge] are; and this, in a wretchedly imperfect way, is what such writings as the present chapter have begun to do” (p. 689).

Despite the various terms, aims, and theoretical motivations of these authors, we can try to bring some order to this history. Though these views obviously differ and each deserves its own deep dive, we should recognize how they all roughly conform to the template set out above. First, through **introspection** or observation these authors try to identify and categorize the tasks that minds are involved in, tasks including sensing the world, remembering past events, and reasoning about future actions. Second, each view proposes a set of entities, which could include physical things like ventricles or abstract entities such as ideas and functions, and then to various degrees they attempt to model how those bits work together within an individual to produce the mental achievements that we take the mind to be responsible for.

We might term each of these disparate views a **cognitive ontology**. They’re ontologies in that they characterize the constituent ENTITIES and *functions* of a domain of study, in this case minds. These views are also all arguably cognitive, in that they decompose the mind into smaller ENTITIES, usually thought to reside in the head, whose operations aren’t readily explained by behavior alone.

3. What we’re doing now—the case of working memory

This template can also roughly characterize the project that many psychologists, neuroscientists, cognitive scientists, and philosophers are doing to this day.

Suppose you meet someone who gives you their online handle; how might you go about remembering it? If you have your phone on you, you're lucky as you can just type it in the relevant app right away; but suppose you don't. You could write it down on a sheet of paper. But suppose you're really out of luck with no paper to be had; what might you do? You might repeat it out loud, like a mantra, until you get to your phone or a pen and paper. That might be awkward, so instead you might repeat it using **inner speech**. If you suddenly get distracted, say by seeing a friend across the street, you risk losing that information, so you have to keep your attention focused on the thought. It seems simple, but this remarkable ability we have to keep thoughts derived from our environment alive in the fore of our minds, so that we can do something with them a bit later, is the core of one of the foundational constructs in contemporary psychology: **working memory**.

How does working memory map onto the features we distilled from earlier accounts? Through a process of introspection (or observation if you see others engage in this ability), we've characterized a mental achievement: the ability to hold things in mind and manipulate them for our goals. We have then started the second step, positing ENTITIES and *functions* that might produce this achievement: we have an initial stimulus to be remembered, a goal to do something with it, and a suite of hypothesized, internalized mechanisms, including ATTENTION and INNER SPEECH, that get us from the stimulus to the goal. Finally, we unite this set of entities and functions under a moniker which helps us organize the tasks that we think test this achievement and the litany of results that they produce, in this case with the label 'working memory.'

Notice that, at least at this level of detail, working memory retains significant similarities with other labels from past theories. Arguably it does much the same job as Aristotle's own use of 'imagination' or Ibn Sina's ventricles. However, in contrast to ancient views, theories of working memory benefit from the methods of contemporary psychology and other sciences of the mind: experimental and statistical best practices, response-time measures, infant and animal studies, not to mention the whole suite of tools from neuroscience and computational modeling. With these advancements, we're arguably in a *much* stronger position to drill down and exhaustively map how constructs like working memory *explain* the successes that are characteristic of our minds.

How can we keep information, like your new friend's handle, in mind? Well, we can use INNER SPEECH to *maintain* that INFORMATION over *delay periods* until we can complete the task. Why is your roommate better at learning math than you (or chess or guitar, whatever it may be)? Perhaps it's because they have a greater WORKING MEMORY CAPACITY. Even better: we can correlate the maintenance of information in working memory with increased neural activation, in the form of **neurons** in the prefrontal cortex increasing their rate of **action potentials**, and we can show that disturbing this activation can often lead to measurable impacts on your ability to remember the relevant information (your friend's handle, remember?). It seems like we've

got a good roadmap to get from the **intuitive** models of the mind that we've been discussing since antiquity to an actual recipe that we might be able to follow—something that could meet those explanatory criteria that we set out at the outset. We might be edging ever closer to being able to *build a mind*.

