The Predictive Psyche

Contents

I use the terms mind and psyche interchangeably.

The Anatomy of the Mind

Predictive Models

I propose that the mind is fundamentally composed of predictive, or generative, models of behaviour (Friston, 2010; Clark, 2013). By generative model, I mean a hierarchical prediction apparatus that attempts to model the underlying causal nature of the world as observations are made (Parr et al., 2022). As observations are made, the activation patterns of these predictive models are altered.

Individual nodes in generative models are also generative models. Thus, the term generative model will be used to refer to both the whole structure of a generative model, and the fundamental unit that generative models consist of. From this lens, generative models are hierarchical and modular. For example, the generative model for the direct input from the senses is usually used as a basis for all other generative models, modularly plugged into them, or used across them simultaneously, as necessary.

Generative models do not have a binary connection or disconnection: they become habitually easier to connect as integration occurs (more on integration later).

As observations are made, the activation of the underlying model, and thus underlying probability distribution for the source of these observations changes. These distributions are used to make predictions about future observations, and thus guide relevant actions. A generative model being 'active' means that it is currently making predictions about the world.

Over time, the models themselves also update based on the prediction errors the model encounters.

Action Systems

Actions are selected from a set of available and similarly hierarchical action systems, based on the goal of the action system (Panksepp and Biven, 2012). Proposed action systems are that of exploration, finding food, attracting mates, self regulation, escaping danger, and many more. Actions taken within these action systems are meant to balance both epistemic gain, to inform future predictions made by related generative models, and the direct goals of the system.

Just like generative models, action systems are also hierarchical. Just like in action systems, where identical action tendencies, or lower level systems of behaviour, are proposed to be able to be used by multiple independent action systems, I propose that sections of the generative models can be taken and applied to different priors. Thus, greater degrees of dissociation produce changes in increasingly fundamental priors in these models. There is often much that is more fundamental beyond the narrative self, or components of the narrative self, such as the vision system, models that control the relationship between the central nervous system and the autonomic nervous system, and much more.

Action systems become deeply associated to generative models. It is proposed that there are psychological fault lines separating action systems and their associated generative models that become slowly less pronounced over development. If typical development is disrupted, the psyche can easily split along these fault lines, as per the Theory of Structural Dissociation (van der Hart et al., 2006).

An important addition to note is that these generative models do not just predict the causal structure of the external world, but also the internal world: interoceptive signals are predicted and observed (Seth, 2013). In this way, I propose that there is no fundamental difference between the internal and external world.

Our relationship to these internal signals is extremely important, more on this later.

Elaboration of Models

The increase of complexity of a generative model is typically referred to as elaboration. As the generative models get increasingly elaborated, they can form increasingly abstract representations for the psyche's interaction with the world. These top level, deeply layered and highly abstract variables can become relatively unchanging over time, as they attempt to model fundamental realities about the interaction between the self and the external world.

For a generative model to be most successful, it must eventually learn to model itself. Let me give a few examples of why this is the case. A generative model that can model its own behaviour can model several steps ahead, knowing its own expected output before acting, and adjusting for it as necessary. Moreover, a generative model that can model itself is able to change itself most deliberately. It can develop a model to more purposefully interpret prediction errors in a domain, and change the underlying model in the way it has learned best. From an information-theoretic perspective, it is always advantageous for a generative model to model itself at as many recursive levels as possible, for each recursive level has the capacity to provide strictly more information than would otherwise be possible.

When a generative model becomes sufficiently self referential, a model of the model's relationship to itself can emerge. This forms what is commonly known as the narrative self. The management of prediction errors with a carefully tuned model is essential for the preservation of narrative selves. The ego is the dominant narrative self at any one time.

This ego forms the centre of awareness. However, awareness is carefully controlled by models the mind uses that are not fundamentally different than any other. Thus, what is in awareness does not necessarily encompass all, or even a significant amount of what generative models are currently active. The broadcasting of particular brain activity into conscious awareness can be seen as being in line with Global Workspace Theory (Baars, 1988, 2005).

I make the distinction between three parts of the psyche, in line with Jung (1959): the Self, the Subconscious and the Ego.

The Self

The Self is the actual centre of our psyche. From it originates the most fundamental drive of human nature. This drive is inherently integratory, and the direction it compels us in can be described as the process of individuation.

