Reproducible, Generalizable Brain Models of Affective Processes

Philip Kragel 1
Tor D. Wager 1,2✉

Email tor.wager@colorado.edu

1 University of Colorado, Boulder, CO, USA
2 Department of Psychology and Neuroscience and the Institute of Cognitive Science, University of Colorado, Boulder, CO, USA

Abstract

The abstract is published online only. If you did not include a short abstract for the online version when you submitted the manuscript, the first paragraph or the first 10 lines of the chapter will be displayed here. If possible, please provide us with an informative abstract.

Recent years have seen dramatic advancement in the measurement of biology at a systems level. In humans, neuroimaging can be used to probe the brain bases of affect and emotion in increasingly sophisticated ways, but the complexity of these measures presents new challenges in maintaining scientific transparency and reproducibility. We describe several new models of the brain bases of affective processes, including models that predict the intensity of pain, negative affect, autonomic responses, and prosocial emotions including empathic care and distress. These models reduce complex, brain-wide neuroimaging data to measures that can be readily replicated and generalized across laboratories, and they can yield correlates of affective behavior that are substantially stronger than those based on single regions from standard brain maps. They can also be used as mechanistic targets for interventions, allowing comparisons across diverse treatments. Most importantly, they can teach us about the brain representations that underlie various forms of affect, in part by providing information about the necessary and sufficient brain bases for predicting affective states and behaviors. The results across the series of studies discussed here indicate that different forms of affect have reliably different brain representations. For example, Somatic pain, romantic rejection, vicarious pain, and empathic care all have differentiable brain substrates. The latter processes are particularly important for empathy and prosocial behavior, and the chapter includes an extended example of how multivariate brain measures can inform us about how we might recognize others’ suffering and take action to help.

Keywords

Brain models
Affective processes
Scientific transparency
Scientific reproducibility
Prosocial emotions
Somatic pain
Empathetic care
Prosocial behavior

Nature has placed mankind under the governance of two sovereign masters, pain and pleasure. On the one hand the standard of right and wrong, on the other the chain of causes and effects, are fastened to their throne. They
govern us in all we do, in all we say, in all we think…—Jeremy Bentham, The Principles of Morals and Legislation

A Q 2

Pain and pleasure are primary motivating forces that underlie much of human behavior and motivation. Take pain, for example. It is defined as an aversive sensory and emotional experience. Generally, we avoid it. Many of our philosophical and religious traditions are focused around advice on how to escape, avoid, manage, or accept its reality without causing additional suffering.

A Q 3

But what is pain, exactly? Pain is multiple things. It is an experience, a motivating force, an elicitor of emotional responses, a driver of decisions. Sometimes it is, more or less, a “negative sensory and emotional experience” caused by activation in nociceptive pathways. But it cannot be only that, because it teaches us to fear pain in the future, and it drives our autonomic systems and the musculature that allows us to escape. It motivates goals, from the simple and immediate—take your hand off the stove!—to the elaborate and complex, even becoming a focal point around which one’s life is organized. It teaches us to fear it, but sometimes also to seek it, as during the expiation of guilt or when we turn to pain to relieve emotional distress. And because it is an experience, it is consciously accessible. We will never fully understand pain until we understand consciousness.

Identifying the brain representations that underlie pain is a crucial part of our understanding of it. There are many fundamental questions that remain unanswered. Is pain caused by a specific neural pathway or region? Is the thing that makes pain conscious the same as the thing that makes us conscious of the visual world, or of our own intentions? Is pain a single type of experience, with one neural mechanism, or a family of experiences with diverse mechanisms? What is the relationship between the processes that cause us to experience immediate pain in the moment and those that drive long-term avoidance? Is the aversiveness of physical pain neurologically similar to the aversiveness of other negative experiences—the empathic distress of seeing another person suffer, or the emotional heartbreak of being rejected by a lover or a longtime friend? Are pain and suffering the same thing, and must the one cause the other?

Without understanding the brain processes involved, our answers to these questions will be definitional. People use similar words to describe pain and emotional distress (MacDonald, 2009)—so by definition they are conceptually similar. But these similarities may say more about the way we organize our thoughts and the pragmatics of communication than about the deeper nature of our being.

How do we even know when someone is in pain? In one sense, it’s simple: We can ask them. But, though it is tempting to take self-report as a ground-truth measure of subjective experience, experience is private. I cannot directly observe whether you are in pain, or whether the color red looks the same to you as it does to me (Chalmers, 2007). And though people’s self-reports are often trustworthy, in many cases they fall short. If you burn your hand and you tell me you are experiencing 7 out of 10 pain, that pain report is very likely related to the degree of nociceptive activity traveling through your spinothalamic tract (and other tracts) up to your brain. But it is also related to your prior belief that you’ve been injured and that it should be painful. It is colored by the emotions you feel—are you cool and objective, or afraid of serious injury? Are you angry at yourself for putting your hand on the stove, or at me for leaving it on? And it is filtered by what you are trying to communicate. Do I look sorry for causing your injury, empathetic, understanding? Or do you need to make sure you are being taken seriously? Your “Level 7” is a communicative behavior that results from a judgment that is made relative to your past experience—how bad should a “7” be?—and your appraisal of the overall context of the situation.

If we are simply to believe all pain reports in all situations, we will have to accept a number of uncomfortable things. We will have to accept that when people report pain after a fake auto crash—a controlled experiment that looked like a crash but involved no real sudden movement—the whiplash pain they report is just as real as the pain from a real accident (Castro et al., 2001). We will have to accept that males who are strongly gender identified really experience less pain than those who are not (Alabas, Tashani, Tabasam, & Johnson, 2012). We will have to accept that if a faith healer reduces a person’s pain report with a sham surgery, magnets, or device
that manipulates “auras,” that is just as good as a drug that reduces pain by a similar amount. We will have to accept that if I give two groups scales with different anchors, and they report different pain intensities relative to those different anchors, they really feel different levels of pain (Schwarz, 1999). And we will have to accept that people with intellectual disabilities who do not communicate pain effectively (de Knegt & Scherder, 2011) are not really experiencing it in the same way as the rest of us. This kind of blind trust in self-reports is in part what underlies the mistreatment of groups thought to be “incapable” of normal pain, including animals and human infants (Fitzgerald & Walker, 2009).

### Brain Representations

The complexity of even “basic” affective processes like pain and pleasure and the limitations inherent in using self-reports as exclusive measures are fundamental issues that have held back the study of motivation and emotion since the inception of the scientific disciplines that study them. The hope, then, is that neuroscientific measures will enable us to identify neurophysiological processes that cause affective experiences—brain representations of affective states.

In some sense, studying a defined brain circuit or process that contributes to pain, or any other affective/motivational state, is much simpler than understanding pain reports (or other behaviors) as a whole. This is because behaviors emerge from the interactions among many brain processes. Identifying particular brain circuits and their relationship to overall behavior is a way of beginning to deconstruct those behaviors and thereby understand the elemental ingredients of affect and motivation.

However, the way in which we have historically approached studying the brain has, in many respects, been oversimplified in ways that do not lead to greater understanding. Some of these simplifications have been embedded in the way we analyze brain data and make inferences about the mappings between brain and mind. This chapter outlines some of these difficulties, anchoring on pain, negative affect, and empathy as exemplars. It also presents a new approach to brain-mind mapping, predictive modeling, that is gaining traction in the field and promises to help overcome some of the limitations of previous work. Finally, we discuss current progress using predictive modeling to understand some of the brain “ingredients” that contribute to pain, negative affect, and empathy, and how the brain processes involved relate to one another.

### Betwixt Simplicity and Complexity: A Middle Road

**Representations and Measures**

One of the major goals in mapping brain to mind is to develop brain measures that capture the underlying representation of a mental event. “Representation” is a theoretical construct that came out of cognitive science over the past decades. A “representation” of an object, an orange for example, is an information structure that describes the properties of the orange (orange-colored, sweet, healthy) and can be linked to actions (eat it, smell it, slice it) and similar objects (lemons, watermelon). We assume that the brain encodes such representations, so that oranges and other objects can be perceived and categorized, acted upon, remembered, and so forth. A brain representation, then, should be an obligatory information structure encoded into the brain; without it, one cannot recognize an orange. It should also be sufficient; if I activate a perceptual representation of “orange” in your brain, you will perceive or imagine an orange. Pain and other affective experiences are thought to be encoded similarly. Activating a representation of pain would create the experience of pain and activate at least some of its associated actions and thoughts; suppressing such a representation would block pain.

If we can identify a brain measure that is closely aligned with a representation, that provides a powerful inferential tool that enables testing of interventions. For example, if we can establish a measure of pain based on fMRI activity, that measure becomes a target for interventions. We can test whether various interventions modify that pattern; if they do, we might infer that they influence the brain mechanisms that generate pain (Fig. 8.1). Such tests would provide objective measures for interventions that are less complex and variable across contexts.
than self-report, and put psychological, drug, and other interventions on a level playing field, enabling direct comparisons of their effect sizes across interventions.

**Fig. 8.1**

Brain measures and brain targets: The specificity problem. Initially, neuroimaging studies defined brain “representations” of a mental construct, or category of mental events, as areas that responded to an instance of that construct. For example, brain “representations” for pain were defined as regions that responded to a painful stimulus (the image shown is a pain-related map from Wager et al., 2004). The logic was sensible: Define brain markers that correspond to a mental construct, and these would become targets for interventions. One could then compare intervention effects on those targets, characterize their changes across time, and more. However, identifying a brain representation that corresponds to a mental construct is much more difficult than we initially realized. One problem, shown in the right panel, is that individual regions or voxels are rarely highly specific for any category of mental event. The figure shows one of the most pain-selective voxels in the dorsal cingulate cortex. It is activated by approximately 200 studies in the neurosynth.org database (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011) that correspond to many different mental categories (Wager et al., 2016). Thus, unfortunately, it is a poor measure of any type of mental event, and when testing the effects of an intervention, observing effects on the cingulate cannot tell us whether it affect pain, emotional, decision, motor, or language, or social cognitive processes.
The trouble is that a representation is an information structure, not a pattern of brain activity. Work on population coding in neuroscience suggests that in many cases, representations of objects, actions, and so forth are encoded in patterns across neurons with different tuning curves (for reviews, see Averbeck, Latham, & Pouget, 2006; Kragel, Koban, Barrett, & Wager, 2018; Sakurai, 1996). But establishing mappings between measurable brain features and “representations” of mental events is very tricky and requires an extensive set of empirical tests—and, where evidence is not available, a leap of faith. Let’s say we identify a set of neurons that respond with increased firing rates to a painful event, for example a hot stimulus on the hand. We don’t yet know that those neurons “represent” anything, or what they represent. One must ask several questions to get closer to understanding what the neural patterns encode. Is the increase in activity highly sensitive to the event, meaning that they respond the same way every time? Is it generalizable across different types of painful events (heat, cold, chemical, ischemia) and across body sites? Is it specific to painful events, or does it respond to other kinds of salient events as well (touch, food, threatening auditory cues)? Is it predictive of the onset, probability, or magnitude of pain behaviors (self-reports in humans or other behaviors in animals)? Is it necessary, so that if we suppress its expression in the brain we eliminate pain behavior? Is it sufficient, so that if we activate it exogenously, we recreate pain behavior in the absence of a stimulus?