So, what's wrong with this picture? After all, it seems to be participating in a long and rich lineage of similar models, each with their **ENTITIES** and *functions* situated in such a way as to give us a recipe for how we might end up at a mental achievement. In fact, if we recall our two rules from the start of this essay—the Goldilocks heuristic and Dretske's maxim—it seems that we're doing well by those two metrics. Want to build a mind in such a way that it can hold a small amount of information that's no longer in the environment but that is important to help guide future behavior? Well, it'd be good to start with dedicated mechanisms whose operation is fine-tuned to the functions that we can identify in the mental achievement that we're interested in. In the case of working memory, it seems quite promising that we have found mechanisms—like increased neural firing—whose processes reflect those functions, such as maintenance of information. We also seem to have approached the thing we want to explain—this ability to maintain and manipulate information—at just the right level: not in a manner that's too abstract or too fine grained. By our own standards, then, it seems that we're well on our way to understanding and being able to model an important piece of our mental lives—one that has analogues in nearly every model of the mind that we've reviewed.

4. Is this model of working memory actually *just right*?

It may be tempting to chalk up this model of working memory as a win and move on to trying to explain the many other **capacities** that we've posited to help explain related mental achievements. In the same neighborhood of cognitive science as working memory we can see scientists using related terms like attention, executive function, motor intentions, and so forth. But before jumping to the next project, it's worth pausing and asking just how successful our story about working memory is. We have a characteristic mental achievement—holding information in mind and doing something purposive with it—and we've decomposed that into a series of functions, like the *maintenance* and *manipulation* of **INFORMATION**, whose activity we've attached to entities like **INNER SPEECH** that we've in turn correlated with **neural processes**, including the increased firing of neurons in the prefrontal cortex. It seems like we have the whole explanation! But we should ask: is this mechanism sufficient for working memory performance? Is it necessary? After all, this mechanism could underwrite other cognitive functions, like attention or preparation for action, and it's also possible that other systems in other parts of the brain can also support these mental achievements.

Indeed, mounting evidence has accumulated against our elegant 'prefrontal dogma' view of working memory almost since it was inaugurated in the

mid-1990s by the neuroscientist Patricia Goldman-Rakic. Lesion studies on animals and studies of patients with damage to the prefrontal cortex show that the activity of neurons in the prefrontal cortex isn't strictly necessary for successful working memory task performance. At the same time, it looks like this brain activity is important for other, related, capacities. That is, we see similar patterns of increased firing in the frontal cortex in the context of what's been termed the 'frontal parietal network,' which covers a good chunk of our neocortex. Activity in this network can be correlated with success not just on working memory tasks, but also tasks of other supposedly separate capacities like motor intention (planning where you will move your eyes) and 'covert' attention (attempting to focus your attention to a zone in your visual field without moving your eyes to it—try it, it's tough).

On the one hand, this is a disappointing turn. The idea that we could tie a capacity that we seem to rely on a good bit of the time—even reading this sentence requires that you can maintain traces of it in your mind to comprehend its meaning—to an elegant neural process that lends itself to being modeled with computational tools would have been an impressive scientific breakthrough. And it wouldn't have been that outlandish a scenario: many key breakthroughs in our understanding of how neurons encode information had a similar explanatory narrative (like Hubel and Wiesel's Nobel Prize-winning work on the cat retina in 1959). On the other hand, it's not so surprising. The functions that we've associated with working memory are abstract; what it means to *maintain* or *manipulate* INFORMATION for goal-driven behavior is rather vague. It's not obvious that we can unify all the ways we maintain or manipulate information across working memory contexts: from rehearsing a phone number using inner speech, to using our 'mind's eye' to mentally walk through a maze, to recalling a recipe or reading this chapter. It would be great if such diverse exercises of this capacity all bottomed out in a single elegant mechanism. But we shouldn't count on the mind working that way.