Jung proposed that there are psychoid, or psychological elements so fundamental they are rooted in biology, components of the Self (Jung, 1969). He proposed some examples of these that can be found in a search for the Jungian Archetypes, but conceded that there may be an infinite number of these. They may also be individually different. He derived these archetypes by studying persistent and repeated cross-cultural symbolism and patterns, such as the idea of a descent into something dangerous and returning a new person, or a divine disaster where only a chosen one remains. There is little doubt that there is something fundamentally biological about these symbols.

The Subconscious

The subconscious contains the generative models that are relegated out of conscious awareness at any one time. Thus, the boundary between the conscious and the subconscious is always mercurial and malleable. Jung (1969) proposed that the subconscious is constantly regenerated from the Self, such that we are always presented with parts of the self to integrate into the narrative self, and broader generative models.

The Ego

The Ego is the part of the part of the psyche at the very centre of conscious awareness. As the conscious awareness expands, the Ego attempts to integrate and thus own what it sees, which is both the source of development and the cause of a lot of ailments. For example, if conscious awareness is expanded too far too quickly, the Ego can

collapse into self hatred and childlike incapability, it can impulsively attempt to own everything it sees and expand into narcissism, or it can destabilise itself in its attempts towards integration and lead the mind into psychosis.

Memories

A memory is a generative model's predisposition to reactivate in a prior way, stored in the weights of the generative model. Traumatic memories are often encoded with extremely high precision priors, meaning that the mind is extremely certain on how to interpret the world, despite it not necessarily matching the ground truth. This can create the experience of flashbacks.

Integration

I propose that the Self's fundamental drive towards integration is enacted on the Ego through interoceptive signals. These interoceptive signals pull the Ego towards integration of specific generative models.

Perhaps the integratory drive we are compelled upon by the Self is a consequence of the shared causal structure of the body. Humans act as a single entity in the world, held together by the body, and thus the body forces integration through interoceptive signals.

Fundamentally, integration is holding two differences simultaneously in conscious awareness, and then following the interoceptive signals sufficiently to combine them. Fundamentally, this is about connecting distributions in generative models through the use of latent variables that explain both, thus forming a cohesive generative model that can predict both phenomena from the same fundamental priors.

Just as the Ego's field of awareness can include or exclude activated generative models, it can be expanded or contracted to include or exclude these interoceptive signals.

However, this is not a precise process. As interoceptive signals are discarded or excluded, other generative models or attached memories may also be incidentally discarded.

Moreover, this is also not a painless process. The purposeful exclusion of interoceptive signals can only be maintained through what I will refer to as suffering.

Buddhist philosophy stated attachment as the root of all suffering (Buddhagosa, 2010). This is commonly explained as attachment to external phenomena, such as our loved ones, our possessions, or this life. I propose that the attachment he was referring to was fundamentally about attachment to generative models. The refusal to integrate accrued evidence of the world into our models, which, as well as being a painful process, would fundamentally change them and our relationship to reality.

Resistance to Integration

The integration of generative models requires holding the interoceptive signals in conscious awareness. For unpleasant signals, this is inherently painful. Thus, we come to a fundamental tradeoff: that of pain and suffering. Our Self's constant pull towards integration requires us to integrate painful signals as we inevitably accrue increased evidence of the world through generative models.

Given this tradeoff, I propose that the integration our Ego engages in is selective and controlled. In fact, a generative model producing our response to interoceptive signals is also formed, no different than any other. Thus, we come to another significant point of self referentiality, and fundamental restriction of outcome: the generative model predicting our internal signals can only be altered by engaging with these internal signals, which may be restricted by the very mechanism that needs to be integrated.

Because of this, a jolt, or shock, may be required for people to begin integrating beyond the narrow capability that the model allows. Such a jolt, or destabilisation of the predictive machinery, may only be able to be provided by neurostimulation technology or psychedelics.

I propose that the system's resistance to integration is based on three fundamental drives: The rejection of unnecessary stimuli to preserve real world accuracy.

One common example of this is stimuli prompting the narrative self to integrate information that would change it. It is beneficial for a narrative self to have this reflexive self protection: a destabilisation of the highly fundamental priors that predicts how the model interacts with the outside world should be done extremely selectively, and only on very persistent and reliable evidence.