Whether we are measuring populations of single neurons or patterns of activity in human neuroimaging studies, the same principles for understanding representations apply. We can develop measures, but to know how they
relate to mental categories and behavior, we have to engage in the series of tests outlined above, and perhaps others. To the degree that a measure is validated in these ways, it can be taken as a provisional proxy for a “representation.”

Viewed in this way, it is not clear whether there is one “representation” of pain (or any other state) in the brain, or many. It is also not clear whether there is a representation of pain itself that is particular to the conscious experience of pain rather than other associated processes (avoidance, withdrawal, autonomic responses, learning, memory encoding, updating of goals). In fact, “pain” is a construct, a category that we have invented. Whether or not it is coherent at the brain level is debatable. That is, does “pain” exist as a useful neurological category, or is it a category we invented for human convenience, like the categories “furniture,” “heavy metal music,” or “actors who played in Hamilton”?

But all is not lost: The set of criteria above outline an empirical framework for testing what any particular brain measure actually measures, and we can use that to make progress in understanding which kinds of mental constructs (including “pain,” “empathy,” “reward,” or their subtypes) we can identify coherent brain measures for, and which we cannot. Humans can invent any categories they want to; some, we will find, map onto neurological systems in coherent ways because they are innate, developed early in life, or developed through human experience shaped by neuroplasticity. Such categories are likely to have predictive and explanatory power in terms of our innate tendencies and predispositions. The biology can also form a basis for agreement among scholars on what the essential constructs are and where the boundaries between them lie. Others do not; these may still be useful for human communication, but they likely will have little predictive power in terms of our innate tendencies and predispositions. And, if we invent psychological categories in a way unmoored to biology, we run the risk of simply inventing our own “convenient truths,” with little agreement on constructs, definitions, and boundaries. Our contention here is that we should anchor psychological constructs to biology, and that the set of criteria outlined above provides a way to validate psychological constructs at the brain level.

Over-Simplified Measures

In the grand timeline of the study of mind and brain, neuroimaging arrived very late on the scene, a few decades ago. It grew up in the intellectual soil of neuropsychological research conducted over the previous two centuries. One of the legacies of neuropsychology and early philosophy of mind was that brain processes are implemented in single, discrete chunks of brain (Lindquist & Barrett, 2012). Some dramatic successes identified patients with focal lesions and distinct, circumscribed cognitive deficits—in, e.g., speech production, language comprehension, perception and action (reviewed in Banich, 2004; Brett, Johnsrude, & Owen, 2002). These successes defined the field, and the way in which neuroimaging data were analyzed—one region at a time, with hopes of finding a single, key region that was crucially involved in the behavior.

A “one-region-one-function” ideology licensed several problematic assumptions. First is the assumption that scientists can understand the brain by examining one region at a time. This is the way the vast majority of neuroimaging analyses are still conducted, with “mass-univariate” outcomes that treat each unit of brain (or “voxel”) as an outcome in a separate analysis. Maps are collections of effect estimates across voxels. Second is the assumption that if one identifies an area that responds to a particular stimulus or correlates with a particular behavior, it is sufficient to take activity in that region alone as a measure. For example, it has been typical to assume that one can identify regions whose activity represents pain by identifying one or more regions that respond to painful stimuli.

The same assumption plays out across different areas in affective science. Because the amygdala responds to negative affective stimuli, it is widely assumed that the amygdala represents negative affect. Amygdala activity has thus been taken as a measure of negative affect and used as a target for manipulations of the social context, mental health interventions, and more. The same is true for the nucleus accumbens/ventral striatum and reward; the dorsal cingulate and pain or “conflict”; the ventromedial prefrontal cortex and value, reward, or emotion more generally; and more.
One way to describe the problems with this assumption is to consider the formal inference being made. If we want to understand which brain areas encode the intensity of a painful stimulus, we need to infer on the probability (or effect size) of activity in a local region given an increase in stimulus intensity. This is called a forward inference, and it is what statistics in standard brain mapping studies are designed to address. The ability to make forward inferences depends on the measure’s sensitivity to the event in question. However, if I am interested in whether activating a brain measure implies that a particular type of mental event (e.g., pain) has occurred, this is a different type of inference. It is known as reverse inference in the brain mapping literature (Poldrack, 2006) because it involves inference on the stimuli or causal events rather than their effects on the brain. The ability to make reverse inferences depends on more than sensitivity; it depends on the positive predictive value of the measure, which depends jointly on sensitivity, specificity, and the base rate of the mental event in question. If a single brain region or network is active during many different behaviors or tasks, inferring mental states based on its activation becomes impossible, because one of many different constructs could be driving changes in activity.

This is precisely the case with brain regions typically used as measures of pain and other affective processes. For example, Fig. 8.1 shows a breakdown of the various types of tasks that activate one of the most “pain-selective” regions of the dorsal cingulate cortex, part of the anterior midcingulate (aMCC). Though the aMCC does contain single neurons that encode the incidence and intensity of noxious events (Hutchison, Davis, Lozano, Tasker, & Dostrovsky, 1999; Koyama, Tanaka, & Mikami, 1998; Sikes & Vogt, 1992), fMRI activity is observed in a wide variety of non-pain-related tasks as well, including cognitive, motor, and language functions. Thus, little can be inferred about the mental processes engaged based on finding activity increases at this location—and if a pain-related intervention influences activity in aMCC, little can be said about which of the processes potentially involved (pain perception, emotion, decision making, etc.) are being represented.

But again, all is not lost. Whether the aMCC is activated is only a small fraction of the information available in the brain image. In addition, we have information about the precise locations within aMCC, the magnitude of activation in each location, and the relative activation across locations. This pattern information can be substantially more sensitive and specific to pain and other mental categories, as we shall see below.

In sum, identifying brain measures for mental constructs is a worthy goal. These measures can tell us a great deal about the physiological architecture that supports the mind, and they can form a useful set of physiological targets for interventions. But problems arise with over-simplified definitions of brain measures and hasty, superficial validation. Showing that the dorsal cingulate responds to painful events, or the amygdala responds to negative images, is only the first in a long series of steps outlined above for understanding what constructs the brain measures represent. The strategy is not wrong, but the development and validation of the brain measures we use as proxies for affective representations is incomplete. In addition, the measures that come from standard hypothesis tests (is region x active in task y) are too coarse to have high positive predictive value for mental constructs.

Hyper-Complexity

The combination of sophisticated machine learning approaches and non-invasive whole-brain imaging has produced brain measures that are increasing in complexity. Rather than basing predictions on the activity of a single brain region, it is possible to develop more complex measures for mental constructs. Currently, many brain measures include on the order of a hundred of thousand parameters to make predictions. And a new class of models based on brain connectivity expands the space of parameters even more dramatically. For example, many studies now develop predictions based on connectivity across pairs of voxels (Dosenbach et al., 2010; Drysdale et al., 2016; Rosenberg et al., 2016; Shen et al., 2017; Turk-Browne, 2013). A standard fMRI scan at spatial resolutions now easily accessible has 62 billion pairs of connections, leading to models with up to 62 billion parameters. And although this may lead to better predictive performance, it comes with its own problems including interpretability and transparency.
The more complex a brain measure, the more difficult it is to interpret and explain how it makes predictions. A simple brain measure based on the activity of single region is easy to interpret: if the brain region is active above a set threshold, then we can make a probabilistic claim about the likelihood of a mental state. There is a single parameter to interpret and model predictions can be explained in a straightforward manner. This model is transparent (Lipton, 2016) because it is easy to contemplate in its entirety, it utilizes a simple (linear) learning algorithm, and each component can be intuitively explained. On the other hand, a complex brain measure that is based on the joint activation of every region of the brain is not so straightforward to understand. Because model parameters are jointly estimated, no single brain area is guaranteed to predict a mental phenomenon on its own. Dependencies across brain regions, estimated by their functional covariation, are also learned by the model in many cases. Often, modeling covariation across regions can help to more accurately predict outcomes of interest (Woo, Schmidt, et al., 2017). This means that the role of individual regions in these complex models can be difficult to infer (Haufe et al., 2014), and inferences are most accurate when the model is examined in its entirety.

The tendency for complex models to outperform simpler ones in many cases has given rise to the popular notion of a tradeoff between prediction and explanation. If prediction is the primary goal, we may not worry much about a model’s complexity. Conversely, if explanation is primary, we may care less about predictive accuracy. However, characterizing prediction and explanation as a simple tradeoff is a very limited way of thinking. Typically, we want both. A model that “explains” a phenomenon without being able to predict new instances of that phenomenon is a myth. For example, “the volcano erupted because the volcano god is angry” is one of many explanations, but if it has no predictive validity, there is no substantive reason to prefer it over any other explanation. Conversely, a model that predicts instances without explaining—the proverbial “black box”—is problematic because without knowing why a model makes one prediction vs. another, one never knows when the model will fail. It might make accurate predictions in one context, but be wildly inaccurate in others. For example, a brain model that predicts ADHD status might be based on head movement (Couvy-Duchesne et al., 2016; Eloyan et al., 2012), which obviously has little explanatory power for the brain basis of ADHD and might have predictive validity only until better motion-correction algorithms are available. Prediction and explanation might be better thought of as the two parents of understanding—sometimes at odds, but more often working together toward a common goal.

Beyond complicating our understanding of how they work, complex models can often be less generalizable. Models that are overly complex tend to characterize idiosyncrasies of the data used to build models to make predictions, rather than core features related to a mental construct. For example, a hypothetical brain measure of a broad construct like negative emotion may rely just not on responses related to unpleasantness or negative motivation per se, it could also be driven by responses that are related to affective arousal or the attentional demands of negative stimuli. This does not suggest that such a model is inherently flawed, simply that it uses many correlated features to make predictions about negative emotion. However, such a model is not likely to be broadly generalizable; it may not fare well in less frequent cases of negative emotion that are low in arousal, do not demand attention, or are superficially dissimilar for other reasons.

Thus, models should ideally be as simple and transparent as possible, but still retain good measurement properties. Determining the balance between simplicity and complexity is a well-known problem in machine learning, known as the bias-variance tradeoff: a model can either be more complex to precisely characterize the training data used to build it (a case of low bias or overfitting) or it can be less complex to minimize the variance in model predictions (a case of low variance or underfitting). Increasing complexity decreases bias but increases variance. The goal in model development is to find the proper balance of bias and variance to optimize model complexity. Many helpful approaches from machine learning have been developed to meet this goal (e.g., cross-validation and regularization, among others), and ideally produce more transparent, interpretable models. However, using tools from machine learning alone does not guarantee psychologically interpretable models. Insights from psychology and psychometrics are also crucial when it comes to developing brain predictors of mental constructs.