Instead, a more plausible view might prioritize cataloguing and explaining the many mechanisms that we rely on to hold and transform information in the service of our goals. In fact, such a view wouldn't be alien to some of the earliest theorists of working memory, who hypothesized that working memory may be manifest in diverse ways both in and out of the head. I call this a mosaic ontology. Just like an artistic mosaic that builds a picture from thousands of individually distinct chips of ceramic, perhaps we can best model a more abstract ENTITY like working memory by detailing the many, distinct pathways that brains take to *realize* working memory's twin functions of maintenance and manipulation. Of course, figuring out all the plausible pathways that brains can take to maintain and manipulate information is a tall order. (A search of the PubMed database reveals that 70,000 articles have been published on the subject of working memory alone!) But by automating aspects of this process using **artificial intelligence** tools for

natural language processing, we might be able to produce such a mosaic-like ontology for many similar constructs in cognitive science.

5. Where does cognitive ontology go from here?

Arguably, however, the situation for working memory and its ilk may be even worse, and even this move towards pluralism—accepting that working memory may be tied to *many* mechanisms—might not give WORKING MEMORY a place among the entities that figure in our best ontology of mind. Russ Poldrack, a philosophically oriented cognitive neuroscientist, has led a project called the *Cognitive Atlas*, which aims to systemize our ontology of cognition by tying neural discoveries to functional descriptions and psychological constructs. Interestingly, other fields have similar ‘ontological’ database projects, like the *Gene Ontology* project in biology. Poldrack used natural language processing tools to determine how many terms in the *Cognitive Atlas* could be traced to William James’ 1890 *Principles of Psychology* and found that something like 23% of terms in the contemporary ontology could be found in James’ work. And a quick review of the chapters and content of James’ text bears this out: many terms—from ‘memory’ to ‘attention’—and experimental paradigms resemble those of contemporary psychology. Comparing this to a similar range of 19th century biology texts, Poldrack found that less than a single percent of terms in the *Gene Ontology* could be similarly traced back. Of course, this makes some sense, since between the late 19th century and today, we’ve made enormous progress in our understanding of biological systems; important topics like molecular genetics and genomics weren’t even close to being studied when classic biology textbooks were being written in the 19th century. In contrast, many psychological constructs, like memory and attention, share a similar form to those not only of the 19th century but, as our earlier review foreshadowed, even ancient views of the mind.

Poldrack uses these and related findings to motivate a dilemma. Either we have special access to the structure of our minds, such that these long held terms capture something roughly true about how our minds work; or, psychology is due for a serious conceptual revolution that will obviate many of the terms derived from our **folk psychological** understanding, much as occurred with biology after the discovery of the genome, or physics after the introduction of quantum mechanics.

If it’s the first case, then it’s not surprising that Aristotle, Ibn Sina, and contemporary psychologists should all share similar terms and roughly similar models of the mind. After all, if we’ve had access to the structure of our minds all along, then it only takes some studied reflection to trace its contours and begin to hypothesize how these entities and functions fit together. In fact, this is precisely what Hume and others say about their method: we introspect our own **experiences** and reflect on others’ behaviors to posit entities and functions that explain those appearances.

If psychology is due for a conceptual revolution, on the other hand, then we're stuck in a waiting game for the right discovery. While some philosophers have hoped that neuroscience would **eliminate** our ordinary ways of speaking about the mind, we aren't replacing our talk of **beliefs** and desires with physical and chemical descriptions of neural activity yet. Just as proponents of Mendelian genetics thought there must be a hereditary carrier of genotypic and phenotypic traits (remember those Punnett squares from high school biology class?), it wouldn't be until our discovery of chromosomes and DNA that we'd have the language and concepts needed to sufficiently explain these **phenomena**. It may be that we need a similar revolution in science that will lead us to use radically different terms and deviate from our everyday introspective understandings in order to make progress in understanding cognition.

In some sense, this is an unfair comparison. After all, we *have* made a lot of progress in understanding how various mental phenomena work. Hubel and Wiesel showed how retinal cells encoded information in action potentials, and we know a lot about how our motor systems encode and act on information (Lu and Ashe, 2005). We've made substantial progress in understanding how brains produce a lot of perceptual and motoric phenomena: we can even make leg prostheses for patients who have lost limbs that use proprioceptive feedback for fine-grained control of walking and other motor tasks (Song et al., 2024). We also are on the brink of a new generation of neural implants that may help us communicate with locked-in patients, who otherwise might be diagnosed as comatose or vegetative (Zhang, 2025).