One other major category is the implicit drive towards complexity reduction. Complexity reduction can be seen as the fundamental drive to ignore irrelevant stimuli, and only engage with stimuli that are meaningful. Only generative models formed from persistent and reliable signals should be integrated.

Dissociation

Dissociation is used to refer to two highly related but distinct phenomenon (van der Hart et al., 2006): a maladaptive contraction of awareness, and a maladaptive lack of integration between generative models. In the literature, a similar distinction is often made when referring to 'detachment' and 'compartmentalisation' as two separate mechanisms of dissociation.

First, let me address contraction of awareness. Contraction of awareness is not always a bad thing: contraction of conscious awareness is the primary mechanism for ignoring irrelevant, destabilising or threatening (from the perspective of the narrative self) information.

Moreover, a lack of integration between generative models can also be beneficial: it is not useful, and thus a waste of energy, to habitually connect generative models associated with reproduction with that of consumption of food.

Learning

One of the primary ways that generative models are updated is relationally. One of the most important generative models that is learned relationally is the generative model of aspects of the self, and aspects of the body.

A disconnect between the learned self and the actual self is the foundation of Fonagy's Alien Self theory of BPD (Fonagy et al., 2002; Fonagy and Allison, 2014).

In a similar vein, a disconnect between the learned and actual capacity for intelligence and rational ability can be thought of as the root of alienation of our own inherent capacity for independent, critical thinking, and ability to self-teach.

Nonrealisation

The core of the psyche, even what is presupposed as a healthy psyche, is nonrealisation of self and others. A healthy ego acts as if it is the centre of the psyche. However, the more developed and mature the psyche is, the more realised that this is ultimately not true. The more that this can be realised, the weaker the secondary interoceptive response to fear that is generated, especially existential fears such as fear of death.

Interoceptive Strength as Overexcitabilities

Dabrowski's theory of Positive Disintegration posited that the speed at which someone can undergo positive disintegration (which can be viewed as the adaptation of socially absorbed priors to better reflect internal experience and landscape) is mediated by overexcitabilities, all of which describe different measures of intensity of internal experience (Dabrowski, 1964, 1967).

This aligns well with this posited model: people with greater interoceptive experiences feel a greater pull towards integration, and thus have a greater ability to dismantle deep rooted existing societal priors in favour of a model more in line with their actual internal reality.

Differentation as Antifragility

With this being said, there is an advantage to the internal model maintaining some lack of integration. The simpler the model, the less robust it is to changes in the environment and unexpected behaviours. Thus, complexity is not something to be minimised. This ties to Nicholas Nassim Taleb's concept of antifragility (Taleb, 2012).

Bibliography

Bernard J. Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, 1988.

Bernard J. Baars. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in Brain Research*, 150:45–53, 2005.

Bhadantacariya Buddhagosa. *The Path of Purification: Visuddhimagga*. Buddhist Publication Society, 4th edition, 2010.

Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

Kazimierz Dabrowski. *Positive Disintegration*. Little, Brown and Company, 1964.

Kazimierz Dabrowski. *Personality-Shaping Through Positive Disintegration*. Little, Brown and Company, 1967.

Peter Fonagy and Elizabeth Allison. The role of mentalization and epistemic trust in the therapeutic relationship. *Psychotherapy: Theory, Research, Practice, Training*, 51(3):372–380, 2014.

Peter Fonagy, György Gergely, Elliot L. Jurist, and Mary Target. *Affect Regulation, Mentalization, and the Development of the Self*. Other Press, 2002.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

Carl Gustav Jung. *Aion: Researches into the Phenomenology of the Self*, volume 9, Part 2 of *Collected Works of C. G. Jung*. Princeton University Press, 1959.

Carl Gustav Jung. *The Archetypes and the Collective Unconscious*, volume 9, Part 1 of *Collected Works of C. G. Jung*. Princeton University Press, 2nd edition, 1969.

Jaak Panksepp and Lucy Biven. *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions*. W. W. Norton & Company, 2012.

Thomas Parr, Giovanni Pezzulo, and Karl J. Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press, 2022.

Anil K. Seth. An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, 2:395, 2013.

Nassim Nicholas Taleb. *Antifragile: Things That Gain from Disorder*. Random House, 2012.

Onno van der Hart, Ellert R. S. Nijenhuis, and Kathy Steele. *The Haunted Self: Structural Dissociation and the Treatment of Chronic Traumatization*. W. W. Norton & Company, 2006.