From Maps to Models of Brain Function

The Difference Between Maps and Models

The aim of brain mapping is to identify which brain regions are consistently activated by different experimental manipulations of mental state. The classical outcome of a brain mapping study is a parametric map that shows how different experimental conditions are associated with fMRI activity spanning the entire brain: perceiving faces (relative to other objects) produces a map with peaks in the lateral occipital cortex, fusiform gyrus, and amygdala; receiving rewards reveals a map with high activation the ventromedial prefrontal cortex, ventral striatum, and amygdala. Brain maps identify the neural correlates of different manipulations, substantiating forward inferences about the brain regions that will be active during a particular mental state.

Superficially, a brain model may not appear very different from a brain map. This is because they both comprise a set parameter estimates from a regression model (or related statistics) across local areas of the brain. Both reveal patterns of brain activity that are related to some mental phenomenon. However, the purpose of brain maps and models is quite different. Whereas brain maps characterize which brain regions respond to different stimuli or mental events, multivariate brain models are designed to make reverse inferences about (or predict) an individual person’s mental state or behavior based on their brain activity (Kragel, Koban, et al., 2018; Woo, Chang, Lindquist, & Wager, 2017).

The utility of brain models lies in their ability to quantify reverse inferences by making objective, testable predictions about mental states based on brain activity. This allows researchers to focus on models that have desirable measurement properties, falsify models by making strong predictions, establish the reproducibility of models, and identify the mental constructs with which a model is most consistent. Brain maps contribute useful knowledge about the function of brain regions, particularly when accumulated across studies and summarized in mega- and meta-analyses (Wager, Lindquist, & Kaplan, 2007; Yarkoni et al., 2011). But brain models are more useful for understanding how the brain constructs mental states and organizes the space of mental phenomena, through their capacity to predict outcomes of interest and characterize brain representations of mental constructs.

Predictive brain models come in many forms (for review, see Kragel, Koban, et al., 2018), but perhaps the most common form is a map of linear associations between voxels in multiple brain regions and an outcome of interest. An outcome of interest can be estimated by computing the dot product of this map, often called a weight map or predictive map (Fig. 8.2), with brain activity measured during scanning. This estimation procedure involves computing the product of the predictive map and measured brain activity at every voxel, and then summing across all voxels. This way, brain activity measured at every voxel in the map contributes to the prediction. Depending on the form of the model, outcomes of interest could be a continuous measure of behavior, a subjective measure of self-report, a probability of a mental state, or a diagnostic outcome such as clinical status.

Fig. 8.2

Brain mapping versus brain modeling, and valid tests of model accuracy. The standard brain mapping procedure (top left) identifies brain areas that respond to known (or assumed) psychological events. It provides a map of local effects, but not a model of how these various brain effects work separately or together to construct a psychological state. The predictive modeling approach (top right) reverses the equation, treating psychological states as outcomes. These outcomes are jointly predicted by combinations of brain regions (including connectivity and integrative network properties if desired). The bottom panels show some desirable schemes for testing the predictive accuracy of a model on new individual participants. One of the most efficient ways is cross-validation (bottom left), in which a subset of participants is held out as a test set. The model is trained (i.e., its parameters estimated) on the remaining participants (the training set), and the trained model’s performance is evaluated on the test set. The procedure is repeated with a different test set until every participant has been tested at least once. If done properly, this provides a virtually unbiased estimate of how well a model trained on one group of individuals will generalize to others. However, there are many ways in which the assumptions underlying cross-validation can be violated, resulting in over-optimistic estimates. For
this reason, it is desirable to also test the model prospectively on an additional hold-out sample that is tested only once. Repeated testing with different models will produce an over-optimistic bias and invalidate the test. Models that hold up to validation in this way can be validated on other samples as well.

To build models that achieve the goals described above, models need to be trained with generalizability and falsifiability in mind. Increasing the amount of training data should improve performance and decrease the likelihood of overfitting. Verifying that the model performs well on data that is independent from that used to train it is essential to be certain it does not only predict idiosyncrasies of the training data (cross-validation is one way to estimate the generalizability of a model, see Fig. 8.2). Training on data from different manipulations that are all conceptually related to the outcome of interest but that differ from one another in superficial ways (e.g., modeling brain responses to aversive images and sounds) is one approach to increase generalization across contexts. Training models on data from a sample (or multiple samples) of independent subjects has multiple advantages. It greatly increases the amount of data that can be used to train models, as data can be pooled across large and diverse samples. It also allows models to be tested prospectively in new studies, enabling new hypotheses about the specificity and generalizability to be tested.

Three Examples: Models of Pain, Affect, and Emotion

To date, we have developed over 18 predictive models focused on basic affective and emotional processes, all of which were designed to generalize to new individuals (for a partial review, see Kragel, Koban, et al., 2018; Woo, Chang, et al., 2017). These have been tested across independent samples to varying degrees, fostering the process of prospective testing and validation. Although these models are at different stages of development, and
some affective processes are easier to manipulate and measure with fMRI, here we focus on several examples that have been successfully evaluated in multiple prospective tests, leading to a better understanding of which cognitive and affective processes they respond to, and which they do not. These models aim to characterize the neural substrates that best describe affective processing related to physical pain, negative affect, and discrete emotional experiences. They show that it is possible to develop brain models that robustly and reproducibly predict individual people’s affective experiences.

The Neurologic Pain Signature: A Neural Marker at the Core of Nociception

Perhaps the most extensively validated brain model of a basic affective process is the Neurologic Pain Signature (NPS, Wager et al., 2013). The NPS was designed to predict differences in subjective pain reports based on fMRI activity in areas commonly associated with painful stimulation (Yarkoni et al., 2011), including dorsal cingulate cortex, insula, somatosensory cortex, thalamus, midbrain, and a number of other regions (Fig. 8.3). The model was developed using brain responses to thermal stimulation and was found to discriminate between painful heat and nonpainful warmth, the anticipation of pain, and pain recall with over 94% sensitivity and specificity. In a prospective test, the NPS was found to be sensitive to the subjective intensity of thermal stimulation in an independent study and specific to physical pain, as the NPS responded strongly to painful thermal stimulation, but showed little response to nonpainful stimulation and no response to “social pain” evoked by viewing an image of an ex-romantic partner and recalling an experience that evoked rejection-related distress (Wager et al., 2013). This was surprising in light of other work highlighting the similarity of somatic pain and rejection (Eisenberger, 2015; Keysers, Kaas, & Gazzola, 2010; Kross, Berman, Mischel, Smith, & Wager, 2011). Because rejection has been considered one of the experiences most similar to somatic pain in its brain representation, finding no NPS response to rejection was a particularly important demonstration of specificity. In addition, many researchers have focused on the similarities between pain-related activation and activation related to other salient events (Legrain, Iannetti, Plaghki, & Mouraux, 2011). But rejection-related stimuli and other stimuli that fail to activate the NPS—e.g., observing others’ pain (Krishnan et al., 2016) and highly aversive emotional pictures (Chang, Gianaros, Manuck, Krishnan, & Wager, 2015) are highly salient, suggesting that the NPS is not tracking general salience, attentional demand, or arousal.

Fig. 8.3

The Neurologic Pain Signature (NPS). The left panel shows the local activity patterns that together comprise the NPS, which was trained on voxels covering about 10% of the brain (Wager et al., 2013). Many regions involved in nociception, like the dorsal cingulate and posterior insula, are included in the NPS. The important thing, however, is that the model comprises specific local patterns in these regions that are difficult to name, often do not respect anatomical boundaries precisely, and specify the relative levels of activity in neighboring voxels. This is very different from a “model” that stipulates that one should find some activity somewhere in the cingulate: It is much more precise. The right panel shows the performance of the NPS in tracking trial-to-trial variations in subjective pain reports, in nearly 400 participants across 11 studies collected in our laboratory. Each data point shows the correlation between pain and brain responses across single trials. Though this is expected to be quite noisy, as single-trial responses are highly variable, the vast majority of participants (over 95%) show a positive correlation between NPS responses and pain reports. The “violins” in the plot show the distribution in these correlation values (y-axis) across individual participants for each study. This shows that the NPS’s correlation with pain is highly reproducible across individuals and samples, without any model re-fitting or any special techniques. The NPS was also reliable, with an odd-even trial reliability of 0.84 on average across the studies. Though accuracy could increase if acquisition and preprocessing procedures were standardized across studies, good performance across studies shows robustness to some variations in paradigm, acquisition (including field strength), and analysis.
Since its development, the NPS has been validated cross-culturally using brain responses to pain in over 34 independent cohorts at the time of writing (e.g., see Zunhammer, Bingel, Wager, & Placebo Imaging Consortium, 2018). Figure 8.3 depicts a recent analysis examining the relationship between trial-by-trial variation in the NPS response and self-reports of pain. Although individual studies vary in terms of stimulation parameters, imaging protocols, sample demographics, and concurrent cognitive and affective demands, the NPS is reliably associated with pain reports in each study, with notably large effect sizes.

The generalizability and specificity of the NPS has been evaluated in an ongoing series of studies (summarized in Kragel, Koban, et al., 2018; Woo, Chang, et al., 2017). The specificity of the NPS has been tested against brain activation during a range of non-painful events, such as viewing aversive images, observing others in pain, performing challenging cognitive tasks, and viewing images of ex-romantic partners. In addition, a number of interventions that influence reported pain have no apparent effects on NPS responses. These include cognitive self-regulation (Woo, Roy, Buhle, & Wager, 2015), most placebo effects (Zunhammer et al., 2018), reward preceding pain (Becker, Gandhi, Pomares, Wager, & Schweinhardt, 2017), and manipulations of expectations (Geuter, Boll, Eippert, & Büchel, 2017; Woo, Schmidt, et al., 2017) and perceived control (Bräscher, Becker, Hoeplli, & Schweinhardt, 2016; Woo, Schmidt, et al., 2017). These null findings suggest that the NPS is not influenced by “top down” effects in most cases, with the exception of some forms of conditioned (learned) influences on pain (Jepma, Koban, van Doorn, Jones, & Wager, 2018) and some manipulations of social context (López-Solà, Koban, & Wager, 2018; Sola, Koban, Geuter, Coan, & Wager, 2019). This suggests it mediates core nociceptive and affective aspects of pain, rather than evaluative aspects that are known to influence pain reports.

Motivated in part by debates about the degree to which the NPS is a marker of pain specifically or is also responsive to stimulus salience (Hu & Iannetti, 2016), current efforts focus on evaluating the sensitivity of the
NPS to a broader array of painful events, including visceral and mechanical stimulation and its specificity against other potentially iso-salient, aversive stimuli such as unpleasant and even “painful” sounds, and breathlessness, among others.