Perceptual and motor phenomena are intimately tied to behavior and the world, and it's this 'objective' aspect that may give us greater leverage in building explanations of these phenomena. In contrast, cognitive processes—if we're willing to cleave cognition from **perception**—don't need to eventuate in any behavior or outward signals at all. Thought can occur without overt action. And in our **idealized** models of thought and its related capacities, like attention, memory, and learning, these processes occur paradigmatically in individual minds.

6. Towards an anti-individualist cognitive ontology

This observation gives us a critical hint at one way we might go about improving the situation. Recall that we begin the process of modeling the mind by introspecting or observing mental achievements; for instance, when doing 'mental geography' we look inward to try to map the contours of our mind. When picking out paradigmatic achievements we often turn to the exceptional: inventors, discoverers, masters of reasoning, and so forth. And this sets up a fundamental assumption at the heart of our models: that modeling the mind is something that's best done at *an individual level*. A single mind should be capable of discovering the genome or inventing calculus. Hence, we start from this point and then begin positing the relevant entities and

functions that should (ideally) allow for such achievements, from reading this page to flexible creative thought.

In turn, some of the cognitive capacities we identify end up making a wide berth over our cognitive lives. Take *learning*, *imagination*, or *working memory*. By their very generative nature, they seem implicated in many if not most robust cognitive tasks. Think of a task that *doesn't* involve the maintenance or manipulation of information for purposive behavior—it's hard. The wide scope given to these abstract capacities may help explain why it's so difficult to give a description that meets our explanatory aims. While working memory incorporates the latest advancements from neuroscience, it doesn't give a substantively better explanation about how our cognitive achievements come about. Rather it just restates our prior intuitions about how our minds work in new terminology. This effectively echoes a similar cautionary note that the philosopher and psychologist G. T. Ladd made regarding the then new science of psychology in 1887, warning that

the application to mental phenomena of uncouth terms derived from the physical sciences ... has simply the effect of repeating certain [mental] facts and laws in a less appropriate way, without adding an item of information regarding the real nature of their physiological basis (p. 534).

One option, then, might be to critically evaluate this assumption at the heart of our cognitive ontologies: that our cognitive achievements are best modeled at the level of individuals. What might this skeptical take look like? Well, some of these achievements may instead be best explained at a *social* level. Take deliberation, for instance. Clearly, we can do this on our own: we can use the powers of our mind to weigh options and decide on a course of action. We might even use inner speech to play out a dialogue weighing two or more choices. But in many contexts, we may maximize our deliberative ends by doing so as part of a group. This is related to the 'wisdom of the crowds' effect, where an average of many different people's guesses can often provide an accurate estimation, for instance of how many pennies are in a large jar. Discussing difficult decisions in a diverse group of agents, each of whom has different stores of knowledge, may help even out the biases that affect us at an individual level. There are also **evolutionary** reasons to think that deliberation and reasoning came about through group processes and collective forces. We might then prefer to think of our achievements of reasoning as paradigmatically occurring with others and only derivatively occurring in the minds of individuals using inner speech and other mechanisms.

What does this move to the social buy us? At least two things: first, it gives us new terrain to survey and target when starting the task of modeling the mind. By treating some achievements, like reasoning and deliberation, as primarily social, we may be able to generate new entities and functions that better explain their characteristic successes and shortcomings. Second, it allows for a kind of **epistemic** humility and flexibility that may not be present

in contemporary models that situate all our achievements as products of individual minds. Perhaps the entities we posit as part of our story to explain the mind don't have to be so well equipped and ultimately flexible. They may instead be suboptimal, provided that there are sufficient social scaffolds and practices to support their operation. We may, essentially, be dumber than our idealized model agent. To put it in the language of working memory, perhaps it is best understood as a plural phenomenon that can happen both *inside* our minds and as something we do with others through language and culture. Though it sounds radical, such a view wouldn't have been unfamiliar to some of the earliest theorists of working memory who claimed that

the kind of working memory that people prefer to use when they are executing a Plan seems to represent characteristic differences. One person will insist on writing things down, running his life from a calendar pad, whereas another person will keep in his own head everything he intends to do" (Miller et al., 1960, p. 120).