These findings also illustrate two additional principles related to model validation across studies. First, the NPS tracks pain in some, but not all, contexts. For example, it responds to pain increases caused by turning up the heat, but not by **imagining** more intense heat (Woo et al., 2015) or expecting more intense pain (Zunhammer et al., 2018). Does this falsify the NPS as a pain-related measure? No, we do not think so! No brain measure can ever measure “pain,” or any other subjective experience. Brain measures measure **brain systems**, which are linked to pain and may play a role in creating it. But these brain systems can be ignored; my “pain systems” may be firing like crazy but I may be ignoring them, unconscious, or just stoic and unwilling to report my experience as “pain.” Since the publication of the NPS, it has become increasingly clear that it reflects one system (with subsystems) that contributes to pain, but other brain systems are important for capturing other aspects of pain, including the change in negative evaluation that occurs when one imagines that a stimulus is damaging or harmless (Woo et al., 2015).

Second, testing a validated pain-related measure can provide a new window into which interventions are effective in shaping the construction of pain. Most psychological and behavioral interventions (e.g., placebo and cognitive regulation) do not affect the NPS, implying that they affect a **later** stage in pain construction or evaluation, or at least different brain processes that contribute to pain reports. Some interventions influence the NPS (e.g., generating expectations of higher levels of pain, see Jepma et al., 2018), which—because the NPS is very sensitive to painful peripheral input—implies a deeper level of influence on earlier aspects of pain sensation and perception. The magnitude of an intervention’s effect on the NPS may turn out to be relatively unrelated to the effects on pain reports, leading to interesting new questions: Is minimizing NPS responses helpful in terms of long-term pain, avoidance, and physiological harms? Or will minimizing self-reported pain always be the sine qua non of pain treatment?

The Picture Induced Negative Emotion Signature: Multiple Brain Systems Engaged in Processing Unpleasant Images

One of the most prominent organizing features of affective states is their valence: whether they are pleasant and positively reinforce behavior or whether they are aversive and act as negative reinforcers. The Picture Induced Negative Emotion Signature (PINES) was developed to characterize the brain systems involved in predicting negative affect generated during picture viewing. The PINES was trained in a sample of 121 individuals, using whole-brain patterns of fMRI responses to scenes and objects to predict subjective ratings of negative affective experience. Cross-validation in this training sample revealed exceptional prediction of negative affect in independent subjects: the root mean squared error was only 1.23 points on a 5-point rating scale, and correlations between observed and predicted ratings were high \((r = 0.85, \text{Cohen's } d = 3.23)\). Holdout testing in 61 participants not used for training the model showed similar performance in a completely independent sample, discriminating between negative and neutral images (Fig. 8.4).

**Fig. 8.4**

The picture-induced negative affect signature (PINES). This brain model was trained to predict the intensity of reported negative affect on a 1–5 scale, across approximately 180 participants (Chang et al., 2015). The pattern is shown in the left panel; it includes local patterns in the amygdala, insula, dorsomedial prefrontal cortex, hypothalamus, periaqueductal gray, and other regions related to affect and social cognition. The middle panel shows that it tracks ratings across all levels of negative affect in an approximately linear fashion, with nearly identical performance for the cross-validation set and prospective holdout set, which was tested only once. This similarity indicates a lack of bias in the cross-validation test. The right panel shows the performance on the holdout set (tested once) on responses to negative versus neutral images. Each pair of dots connected by a line represents data from one participant. The model correctly predicted which image was the
negative one in 100% of the participants, with a massive effect size of $d = 3.3$. This is an unbiased estimate because the model was tested only once on these data, without re-fitting parameters. Measures with large effects like this one offer substantially more power than single-region effects as targets for interventions (e.g., Gilead et al., 2016; Koban, Kross, Woo, Ruzic, & Wager, 2017; Reddan, Wager, & Schiller, 2018).

Because high levels of negative affect are often associated with salience or arousal, a follow-up test was conducted to examine the specificity of the PINES to another unpleasant experience that is highly arousing: painful thermal stimulation. This test revealed that while the PINES is sensitive to the intensity of negative affect evoked by images, it is not sensitive to thermal stimulation, ruling out the possibility that a simple common factor such as arousal adequately describes the mental processes characterized by the PINES. Since its development, efforts are underway to determine if the PINES is better thought of as a general marker of negative emotion or is fine-tuned to a particular set of appraisals, such as evaluations of threat or prospecton about negative outcomes. In particular, work has focused on whether the PINES is sensitive to affect evoked by aversive sounds and to positive images to see whether it captures appraisals common to highly salient stimuli.

Relatedly, the PINES has been validated in a prospective test examining whether perspective taking can modulate affective responding to negative images (Gilead et al., 2016). In this study, participants were presented negative and neutral images, and were instructed to either take the perspective of a tough individual who feels little emotion or a more emotionally sensitive and squeamish person who is more prone to responding emotionally. The PINES robustly generalized to this independent sample; brain responses to negative images evoked greater PINES responses compared to neutral images (Cohen’s $d = 2.3$). Moreover, PINES responses were diminished by perspective taking. Responses were lower when participants took the perspective of a tough individual compared to the perspective of someone with high levels of emotional sensitivity, although with a
smaller effect size (Cohen’s $d = 0.37$). This result demonstrates that, unlike the NPS, the PINES is sensitive to cognitive self-regulation—making it a potentially useful target for clinical interventions.

**Brain-Based Markers for Emotion Categories: Distributed Representations Identify Qualitatively Distinct Kinds of Emotional Experience**

In addition to models that characterize continuous affective dimensions, such as the intensity of pain and negative affect, predictive models have also been developed to identify brain states that distinguish emotional experiences that are rated as being categorically distinct (Kragel & LaBar, 2015). These brain-based models of discrete emotions (Fig. 8.5) were identified by modeling whole-brain patterns of fMRI response to cinematic films and instrumental music that participants rated as evoking distinct feelings of either contentment, amusement, surprise, fear, anger, sadness, or the absence of emotion which was rated as neutral. The decision to include both films and music in the training dataset for these models illustrates a powerful principle: To train models to be maximally generalizable, it is a good idea to include examples of the constructs (here, emotion categories) that are as distinct as possible from one another on superficial features, such as the sensory modality used to elicit emotion. This reduces the chances that the model picks up on confounding characteristics (e.g., different visual properties of “angry” vs. “happy” movies) and increases the chances that its predictions will generalize to new stimulus sets and tasks.

**Fig. 8.5**

Brain-based markers for multiple emotion categories. These distributed brain models were trained to predict the emotion category labels assigned to both emotional videos and music clips. As with other models, the peak regions are shown, but the models included a broader pattern across the brain. The models were able to classify single-trial responses into one of seven emotion categories with an average accuracy of just over 37%—nearly triple the chance level of performance. As shown in the lower left panel, pairwise classification of single trials revealed effects robustly above chance levels, with an average area under the ROC curve = 0.652. In terms of self-reported experience, the frequency of model classifications explained 57% of the variance in self-report across the seven emotion categories (based on a binomial regression model). Thus, even though the models were trained using a categorical framework, and did not include information about self-report, they are sensitive to differences in emotional experience across a priori categories.
Cross-validation across independent subsamples of subjects revealed that brain responses to single movie and music clips could be classified into one of these seven categories of emotional experience with a medium effect size (Cohen’s $d = 0.55$) and that the models could predict 57% of the variance in self-reported emotional experience.

Given the initial success classifying brain responses to films and music in independent subjects, the generalizability of these models was prospectively tested in the absence of stimulation to see if the brain states identified aspects of emotional experience that were stimulus-related, or if they captured more general aspects of emotion that are shared across internally and externally generated feelings. This test was conducted by classifying brain activity during resting-state scanning (in a sample of 499 participants in the Duke Neurogenetics Study — Kragel, Knodt, Hariri, & LaBar, 2016) and evaluating the relationship between individual differences in emotional states (anxiety and depression) and traits (anxiety, angry hostility, and depression). Associations were found between participants’ state anxiety and model-predicted fear, depressive symptoms and model-predicted sadness, trait anxiety and model-predicted fear, trait angry hostility and model-predicted anger,
and trait depression and model-predicted sadness. Although effect sizes were modest (on the order of Cohen’s $d = 0.1$), which is common when examining individual differences, these findings validated the emotion markers by showing selective correlations with conceptually related state and trait measures of emotion.

**Learning About the Brain, Learning About the Mind**

Key questions in affective science have traditionally focused on either the mind or the brain. Ongoing debates in psychology concern whether affective states and constructs should be characterized as points in a multi-dimensional space, represented as different kinds or categories, or as some combination of the two. These types of debates span multiple areas of affective science, including pain (Davis, Kucyi, & Moayedi, 2015) and emotion (e.g., Barrett, Khan, Dy, & Brooks, 2018; Cowen & Keltner, 2017, 2018). Until recently, evidence from neuroimaging has played a relatively minor role in constraining theories of pain and emotion. Analogously, debates in affective neuroscience tend to focus on mapping different psychological processes onto different neural substrates: does the amygdala selectively process information related to valence, threat, fear, or salience? Is the dorsal cingulate a pain selective region (Lieberman, Burns, Torre, & Eisenberger, 2016; Lieberman & Eisenberger, 2015), or does it integrate multiple different computations involved in valuation and action (Apps, Rushworth, & Chang, 2016; Brown & Alexander, 2017; Kolling et al., 2016; Kvitsiani et al., 2013; Shenav, Straccia, Cohen, & Botvinick, 2014; Wager et al., 2016)? For the most part, hypotheses about the brain and mind have been separate, making it difficult to use our knowledge of the brain to advance our understanding of the mind, and vice versa. The predictive modeling framework aims to overcome this issue by making links between psychological theory and brain models explicit, with the goal of simultaneously uncovering knowledge about both the brain and mind.

**Learning About the Brain: Using Models to Understand Brain Representation**

Predictive brain models can be used to answer many different questions about brain representation. One line of questions explores the relationship between brain structure and mental events: Which neural structures are important for a mental construct? Are certain networks or groups of brain regions more important than others? Is the brain representation of a construct distributed, or is it engendered by local codes? These questions reveal insight into the nature of brain representations, providing a rich way of comparing predictive models of mental phenomena.

These questions can be answered by crafting models using different approaches and comparing the results. To understand which brain regions are important for a mental construct, multiple models can be built using brain activity from different sets of areas. If a single region, or as is more often the case, if a set of brain regions is sufficient for predicting an outcome of interest (e.g., the intensity of negative affective experience), then increasing the complexity of the model by including signals from additional brain regions should not improve the accuracy or performance of the model. Conversely, if a brain region is necessary for predicting an outcome of interest, then any predictive model that does not include it should perform worse than if it had been included in the model.

As an example, consider two predictive models that are both good predictors of an outcome of interest that were trained using two non-overlapping brain regions. Model performance is the same regardless of which brain region is used to build a predictive model. Thus, either brain region is sufficient for prediction, but neither brain region is necessary. In this scenario, the outcome of interest may be coded similarly in each of these brain regions, because no information is gained by adding signals from both regions to a common model. In this case, the regions could be considered redundant from an information theoretic perspective.

Related to the problem of identifying which brain regions are necessary and sufficient for prediction, brain representations can be characterized either as local or distributed codes. Local representations are spatially restricted to a single brain region (or circumscribed neural circuit). Distributed representations are spatially extended, and contain multiple codes that on their own do not directly reflect the outcome of interest but only do so when considered together. The distinction between local and distributed representations could apply to coding
in single neurons vs. populations of neurons in a brain region (Averbeck et al., 2006), or to single brain regions vs. large-scale distributed networks or combinations of networks (Kragel, Koban, et al., 2018).

The representation of objects in inferotemporal cortex is one particularly well-studied example of distributed representation. This brain region contains neurons which code for different high-level visual features, such as color and form (Tanaka, 1996). Individually, these neurons cannot effectively represent an object. However, when considered jointly, populations of these feature-selective neurons can be used to code for many different types of objects. Consider, for instance, populations of neurons that selectively respond to objects that are orange in color, or objects that have curved edges, or are somewhat glossy, or that have a dimpled texture, and so on. Any single one of these features is not sufficient to represent the fruit “orange,” but when enough of these features are combined, they can form a distributed representation for “oranges.” In addition, a number of studies have shown that the individual neurons that respond most strongly to a given object type (e.g., oranges) are not sufficient to decode object categories—i.e., to discriminate oranges from others (Kiani, Esteky, Mirpour, & Tanaka, 2007). Distributed population codes also appear to be crucial in other areas as well, from motor control to emotion (reviewed briefly in Kragel, Koban, et al., 2018).

With fMRI, this logic can be extended to the analysis of distributed codes that span the entire brain. A goal in model development is to identify the full set of brain regions that are internally consistent (i.e., that reliably code for a single feature) and which improve performance when added to a predictive model (i.e., is not redundant with other brain features). In this case, each individual feature is necessary for the distributed representation, but no single feature is a sufficient prediction. Characterizing the different aspects of distributed representations can help characterize the nature of complex mental constructs. As an example, there are many different features related to negative affect: valuation of poor outcomes, unpleasant feelings, high levels of arousal, increased attention, motor activation, and so forth. This kind of distributed representation of negative affect would not likely be coded in a single brain region, but would be processed in parallel by multiple systems specialized for different processes. Predictive modeling of negative affect using fMRI provides evidence for such a representation: the PINES, which predicts the intensity of negative affective experience, is composed of multiple subnetworks (including visual, somatosensory, limbic, subcortical, among other brain regions). Although each of these subnetworks independently contributes to predictions of negative affect, no single region is necessary or sufficient for prediction—providing evidence that brain representations of negative affect are distributed in nature.

Our model predicting negative emotion from brain activity patterns (Chang et al., 2015) exhibited characteristics of a broadly distributed process: No single resting-state network was either necessary or sufficient to predict the intensity of reported negative affect. In addition, a model that combined voxels across multiple large-scale networks was vastly superior to models restricted to any single region (Fig. 8.6). Likewise, models of somatic and vicarious pain constructed from territories spanning multiple brain networks outperform those constrained to single brain regions.

Fig. 8.6
Testing the necessary and sufficient basis for prediction. Comparing predictive models and testing their relative accuracy can inform us about the features of a model that are necessary and sufficient for prediction. One important aspect is the question of how much brain “real-estate” is needed to accurately predict an outcome? Perhaps a single region, like the amygdala, is enough. Perhaps the critical voxels are all contained within one coherent network, like the “default mode” network. Or maybe the situation is more complex and multiple networks are required. Though distributed models appear to produce more accurate results with larger effect sizes in many studies—with outcomes ranging from memory to sustained attention to pain and emotion—the benefits of distributed models are rarely tested systematically. Here, we show two examples of such comparisons, for negative emotion (top), somatic pain (middle), and vicarious pain (bottom). In each case, a model including the whole brain substantially outperformed even the best single regions identified in searchlight analyses across the brain (e.g., the amygdala for emotion, or posterior insula for pain).
analyses show that for both outcomes, negative affect is truly encoded in a distributed network, and no single region is adequate.

Validating predictive models can additionally be used to show which contexts and variables a brain representation generalizes to or is specific against. For example, the amygdala is an important structure for acquiring conditioned skin conductance responses to tones paired with aversive outcomes. However, is the same amygdala representation also critical for fear-potentiated startle? Is the same representation involved in learning the negative value of certain tones also utilized in learning which tastes should be avoided? Often, we do not know which features are important for affective behavior; we make assumptions, but the boundaries of generalization are usually untested.

Prospectively testing predictive brain models moves beyond assuming brain representations are shared across these factors by making tests explicit. Showing that the NPS responds robustly with the intensity of thermal and
mechanical stimulation, but not other emotionally salient events like “feeling” another’s pain or viewing aversive images makes it clear that the NPS is not *just* a model of exteroceptive salience, but that it is uniquely predictive of intense sensory events that lead to physical pain. Just as these tests can tighten the boundaries of a predictive model, by showing specificity, so too can they broaden the limits of presumed generalization. For instance, predictive brain models for distinct kinds of emotional experience (e.g., fear and sadness) respond not only to rich stimuli such as narrative film, but also to individual differences in self-generated feelings in the absence of stimulation.

These examples show how systematically evaluating predictive models—whether testing in an independent subject or a different population, and testing the response of a model to related psychological manipulations—provides insight regarding whether a mental construct has a reliable brain basis, and what the nature of brain representation might look like.

**Learning About the Mind: Distinct Systems for Different Types of Affect**

Comparing brain representations to one another sheds light on which constructs are more similar to one another, and may be conceptually linked. Often, we make psychological distinctions based on behavior, language, subjective experience, or more generally based on long-held assumptions about how the mind works. Comparing and contrasting models based on the brain can shed new insight into the structure of the mind.

As an example, consider relationships among emotional experiences. Most of the time, correlations among self-reports and judgments of conceptual similarity show that anger and sadness are more similar to each other than to happiness. This is assumed to be the case because anger and sadness are both associated with negative affect whereas happiness is a positive emotion. But our assumptions are often not holding up when validated against human brain activity. Both meta-analyses (Murphy, Nimmo-Smith, & Lawrence, 2003; Wager et al., 2015) and individual studies (Kragel & LaBar, 2015) have shown that sadness and happiness are, relatively speaking, more similar to one another, and that anger is farther away in brain space. This challenges the notion that the emotions are organized primarily based on valence, and that other dimensions of appraisal and affective experience, such as self-relevance and internal orientation (Wager et al., 2015), may be equally if not more important in organizing emotions. Thus, understanding the brain can be used to update current theories about how the mind works by identifying commonalities and differences between mental constructs.

When we begin to compare models that predict various kinds of affective outcomes, a very interesting pattern emerges. The models are largely distinct, suggesting that different affective outcomes are related to different patterns across brain systems. The similarity in the spatial patterns for 18 predictive models developed in our lab is shown in Fig. 8.7. The matrix of intercorrelations shows that the maximal correlation among any pair of models is around $r = 0.2$, suggesting that each model is distinct. There are caveats; comparing cross-prediction of outcomes is a stronger criterion for assessing similarity across models (see Fig. 8.7 legend for discussion), and the low correlations could be due in part to noise. Nonetheless, when paired with the dissociations in outcome prediction we have observed, the results suggest that there is much more differentiation among affective brain processes than we, at least, had previously imagined.

**Fig. 8.7**

Similarity across brain models of affective processes. The matrix image shows the correlations in the spatial patterns (weights predicting affective outcomes) across 18 models, each designed to predict a specific affective outcome across participants. These models are remarkably dissimilar from one another: The maximum correlation between any pair of models is around $r = 0.2$, suggesting that each model is distinct. We must be cautious here, as two brain models can be spatially dissimilar but make the exact same predictions (it’s true!). Thus, spatial similarity is only part of the story, and cross-prediction and tests of separate modifiability provide a stronger way to evaluate whether two models predict different things. Nonetheless, the picture that emerges from both comparing spatial similarity across models
and patterns of separate modifiability within individual studies is that different affective outcomes are predicted by different brain patterns. The dendrogram shows the group of the models; those closest together are most similar. The model groups models for similar outcomes together: neighbors include models of pain (NPS and SIIPS), empathy (care and distress), autonomic responses to stress (heart rate [HR] and galvanic skin response [GSR]), similar emotions (fear and surprise), and fibromyalgia (pain and multisensory responses). Some are less similar than expected: the Vicarious Pain Signature (VPS), trained on pictures and validated on non-visual stimuli, is different from a model of empathic distress trained on auditory narratives and linked with charitable donation.

Empathy: A Case Study
Recognizing others’ distress can involve at least two kinds of processes. One is fast and relatively reflexive and automatic, involving little conscious thought. It tends to produce “experience sharing” or “state matching” (e.g., I am distressed at your distress), which is believed to underlie emotional contagion and some specific forms of helping (and aggressive or fearful) behavior in animals (Preston & de Waal, 2002). At the brain level, representing others’ actions, and perhaps emotional states, is associated with “mirror neurons” in the premotor and inferior frontal cortex. Additionally, perceiving others’ distress is associated with activation of the anterior insula and cingulate cortex (Lamm, Decety, & Singer, 2011). The second process is slower, more deliberate, and is thought to involve higher-level cognition and, in particular, mentalizing about others’ distress. Mentalizing requires the ability to recognize that another’s mind is distinct from one’s own and conceive a theory about their mental state based on (potentially) multiple context clues. At a brain level, it is thought to rely on cortical networks associated with social cognition, including dorsomedial prefrontal cortex and superior temporal sulcus. It remains largely an open question how important each of these facets of empathy are when it comes to generating feelings for others and helping behavior.

For our purposes here, we’ll focus in on one aspect of this broad picture: How similar are brain responses to one’s own pain vs. vicarious “pain” from observing others? The idea that they activate the same brain systems, particularly the anterior insula and cingulate (Singer et al., 2004), has been used to argue that state-matching mechanisms are important for empathic care and helping responses in humans. The implications of the answers to this question go beyond understanding how empathy works; as David Brooks wrote in his article “The Archipelago of Pain” (Brooks, 2014), if emotional pain is the same as somatic pain at a brain level, why treat them differently in our legal system and policymaking endeavors? Brooks argued that if social isolation is like being physically tortured, maybe it should be just as illegal. And if observing someone else in pain activates our pain circuits, does that qualify it as a type of harm as well?