While psychology may still require a conceptual revolution, we can already do better by loosening the grip of individualism. Many of our most impressive achievements—reasoning, deliberation, memory—are not solitary feats but socially scaffolded practices. Recasting them as plural phenomena, enacted across brains, bodies, and groups, enlarges our ontology and grounds our explanations in how minds actually succeed. If building a mind has seemed an impossible task, it may be because we have been looking in the wrong place. We should try looking outside the head, to the shared spaces where thinking lives.

RESPONSE TO GOMEZ-LAVIN: HOW SHOULD WE STUDY SOCIAL MINDS?

Tim Elmo Feiten and Anthony Chemero

Javier Gomez-Lavin combines an engaging introduction to the project of cognitive ontology with a thought-provoking suggestion: what if our intense and diverse efforts to study the mind have hitherto been hampered by a shared underlying assumption that thinking is something that happens inside of individual minds, rather than an activity that consists in the interactions of minded organisms in shared environments—a fundamentally social phenomenon?

This proposal raises larger questions about what cognitive science would look like after a 'social turn,' and our goal here is to point out some of these questions for further exploration. We are largely sympathetic to the proposal, and want to provide additional context and distinctions that may help clarify how big Gomez-Lavin's challenge is, and the different ways in

which we might approach it. Among Gomez-Lavin's contributions are: a) the approach of explaining what a mind is by building one; b) an account of cognitive ontology as a transhistorical research project; c) a critique of the current state of functional localization in neuroscience; and d) a proposal for a social turn in cognitive science. We will touch on each of these, but in a different order. First, we want to reinforce Gomez-Lavin's concerns about the state of functional localization, but add that the massive amount of data that has been collected encodes the methodological limitations in a way that cannot be washed out by large-scale computational analysis. Second, we supplement the history of cognitive ontology with a brief sketch of social views of cognition in the 20th century, which might inform the proposed social turn. We then highlight how this social turn dovetails with recent methodological critiques within neuroscience and comment on the role that computational tools and **metaphysical** assumptions play in designing a social turn.

The lion's share of the empirical data that has been collected by cognitive neuroscientists has been collected under the assumption that cognitive functions are carried out by specialized neural structures. The aim of the project of localization has been to conduct experiments for the sake of establishing mappings between these functions and structures. Assessments of this project as a whole have concluded that **one-to-one mappings** are generally not possible, and the collective results of cognitive neuroscience instead show many-to-many relationships—neither are the neural structures specialized, nor are the cognitive functions localized in the way we had imagined. But does this really show that localization is misguided, or did we just look for the wrong functions? Maybe we could replace the 19th century concepts we have inherited for talking about minds with new ones, which *can* be localized in the brain? Gomez-Lavin suggests that those novel concepts might be mined from the massive body of published papers in cognitive neuroscience, but we want to add a grain of skepticism here. The empirical data available to us is not agnostic to the conceptual assumptions that guided its collection. Quite the opposite: at every stage, from the design of tasks and studies, to the collection, cleaning, and analysis of data, and not least of all in the writing of the actual papers, these conceptual assumptions shape science and its outcomes. The existing archive of research is a gigantic achievement, but it is also a reflection of the specific history of functional localization rather than a neutral starting point for computational meta-analysis. Worse, this problem gets more severe the more seriously we take Gomez-Lavin's call for a social turn: cognitive neuroscience as an experimental science depends almost entirely on highly constrained, artificially sparse, largely contextless tasks performed by individuals in a lab setting. To the degree that minds are social, no AI-driven analysis of past research outputs will find them, because they just aren't to be found in the place from which we've been getting (almost all of) our data.