Studies comparing self-pain with other-pain (or vicarious pain) reveal very different patterns of brain overlap depending on whether one is looking at overlap in univariate brain responses to stimuli or comparing patterns that decode experiences of distress. Most early studies have found overlap in the anterior insula and cingulate, but there are a couple of problems with concluding that this reflects a “shared mechanism.” First, these studies generally do not establish a strong link between activity and vicarious pain, and they do not establish that the responses are specific to “painful” events rather than other classes of emotional, cognitive, language, decision, and motor processes. Thus, if there is overlapping activation when experiencing heat pain and looking at pictures of others in pain, what does this overlap mean in terms of which processes are shared? Any set of mental processes common to self- and other-pain and reduced in the control tasks (generally non-painful warmth and neutral pictures) could be driving shared activation. This includes greater attention, greater salience or relevance, stronger autonomic responses, and greater demand on action planning mechanisms. To infer that what is shared is specifically related to empathy involves ruling out these and other alternatives.

A second problem is that these studies typically focused on the overlapping areas and assume all the differences are essentially due to noise. For example, if self-pain and other-pain overlap in 5% of the voxels tested, does this mean that a “shared mechanism” has been found? It might be prudent to consider the differences as well—perhaps it means that self- and other-pain are only 5% similar. But this conclusion would be premature as well, for at least three reasons. First, overlapping voxels are a poor way to assess similarity in mental processes because the number of activated voxels is not a measure of any particular process, as described above, and may not be related to empathy at all. Secondly, it ignores the magnitude of the responses. What if the voxels in common are activated twice as strongly by one condition than another? Should this still be counted as an
identical response for purposes of assessing overlap? Third, the similarity metric will be strongly influenced by measurement noise as well as the similarity in the underlying processes.

The first of these is the most conceptually profound, and points to some fundamental uncertainties in how we should use brain similarity to infer similarity in mental processes. Even if we quantify the degree of overlapping vs. non-overlapping voxels, we need to understand the relationship between activation patterns and the behavior we are interested in (e.g., vicarious pain). This is a conceptual problem, not a problem with measurement noise. An illustrative example comes from a recent study by Carrillo et al. (2019). They quantified the proportions of cells in rodent dorsal anterior cingulate cortex (dACC) that respond to (1) painful shocks, (2) observation of another rodent receiving painful shocks, and (3) a threatening sound conditioned to painful shock. Some neurons responded to each of the three motivationally relevant conditions, and subsets responded selectively to pairs of conditions or to all three. They interpreted those neurons that responded to self-pain and other-pain as “empathy selective.” This is appropriate—but does it mean that the dACC contains a “mechanism” for empathy? Maybe, but we have to make an additional assumption: We must assume that shared neural activation implies a shared process. However, this need not be the case. A wealth of population-coding evidence in neuroscience indicates that in many domains, the pattern across neural population carries information about stimulus types, mental categories, and motor responses—not the individual neurons. If we accept the “empathy mechanism” account of neuronal overlap, we must also posit that there is a “pain and emotion but not empathy” process also represented in the cingulate, and a “empathy and emotion but not pain.” A more parsimonious alternative is that the cingulate contains distributed representations related to each of the three types of events. These may share some neurons in common, but involve distinct or even completely independent neural patterns. In this case, a natural similarity metric would be the spatial similarity across the populations of neurons—do they share 5%, 50%, or 90% of their neurons in common? This can be captured by calculating the spatial correlation across neural patterns, perhaps considering the continuous intensity of the neural responses to each type as well as whether or not they pass a statistical threshold and are thus considered “responsive.”

We must also recognize that linear spatial similarity may be insufficient and other metrics apply—or even that the overlap in individual units has little bearing on the population-level representation. As an analogy, consider three words: GRASS, FLOWERS, and BAGS. Imagine that each word represents a mental process, or construct, and each letter a neuron that fires in response to that process. Some neurons (“S”) respond to all three, and some (“A,” “R,” “G”) to only two. If we summarize these overlaps, we will find that there are some “grass-flower” units (“R”), which we might capture a mental process common to the two conditions (R encodes “living plants”). If we calculate the similarity across units, we might infer that GRASS and FLOWER share slight overlap, but GRASS and BAGS are very similar. In fact, they are! They are orthographically similar, but they are not semantically similar. The relationships between letters and conceptual meaning are not linear, and one cannot construct a function of the similarity in letters and come up with an answer for the similarity in meaning. This example shows us that counting overlapping neural populations, or even assessing spatial neural similarity, may not always give us the right answers when it comes to inferring similarity in mental processes.

We are not particularly nihilistic about the situation, and there are solutions. Multivariate pattern analysis provides a complementary way of looking at neural populations—whether fMRI voxels or individual neurons—that partially solves the problems raised above. First, multivariate decoding provides a set of predictive models that can quantify how much variance in an outcome (e.g., reported vicarious pain experience) is captured by the model. If the predictive validity is high, then we can be more certain that we are studying brain measures related to empathy. In addition, if the predictive models can be tested across studies, some alternative processes can be ruled out—e.g., if other tasks that enhance attention do not increase responses in the model, enhanced attention can be ruled out as an explanation for what the brain model is measuring. Second, we can provide an unbiased estimate of brain similarity, in two ways. We can assess the spatial similarity across two multivariate models, as described above. Or we can assess cross-prediction: How much variance in one outcome (e.g., somatic pain experience) is predicted by a model of a comparator outcome (e.g., vicarious pain experience), and vice versa. Third, cross-prediction provides an unbiased estimate of similarity, controlling for effects of measurement noise. For example, if Brain Model 1 explains 25% of the variance in somatic pain ratings and 25% of the variance in vicarious pain ratings, we might infer that it reflects both somatic and vicarious pain. If it explains 25% of somatic pain ratings but only 5% of vicarious pain ratings, we might infer that the effects are five times stronger.
for somatic pain. And if there is more noise or the outcomes are unreliable, we can still assess the relative predictive power—e.g., 5% for somatic but only 1% for somatic pain. Finally, assessing cross-prediction avoids some of the problems with the mapping between neural units and conceptual categories discussed above, because it assesses the pattern as a whole, and whether that pattern is specific for one condition or general across both. We need not assume that the units of the model (voxels or neurons) are individually interpretable in relation to the mental category that the pattern reflects.

In a series of studies comparing pain and empathy, we tested whether self- and other-pain activate similar brain representations. We identified whether multivariate brain patterns that predict pain and vicarious pain are similar or different, and analyzed both spatial pattern similarity and cross-prediction to compare the models that predicted each. The design of the first study (Krishnan et al., 2016) involved presenting a randomized series of trials of self- and other-pain. In one fMRI session, participants experienced three levels of somatic pain (heat at 44, 45, and 46 °C), selected to reliably sample the range from nonpainful/barely painful to moderately painful (Green, 2004), on two body sites: the upper forearm and foot. In a second session, participants viewed pictures of painful events on others’ hands or feet (e.g., a toe being caught in a door), which were selected in pilot testing to span three levels of vicarious pain approximately matched to the heat in subjective intensity. We selected these images because they have been used extensively in past work (Jackson, Brunet, Meltzoff, & Decety, 2006), and have been shown to activate areas in the dorsal cingulate and anterior insula that putatively encode the affective dimension of pain. Participants were also instructed to take the perspective of the experiencer and imagine the painful stimulation was happening to them; this has been found to increase dorsal cingulate and anterior insula activation as well (Jackson, Meltzoff, & Decety, 2006). Before each stimulus, participants saw a cue instructing them to get ready for the upcoming trial, which allowed us to identify anticipatory activity and compare it to activity during stimulus viewing. After each trial, following a time-varying delay that allowed us to separate stimulus-related from rating-related brain activity, participants rated the subjective intensity of the experience. Though we selected stimuli at three levels of intensity for each of the arm and foot stimuli in each of the self- and other-pain modalities, our analyses focused on predicting variation in trial-by-trial intensity ratings in each modality.

Four other design features are particularly important for our ability to compare self-pain and other-pain brain representations. First, we developed models predicting within-person variation in reported experience. This focuses on brain patterns that are related to experience, and furthermore is much less noisy and subject to fewer confounds than predicting individual differences in reported experience. Second, we developed models that can make predictions about previously unobserved individual participants, or “population-level” models. This allows the model to be tested for specificity and generalization across different task variants by testing the model on new samples. This validation tells us much more about what each pattern measures than any single study is likely to be able to do. We compared the population-level models to idiographic models, in which the brain patterns predicting self-pain and other-pain are customized for each individual; the performance of these models was only marginally better than the population-level model, indicating that the brain patterns that predict experience are stable across individuals. Third, before the main fMRI study, we conducted pilot studies to select self-pain and other-pain stimuli that are matched in subjective intensity, eliminating a potential confound. In addition, subjective intensity is controlled for in the analysis, in the sense that we are looking for similarities and differences in brain measures that predict subjective intensity. And fourth, we built in two design features that allow us to test the representation of subjective intensity: (1) selecting stimuli that spanned the range of low, medium, and high subjective intensity in each of the self-pain and other-pain modalities; and (2) testing two body sites (upper and lower limb) in each of the self-pain and other-pain modalities, allowing us to analyze somatotopy and compare it to the established somatotopic organization of pain-related areas.

So what did we learn about shared brain representations for self- and other-pain? When we simply analyzed high-intensity stimuli vs. rest, we observed strong overlapping activation in the dorsal cingulate and anterior insula, among other regions, replicating the pattern found in previous studies. But when we compared the models trained to predict experience, the brain representations for self-pain and other-pain were distinct, involving many different areas across the brain and different local patterns within the dACC, insula, and other regions. We can see the same qualitative pattern across a series of analyses, each providing a slightly different window into shared representation. We’ll walk through the main analyses here.
A first way to look at shared representation is to compare whole-brain, population-level models. We have found that because such models capture patterns of activity within local regions and across large-scale systems, they often more accurately predict affect ratings than any single local region or individual “network” (see above for an analysis and examples). Testing the NPS, which was previously validated to track pain across multiple studies, we found that NPS responses strongly tracked stimulus categories and predicted pain ratings across both upper- and lower-limb body sites. This makes sense, because while some areas are somatotopically organized (particularly somatosensory S1, S2, and dorsal posterior insula), painful stimuli activate broad, bilateral patterns that overlap across body sites, and many individual nociceptive neurons have broad receptive fields that span body sites. However, though the NPS tracked somatic pain intensity strongly, it showed no response to images of others in pain (Fig. 8.8). In fact, NPS responses were significantly below zero. Further analysis revealed that this is because the NPS includes negative weights in some regions activated by pictures, e.g., visual cortex, and that focusing only on areas with positive pain-predicting weights (e.g., dACC, insula, S2, posterior insula, which are also nociceptive targets) showed no significant activation or deactivation for observed pain. Conversely, a population-level model trained to predict vicarious pain intensity (Fig. 8.8, right) showed very strong out-of-sample prediction of vicarious pain in new individuals, but showed no response to painful somatic events.