If the arguments for the social turn in cognitive science are correct, we need new concepts for explaining mental processes and new methods for

doing empirical research. Fortunately, we don't have to start from scratch, but can instead draw on some historical precursors to the kind of view Gomez-Lavin advocates for. Like cognitive ontology, the social view of cognition has its own history and was especially popular in the years leading up to the **cognitive revolution** in the latter half of the 20th century. We sketch only the most important aspects very briefly, because they provide some rich material for imagining what a future cognitive science might look like that recognizes its subject matter as essentially social. **Phenomenological** philosophers argued that having experience at all depends on the presence of others. Edmund Husserl and Maurice Merleau-Ponty argued that human thinkers are "in one another" (*Ineinander*) so that individual minds are intertwined with one another. Martin Heidegger argued that being a human subject is first and foremost being with others. **Pragmatist** thinkers like John Dewey and George Mead thought that thinking and learning are deeply wrapped up with talking to others. Among the things that Dewey learned in his observations of children at the Laboratory School at the University of Chicago was that learning is intensely social. Children learn by spontaneous doing, always accompanied by a lot of talking and listening. Mead learned the same lesson in his work at Chicago's settlement houses. Mead ties the emergence of the individual mind closely to the development of linguistic abilities: "out of language emerges the field of mind" (Mead, 1934). Perhaps surprisingly, at the same time, but far from the settlement houses and the capitalist context in which pragmatism emerges, Soviet psychologist Lev Vygotsky came up with an understanding of the mind and self very similar to Mead's. Vygotsky thought that individual mental lives resulted from the internalization of cultural tools during development. Like Dewey, Vygotsky understood learning as an intensely social process. This is exemplified in what he calls the 'zone of proximal development,' which is the difference between what a child can do on their own and what they can do with the help of adults or older children. As infants, children can do essentially nothing on their own, and can only get things done with the help of a caregiver. That is, all their early problem-solving skills are social. There is also significant evidence from the cognitive sciences that cognitive systems are necessarily social.

The claim that cognition is a social phenomenon comes in different versions. A stronger version holds that cognitive phenomena are literally **constituted** by processes that occur among and between groups of humans, rather than being confined to what happens inside of individuals. A weaker version might claim only that those social processes play a very important **causal role** in shaping cognitive phenomena (wherever we take them to be located). It's important to note that the social turn, as a methodological course-correction, would still be necessary even if only the weaker version were true. By reducing the range of cognitive phenomena to just those that occur when isolated individuals are presented with tasks in a lab, we would still be excluding from our investigation the main form in which our phenomenon of interest occurs in nature. This criticism dovetails with calls for cognitive science to

take seriously the fact that animals are living, embodied subjects inhabiting complex environments (this is also the context in which the causation/constitution debate takes place). Most current cognitive neuroscience involves studying very artificial cognitive phenomena in the lab. This is because the process of isolating explanatory factors for a cognitive achievement requires ruling out the myriad other concurrent influences that make up the complex contexts in which cognition occurs naturally. Outside the lab, cognition involves moving our bodies through environments in loops of action and perception, rather than sitting still and responding to individual stimuli, and variability is a tool that helps us achieve our goals, instead of a source of noise that must be tamed via statistical analysis. This means that in order to build better models of the social mind, we will also need to come up with different kinds of experiments than the ones that have traditionally dominated laboratory science. This is a big challenge, but it also comes with some advantages: because it views cognition as information processing that happens inside the brain, orthodox cognitive science often has to posit complex **computations** to explain behaviors that appear far simpler once we take into account how bodily movement is not just reactive to sensory inputs, but shapes future inputs in a closed loop (see, for example, Gomez-Marín and Ghazanfar, 2019 on rats; and Vasas and Chittka, 2019 on bees).

The different role played by computation in these two approaches to cognitive science—broadly, orthodox **cognitivism** and **embodied cognitive science**—brings us to our final point: building a model of a mind may not be the same thing as building a mind. When outlining the necessary steps for “how to build a mind,” Gomez-Lavin ends up with “an explanatory model of the mind.” But what are we really building, a mind or a model? On one popular metaphysical view, called **computational functionalism**, building a model that computes the same functions as a human mind would result in a model that is literally also itself a mind. We think it’s important to highlight that this only follows from this peculiar metaphysical view. After all, we generally don’t think that scientific models are automatically instances of the phenomena they **represent** or explain, no matter how good they are at doing this job. No matter how good a computational model of fluid dynamics we develop, the model itself won’t be wet. So why would a computational model of a mind count as a mind in its own right? The reason is that orthodox cognitive science assumes that cognition *just is* a form of computation: if the model performs the right kind of computation, it is by definition engaging in cognition. Importantly, in the history of cognitive science this computational view is closely tied to the focus on individuals. In contrast, the social views of cognition we have mentioned above are far more sympathetic to the newer forms of embodied cognitive science, which do not (generally) subscribe to the metaphysical view of computational functionalism. This is all to say that, within a cognitive science that is built atop a social view of minds, building a model of a mind may very well not count as building an actual mind.

We agree with Gomez-Lavin that a social turn in cognitive science is necessary. Some aspects and components of such a new approach already exist, both in early-to-mid 20th century social views of the mind and then in embodied alternatives to classical, computational cognitive science. Although we raised some doubts about the plan of using computational tools to excavate the building blocks for a new paradigm from the archives of the old one, we don't intend these doubts as a general skepticism towards new roles for computational tools (such as **machine learning**). We do think that new forms of experimental research will be necessary to capture the rich reality of social minds, but machine learning tools are beginning to show some potential in this area, too, where they are combined with concepts from embodied cognition to enable multimodal analyses of complex behaviors (e.g., combining data from wristbands, eye-tracking, and cameras in a single analysis).

Gomez-Lavin is guided by Dretske's maxim: "if you can't make one, you don't know how it works," but clarifies that this doesn't necessarily mean that we have to be able to physically assemble the system we are studying (otherwise we'd never be able to understand how the sun works, for example). We'd like to suggest that the social view of minds adds another quirk to our understanding of Dretske's maxim: if minds are truly social, all of us are always already jointly engaged in making them. For our understanding of what minds are, building models might then just be one of many scientific and philosophical avenues of inquiry that can help reveal the astonishing richness and complexity of the minds we make together.

Discussion questions

1. Dretske's maxim states that "if you can't make one, you don't know how it works." What does this maxim mean, when applied to the philosophy of mind? Do you think it's true?
2. The Goldilocks heuristic states that philosophers and scientists should seek explanations that are "just right": neither too sketchy nor too minutely detailed. Think about something you know really well (maybe how to cook your favorite dish, or how to play your favorite sport). Imagine needing to explain that thing to a novice. Try to come up with three explanations: one that would be too sketchy for that purpose, one that would be too detailed, and one that would be just right.
3. Now think about scientific explanation. What reasons do scientists have for following the Goldilocks heuristic, rather than always giving explanations that are as deep in the weeds as possible?
4. In section 2 in his essay, Gomez-Lavin provides a whirlwind history of cognitive ontologies. What's the main point or upshot of this history, for Gomez-Lavin's purposes?
5. Gomez-Lavin suggests that we try to "think of a task that *doesn't* involve the maintenance or manipulation of information for purposive

behavior—it's hard." Can you think of one? What conclusions can we draw about working memory based on the difficulty of this exercise, if any?

6. Feiten and Chemero note that "cognitive neuroscience as an experimental science depends almost entirely on highly constrained, artificially sparse, largely contextless tasks performed by individuals in a lab setting." Why would this fact about scientific practice make it difficult for cognitive neuroscientists to assess Gomez-Lavin's claim that cognitive achievements may be best understood at a social level?
7. Feiten and Chemero distinguish between stronger and weaker versions of the claim that cognition is a social phenomenon. What are these two versions, and what, exactly, is the difference between them? The three authors of this chapter all clearly believe the weaker claim. Do you agree with them? Do you think the authors also believe the stronger claim? Do you? Why or why not?