Fig. 8.8

Vicarious vs. experienced pain: Separate modifiability. This figure shows data from Krishnan et al. (2016), who tested three levels of each of somatic (heat) pain and vicarious pain (images of others in pain) in each of two body sites (upper and lower limb). The left panel shows the NPS (top), and brain responses in the NPS to somatic and vicarious pain stimuli, in warm and cool colors, respectively. The NPS responded only to somatic pain stimuli, across both body sites, with a magnitude proportional to stimulus intensity (and reported pain intensity). The right panel shows the “Vicarious Pain Signature” (VPS), a model trained to predict the intensity of vicarious pain ratings across both body sites. It responded in a graded manner to vicarious pain, but showed no response to somatic pain. This pattern of results, termed separate modifiability, indicates that neither pattern is strongly driven by shared psychological processes common to both conditions, including salience, arousal, and demand on attention. The NPS shows separate modifiability with other patterns for romantic rejection (Wager et al., 2013) and negative emotion (Chang et al., 2015) as well, suggesting that these various types of salient, arousing events have distinct brain bases. ***P < 0.001, **P < 0.01, *P < 0.05, error bars reflect within-participant standard error of the mean.
This pattern of cross-prediction results, in which one brain pattern tracks one and only one effect (self- or other-pain), is called “separate modifiability” (34). This pattern provides strong evidence for the separability of the brain processes underlying each type of “pain.” In particular, it helps to rule out the presence of potential shared, confounding processes like enhanced salience or attention. Imagine that each pattern was driven by a common process, which we’ll call “salience.” Salience is enhanced for high-pain vs. low-pain stimuli in each modality; is this what our models are capturing? If so, we should see at least some cross-prediction from self- to other-pain and vice versa. That is, a model trained on self-pain that actually captures salience should respond to the higher-salience vicarious pain stimuli more strongly than the lower-salience ones. But this is not what we observed. Therefore, these data suggest that the two brain patterns are not representing any common process that is shared by painful heat and observation of others’ pain.

The two patterns also involve different regions, and different patterns within regions. By taking bootstrap samples and re-running the predictive model many (e.g., 5000) times, we can obtain $P$-values for how reproducible each voxel’s contribution to the overall prediction is across participants. This allows us to interpret the statistically significant areas for both somatic and vicarious pain, which are shown in Fig. 8.8 at $q < 0.05$ false discovery rate corrected. The somatic pain signature (NPS) most strongly involved many areas that receive nociceptive information from the body. The vicarious pain signature (VPS) most strongly involved some areas related to mentalizing and social cognition, including the dorsomedial prefrontal cortex, and other areas less important for pain prediction here, including the amygdala. In addition, the global patterns and local patterns within dACC and other regions were uncorrelated. We can perform a more systematic analysis of the large-scale differences in the networks involved by comparing each predictive pattern (across all voxels, not only the significant ones) to identified resting-state cortical networks. This is a useful technique for helping to interpret the brain patterns. As Fig. 8.9 shows, the somatic pain-related NPS showed a concentration of positive weights in the “ventral attention” and “somatomotor” networks, and negative weights in the “dorsal attention,” “visual,” and “default mode” networks as defined by Yeo et al. (2011). The vicarious pain-related VPS displayed a very different pattern, including positive weights in the “default mode” and (to a lesser degree) “ventral attention” networks and negative weights elsewhere, including in the “somatomotor” network associated with somatic pain.
Different types of affect, different brain networks. Interpreting multivariate models in terms of identified systems is a challenge. One lens through which to view them is their loading on (spatial similarity to) large-scale resting-state networks. Here, we show the similarity of the NPS (left) and VPS (right) to each of seven resting-state networks identified by Yeo et al. (2011). The inner dark circle on the polar plots marks the zero-correlation point, so that points outside it show positive correlations between the brain model and network, and points inside it show negative correlations. These plots reveal quite different patterns across large-scale networks for the NPS and VPS, with similar positive loadings on the “ventral attention” network and negative loadings on the “visual” network for both, but very different (usually opposite) correlations with each of the other networks. These networks do not fully capture the models, and similar correlations with the “ventral attention” network does not imply that the two models activate the same locations or patterns within this network—but the pattern of differences across networks illustrates that the two models are different in their macroscopic as well as their mesoscopic (local pattern) organization.

Brain models for empathic care and distress

Prior literature: Neurosynth meta-analyses

Of course, pain and images involve different sensory modalities and cognitive processes, and so it should not be surprising that their brain patterns are differentiable. But the claim these data support is stronger than that: They suggest that the patterns that predict experience are not shared, even within the regions that are thought to encode common representations related to pain affect. A series of additional analyses provided converging evidence for this basic conclusion (Krishnan et al., 2016):

1. A “searchlight” analysis of local regions revealed that though some local regions predicted ratings in each modality, no brain regions showed substantial evidence for cross-prediction, and cross-prediction results were much weaker than training within-modality. These tests are fair and unbiased because whether a model is trained on the same modality as the test data (e.g., somatic pain model predicting somatic pain test data) or a different modality, the test data are taken from new individuals not used in model training.

2. Quantifying the effect sizes shows successful prediction within-modality, but weak cross-prediction. Training on vicarious pain explains 9 times less variance in somatic pain and training on somatic pain explains about 500 times less variance in vicarious pain.
3. Training a vicarious pain model without the visual cortex resulted in prediction that was just as accurate as the whole-brain model, and is not driven simply by visual activation or attention.

4. Re-training both pain-predictive and vicarious pain-predictive models within this study resulted in the same pattern of separate modifiability.

5. Analyzing the time-courses of NPS and VPS responses before, during, and after stimulation revealed that the responses were specific to the stimulus period in both models, and did not respond to either anticipation or post-trial response selection and reporting periods.

6. Training models with patterns customized for each person revealed the same pattern of separate modifiability. Such models are more susceptible to confounds, as they are much more flexible (different patterns for different individuals) and can pick up on different types of artifacts and confounds for different individuals (Todd, Nystrom, & Cohen, 2013). Here, customizing the models yielded little benefit in predictive accuracy.

7. Somatotopy models trained to predict whether stimulation was on the upper or lower limb in each modality showed strong somatotopy for somatic pain in sensory cortex and posterior insula, with hand and foot regions corresponding to those found in previous studies. These patterns could predict whether stimulation occurred on the upper or lower limb for an individual (averaging over same-site trials) with 90% accuracy. But no such somatotopy was observed in the vicarious pain condition, indicating that self-pain nociceptive pathways are not activated by observing others’ pain. Surprisingly, we also observed differentiable brain patterns for observed pain on hands and feet, which could be predicted with close to 90% accuracy from brain data—but these patterns did not transfer to somatic pain. Thus, somatotopic representations for self- and other-pain are also qualitatively distinct.

As may be evident by now, meaningful tests of shared representation across two types of mental events is possible, but it requires a number of analyses from different angles. Most importantly, the tests are only as good as the models: To compare brain representations of somatic and vicarious pain, one must develop models that are validated to track each type of pain, and ideally rule out other kinds of processes that the models might capture. In any domain, this will likely require prospective tests of pre-identified, population-level models like the NPS and VPS across multiple studies. Fortunately, this also seems possible.

Part of our process of testing the NPS and VPS was to test their performance, and in particular their separate modifiability, across studies. For example, is the NPS really an adequate model of pain? It could be that vicarious pain is like a different kind of pain, not heat, with brain patterns similar to that kind of pain. In Studies 2 and 3 of Krishnan et al. (2016), we tested the NPS on mechanical and electrical pain, respectively, and found robust responses. Subsequently, it has been generalized to other types of pain as well, as described above. Study 3 included both painful shocks and pictures of others in pain, allowing us to conduct a prospective test of whether the NPS and the VPS trained in Study 1 show separate modifiability in a new sample. They did.

Another type of test involves testing whether the VPS really captures “vicarious pain” in general, or whether it is capturing something related to the particular images we used or emotionally intense images in general. We tested this in a subsequent study, which also replicated the separate modifiability pattern for somatic and vicarious pain (López-Solà, Koban, Krishnan, & Wager, 2017). In this study, heterosexual women arrived for the fMRI session with their male romantic partners. The partners sat in the scanner room and a thermode was attached to their arm; the women viewed the male partner with pain-induction device attached through the scanner mirror. In the somatic condition, women experienced painful heat during fMRI. In the vicarious pain condition, a small change in the fixation cross indicated that the partner was receiving pain, and there were no other sensory cues. Even in this conceptually driven “cued empathy” situation, the VPS responded strongly and specifically to vicarious pain, and the NPS responded strongly and specifically to self-pain. The separate modifiability criterion held,
supporting the independence of the brain processes involved and the generalizability of the models to new samples and task variants.

The space of “prosocial emotions” like vicarious pain is still relatively unknown territory, especially when it comes to their brain bases. Recognizing others’ pain could just as easily lead to schadenfreude (joy in others’ suffering) and motivation to harm others as to empathic distress and helping. In some cases, this decision may be instinctual, but often—and particularly in humans—the decision to help others requires a deliberate choice (Schumann, Zaki, & Dweck, 2014). In addition to mentalizing about others’ suffering, there must be an act of affiliation, a recognition of the suffering other as worthy of help and comfort. Empathic distress is not the only, or perhaps even the primary, emotion that motivates helping behavior. It has an ambiguous and context-dependent relationship with helping motivation, as distress can lead to disengagement and burnout. In some caregiving professions and spiritual traditions, practitioners are taught to avoid getting “lost” in personal distress. Feelings of warmth and tenderness—or empathic care—may be more consistently related to helping. They may also be more sustainable, as they may be rewarding for those who experience them. Empathic care is intertwined with affiliation, a sense that another is close to or aligned with oneself.

In our work, we have developed models predicting how much people will donate their experimental earnings to charity. In one study, we created biographies of potential aid recipients—pictures and stories—that varied along a number of dimensions. The pictures varied on whether the recipient was old or young, black or white, male or female. The stories varied on whether the recipient was prosocial (e.g., helping others, volunteering), whether they were more or less responsible for their hardship (e.g., contracting AIDS because of a childhood blood transfusion or by injecting illegal drugs), whether monetary aid would have an instrumental value (be likely to help improve their condition), along with clues about the political and social identity of the recipient. By creating a “grammar” of statements that can be recombined in many ways, we created hundreds of unique stories, allowing us to investigate which variables predicted higher donation amounts to a target. One recipient was selected at random and the decision enacted; the money participants gave was, in fact, given to charity.

The results of this study indicated that giving was predicted by a combination of emotions and social cognitive judgments and attributions. Both personal distress and empathic care led to more giving, as did judgments that the person was not responsible for their hardship and that giving would have instrumental value. Perceived similarity—whether external (race, gender, age, socioeconomic status) or internal (values and attitudes)—did not predict donation amounts. A quantitative model combining feelings and judgments correlated over $r = 0.6$ with within-person variations in donations.

In a subsequent fMRI study, we asked whether empathic care and distress could be predicted by distinct multivariate brain models. Participants listened to 30-s audio stories of real individuals taken from charity websites, then subsequently received a reminder about each story and were asked to donate up to $100 (100%) of their experimental earnings to a charity that would help individuals in similar situations. After the fMRI session, participants made second-by-second ratings of empathic care or personal distress, which were averaged across participants to create normative time-courses for each biography. We trained multivariate models to predict the time-courses of each prosocial emotion, testing the models for sensitivity and specificity to each emotion in held-out participants (i.e., using cross-validation). The results identified a pattern that robustly predicted ratings of both empathic care and distress.

The care model, which was selective for care (not distress), involved increased activity in parts of ventromedial prefrontal cortex (vmPFC) and ventral striatum (VS), posterior cingulate, temporal-parietal junction, and anterior temporal cortex (Fig. 8.10). Among these, the vmPFC and VS are particularly associated with appetitive value and reward, including vicarious reward to others (Zaki, Schirmer, & Mitchell, 2011) and prediction of purchasing decisions (Grosenick, Greer, & Knutson, 2008; Knutson & Genevsky, 2018) and donations to charity (Genevsky & Knutson, 2015; Genevsky, Yoon, & Knutson, 2017; Hare, Camerer, Knoepfle, & Rangel, 2010). The vmPFC is particularly associated with self-referential thoughts (Denny, Kober, Wager, & Ochsner, 2012), perceived closeness (Krienen, Tu, & Buckner, 2010; Tamir & Mitchell, 2010), and compassion (Klimecki et al., 2014), and increases after compassion meditation training (Engen & Singer, 2015). And the portion of the TPJ...
and posterior cingulate activated are particularly associated with social cognition and recognizing others’ actions and intentions (Carter & Huettel, 2013; Miele, Wager, Mitchell, & Metcalfe, 2011). Thus, the brain model that predicts empathic care brings together elements of self- and other-oriented cognition and positive valuation.

**Fig. 8.10**

Brain models for empathic care and distress. In this study, participants listened to audio narratives of other hardships. Two models were trained to track normative (group) moment-by-moment ratings of empathic care (warmth and tenderness) and distress. The empathic care-selective model (purple, top) included high weights in regions associated with reward, value, and self-relatedness. We compared this with a neurosynth-generated meta-analytic map for “value” (purple, bottom), and identified strong overlap in the voxels included. The empathic distress-selective model (orange, top) included premotor and inferior frontal regions associated with self-other action mirroring and negative affect in other models (e.g., for romantic rejection). We compared this with neurosynth’s “mirror” map (orange, bottom) and identified substantial overlap. These findings suggest that these two empathy-related emotions have distinct brain bases. The right panel shows how these models can be further explored and validated in new studies. Because they were trained on group rating data, we were able to conduct an online study in which participants rated the same narratives on other emotions, shown at right. These time courses were regressed on the fMRI time courses for the two models. The empathic care model correlated most strongly with reported “care” in the online sample, and less strongly with other emotions. The empathic distress model correlated positively with a range of negative emotions, and negatively with ratings of “positive” and “happy.” Thus, the brain systems underlying empathic care in particular may be relatively unique to care as opposed to other positive emotions.
The personal distress model strongly tracked distress but was not highly selective, as it also tracked empathic care. This model involved premotor regions associated with “mirror neurons” and recognition and imitation of others’ actions (Keysers, 2009; Losin, Iacoboni, Martin, Cross, & Dapretto, 2012; Losin et al., 2015). To test whether these two models were reliably different and help interpret them, we compared each model to two meta-analytic patterns derived from neurosynth.org, an online meta-analysis tool that links reported coordinates from thousands of neuroimaging studies to terms and topics used in the papers (Yarkoni et al., 2011). The empathic care pattern was very similar to the meta-analytic map for “value,” as indicated by a Dice coefficient—a measure of overlap across binary maps—of 22%, but not to the meta-analytic map for “mirror” (1%). The empathic distress pattern showed the opposite, overlapping substantially with the meta-analytic map for “mirror” (52%) but not with “value” (0%).

Another open question concerns which emotions these models track most strongly. Does the empathic care model track something different than “happiness” or “positive affect” in general? An advantage of training normative, population-level models is that they can be annotated with additional data. In this case, we collected second-by-second ratings of each biography from 200 additional participants in an online study. The time courses of each brain model were then correlated against ratings of each of ten different emotions (Fig. 8.10). The care model correlated with empathic care and a blend of other emotions including surprise and positivity,
suggesting that it is a relatively unique experience that is not reducible to generic positive affect or emotional salience. The distress model correlated with a range of negative emotion ratings approximately equally strongly, but less strongly with positive emotions, suggesting that it tracks a relatively undifferentiated form of negative affect. This illustrates how population-level brain models can be retroactively tested by correlating them with measures collected on the same stimuli in subsequent experiments.

This work leaves a number of questions to be addressed in future studies: What is the relationship between vicarious pain and empathic distress across stimulus types? The VPS and empathic care patterns appear to be qualitatively different; is this a function of the types of stimuli used, the social judgments made, or participants’ intentions in the context of the different types of studies? And what is the relationship with helping behavior? Stronger activity in both models predicted donations, but only weakly, suggesting a variable relationship between the brain systems underlying these feelings and action. Our view is that helping decisions are complex, and depend only partly on empathic feelings and attributions—they are also influenced by cognitive policies about how much to give, relative comparisons about how much one has given previously and whether one has “done enough,” personal thoughts about the value of keeping the money for oneself, and likely other factors.

Conclusions and Implications

Given the complexities and variability involved in human affect and decision-making, it is remarkable that it is possible to identify consistent brain correlates of multiple types of affect across individuals. This appears to be true for complex social emotions like empathic care as well as basic, evolutionarily conserved processes such as pain. This did not have to be the case; the construction of pain, feelings of rejection, vicarious pain, anger, sadness, and other emotions could have been very different across different individuals, making it impossible to identify stable brain predictors. In a famous example from philosophy, one can never be sure whether my experience of the color “red” and yours are similar or completely different; we have learned to label them the same way regardless of our inner experience (Chalmers, 2007). However, in this case, the brain systems most closely linked to feelings—and which presumably play a central role in their construction—appear to be relatively conserved across individuals. This has been much more extensively tested for some forms of affect (experimentally evoked pain) than for others (empathy), and there is much more work to do, but the way forward seems promising.

These brain models reveal another interesting conclusion about how affective processes are organized in the brain. We are used to thinking of rejection, vicarious pain, somatic pain, hunger, thirst, and disgust as birds of a feather in some respects: They are all negative experiences, shaped over the course of evolution to elicit escape and avoidance. Descriptive models that group human judgments have consistently found that negative emotions are grouped together in our conception and set in contradistinction to positive emotions. But this conceptual similarity need not reflect similarity in the underlying brain systems, which may more precisely determine how they jointly or separately arise and how they might interact with one another. Put simply, there may be no common representation of “negative affect” in the brain. Different systems might, having evolved to respond to particular environmental demands, “feel like something” (or, in more technical terms, be associated with conscious “qualia”). Viewed in this light, it is easy to imagine one system that evolved to escape from the imminent damage caused by coming too near a fire, and a different one evolved to avoid dying of thirst or being attacked by an angry group member. We may have many systems that represent many types of negative and positive affect, organized in part by the eliciting stimuli, the canonical organism-environment relationships or “situations” involved, the types of actions afforded and the time scale involved, and more. The critical variables that carve the affective brain “at its joints,” determining when one affective brain system versus another is engaged, remain to be determined.

This “multiple affect systems” view stands in stark contrast to recent conceptions of the organization of the affective brain. Many studies have highlighted the broad convergence of multiple types of affect on the anterior insula and cingulate (Klimecki et al., 2014; Lamm et al., 2011). Lieberman et al., for example, proposed that the dACC reflects a unitary system that responds to events “relevant for survival” (Lieberman & Eisenberger, 2015). Eisenberger and Cole relate physiological responses in the endocrine and immune systems to a general “alarm

https://eproofing.springer.com/books_v2/printpage.php?token=KSks_811CEb-2TrVqJrl4ySCIWRQhAp7SiaDc8oWE8
system” in the brain (Eisenberger & Cole, 2012). And our own work has highlighted commonalities in the brain systems involved in constructing multiple types of emotional experiences (Ashar, Andrews-Hanna, Dimidjian, & Wager, 2016; Kober et al., 2008; Roy, Shohamy, & Wager, 2012). There may indeed be commonalities at the level of broad systems and concepts like constructing representations of schemas and assigning personal meaning to actions and events. But at the circuit level, the brain processes that mediate particular forms of negative affect—and our behavioral and physiological responses to them—need to be defined with increasing specificity. In animals, there is now overwhelming evidence that specific neural populations mediate specific affective behaviors in response to specific contexts (Lammel, Tye, & Warden, 2014). For example, different populations of neurons in the dACC mediate different types of foraging behavior (Kvitsiani et al., 2013) and specific aspects of pain (Dale et al., 2018; Tan et al., 2017; Zhang et al., 2018). In humans, different local patterns within the dACC track evoked pain and specific types of negative affect and cognitive control (Kragel, Kano, et al., 2018), though there appears to be a pattern that generalizes across multiple types of pain. Even within autonomic responses to stress, largely different systems mediate increases in skin conductance and heart rate (Eisenbarth, Chang, & Wager, 2016), though there appears to be a common core related to the vmPFC.

The implications of this “multiple affect systems” view are not merely academic abstractions. They bear concretely on how we move forward in identifying the systems that confer resilience or risk to disease, and target those systems with biological and psychosocial interventions. If there is a unitary system for “negative affect,” one can measure its function with multiple behavioral readouts in humans and animals, and probe it with a variety of interchangeable stimuli (faces, sounds, memory cues). We could identify the molecular substrates of this system and develop pills to cure depression, anxiety, and pain. We could characterize genetic and environmental precursors that lead to its dysfunction, and thus direct prevention and treatment resources to those at greatest risk. But if biology has taught us one overarching lesson, it is that this way of thinking will not work. Biology is complex, and its pathways and interactions are myriad. Molecular mechanisms that operate in one strain of mouse may not be operative in another, much less in other species. As a result, advancing treatments for mental health disorders has proven extremely difficult. By some accounts, there are no new classes of drugs for depression, anxiety, or pain, in spite of over a trillion dollars spent on drug development over the past decade. Identifying specific pathways and targets, and relating these to specific treatments, is a promising way forward. The future of neuroscience lies in embracing complexity—but in a limited way, simplifying and generalizing where we can, and throughout hewing to the lines nature draws for us that carve the affective brain at its joints.

Acknowledgments

We are grateful for support from the National Institutes of Health, including grants R01DA035484 (Tor Wager), R01MH076136 (Tor Wager), R01MH116026 (Luke Chang), T32DA017637 (Philip Kragel), and U01 50047-78051 (Lisa Feldman Barrett). MATLAB code for analyses used in papers and figures is available at: https://github.com/canlab.

References


https://eproofing.springer.com/books_v2/printpage.php?token=KSks_811CEb-2TrVqjr14ySCIWRRQhAp7SiaDc8oWE8