Suggestions for further reading

For Dretske's discussion of his maxim:

- Dretske, F. (1994). "If You Can't Make One, You Don't Know How It Works." *Midwest Studies in Philosophy* 19 (1), 468–482.

For a deeper dive into Gomez-Lavin's views on working memory:

- Gomez-Lavin, J. (2021). "Working Memory Is Not a Natural Kind and Cannot Explain Central Cognition." *Review of Philosophy and Psychology* 12, 199–225.

For more on the relationship between the methods of cognitive neuroscience, the practice of neural localization, and the project of cognitive ontology:

- Poldrack, R. A., and Yarkoni, T. (2016). "From Brain Maps to Cognitive Ontologies." *Annual Review of Psychology* 67, 587–612.
- Gomez-Lavin, J. (2024). "From 'Blobs' to Mental States: The Epistemic Successes and Limitations of Functional Magnetic Resonance Imaging (fMRI)." In Heinzelmann, N. (Ed.), *Advances in Neurophilosophy*. Bloomsbury Academic.
- Anderson, M. (2014). *After Phrenology*. MIT Press.

For studies of the role of social factors in the evolution of the human mind:

- Mercier, H., & Sperber, D. (2011). "Why Do Humans Reason? Arguments for an Argumentative Theory." *Behavioral and Brain Sciences* 34 (2), 57–111.
- Sterelny, K. (2012). *The Evolved Apprentice: How Evolution Made Humans Unique*. MIT Press.

For a fuller account of Chemero's views on how cognition is a social phenomenon:

- Chemero, A. (2026). *Intertwined Creatures: The Embodied Cognitive Science of Self and Other*. Columbia University Press.

For an introduction to the philosophical tradition of phenomenology, including the claim that the possibility of experience depends on the presence of others:

- Käufer, S., and Chemero, A. (2021). *Phenomenology: An Introduction*, 2nd ed. Polity.

For an introduction to the philosophical tradition of pragmatism, and especially its relationship to the sciences of the mind:

- Gallagher, S. (2022). “Pragmatism and Cognitive Science.” In Aikin, S. (Ed.), *Routledge Companion to Pragmatism*, 239–251. Routledge.

For a study of how machine learning tools can be combined with concepts from embodied cognition to help scientists understand complex behavior:

- Feiten, T. E., Holland, K., and Chemero, A. (2022). “Scalar Symmetries and Strange Tools: Learning Analytics and Embodied Design as Techno-Cultural Confluence.” *International Journal of Child-Computer Interaction* 31, 100452.

Additional references

- Goldman-Rakic, P. S. (1995). “Cellular Basis of Working Memory.” *Neuron*, 14 (3), 477–485.
- Gomez-Marin, A., and Ghazanfar, A. (2019). “The Life of Behavior.” *Neuron* 104 (4), 25–36.
- Hubel, D. H., and Wiesel, T. N. (1959). “Receptive Fields of Single Neurones in the Cat’s Striate Cortex.” *The Journal of Physiology* 148 (3), 574–591.
- Ladd, G. T. (1887). *Elements of Physiological Psychology*. Charles Scribner’s Sons.
- Lu, X., and Ashe, J. (2005). “Anticipatory Activity in Primary Motor Cortex Codes Memorized Movement Sequences.” *Neuron* 45 (6), 967–973.
- Mead, G. H. (1934). *Mind, Self, and Society*. University of Chicago Press.
- Miller, G. A., Galanter, E., and Pribram, K. H. (1960). *Plans and the Structure of Behavior*. Holt, Rinehart and Winston.
- Song, H., Hsieh, T. H., Yeon, S. H., et al. (2024). “Continuous Neural Control of a Bionic Limb Restores Biomimetic Gait after Amputation.” *Nat Med* 30, 2010–2019.
- Vasas, V., and Chittka, L. (2019). “Insect-Inspired Sequential Inspection Strategy Enables an Artificial Network of Four Neurons to Estimate Numerosity.” *iScience*.
- Zhang, S. (2025). “The Vegetative Patients Who are Still Aware.” *The Atlantic*, June.