Explaining Myself Away

Predictive Coding and The ‘Social-Self’

Sarit Pink-Hashkes

Neurophilosophy, CNS Master Program Radboud University

**Abstract**

The predictive coding framework (Friston & Kiebel, 2009) claims that a core task of the brain is to represent the environmental causes of its sensory input. The brain does this by creating hypotheses that predict what the sensory input should be. The hypothesis that generates the best predictions then determines perceptual content. This framework has been gaining increasing attention as a means to explain many processes in the brain. This paper will use the predictive coding framework to shed light on the brain’s ability to create a singular higher level ‘social-self’ which is commonly called a personality.

Researchers (Apps & Tsakiris, 2013; Limanowski & Blankenburg, 2013) have used this framework to explain the creation of a 'minimal-self', existing even in primitive life, allowing for differentiation between the organism and the environment. They claim that this minimal-self arises by the brain creating top down multi-model probabilistic representations that explain away much of the separate sensory information, binding together these information streams based on prior probabilities and thus minimizing prediction error.

This minimal-self has been shown to be flexible and prone to mistakes by examining illusions such as the Rubber Hand Illusion. In these types of illusions the most probable explanation of two (or more) data streams occurring at the same time result in the wrong integration of sensory data. For instance, the visual stream of a rubber hand being touched and the proprioception stream carrying the sensation of touch from sensors in the hand being stimulated result in a top-down explanation that the rubber hand is part of the body and the ‘minimal-self’.

The goal of this paper is to take this idea one step further and use the predictive coding frame work to explain the creation of a higher ‘social-self’, or personality, showing that the ‘social-self’ is a construct created to minimize prediction error from social interactions.

The second part of this work will examine the flexibility of this ‘social-self’ and how changes in social interactions can affect the ‘social-self’.

**Introduction to predictive coding:**

*“Life is interposed between two energy levels of the electron”* [*Albert Szent Györgyi*](http://en.wikipedia.org/wiki/Albert_Szent-Gy%C3%B6rgyi)

Life on earth arises from the sun’s photons exciting electrons to a higher-level. The survival of life depends on slowing the downwards cascade of this energy to maintain a local minimization of entropy, differentiating between inside and out, between environment and organism (Cook, Carvalho, & Damasio, 2014).

Interaction with the environment is costly but needed for survival. The ‘free-energy’ principle explains how [biological systems](http://en.wikipedia.org/wiki/Biological_systems) “maintain their states and form in the face of a constantly changing environment” (Friston, 2010) by restricting themselves to a limited number of states. The framework suggested to implement this is predictive coding, which claims the brain’s core task is representing the environmental causes of its sensory input for simulating the best interaction with the environment.

Predictive coding explains that the cognitive system is ordered hierarchically, in levels (Bastos et al., 2012). For any pair of levels, the higher-level will have hypotheses predicting the bottom–up signals from lower-levels. If the predictions are good, the bottom–up signals will be ‘explained away’. Only discrepancies between the winning prediction and the bottom–up signal remain as ‘prediction error’.

This framework explains that sensory information is processed probabilistically, with prior predictions and posterior inferences made based on Bayesian optimized probabilities. Past contextual information is part of competing hypothesises. An example is Dr. Jekyll and Mr. Hyde (Kilner, Friston, & Frith, 2007) both holding a scalpel. Two competing hypothesis are created, predicting a helpful operation or a gruesome murder. Different higher-level contextual priors are used to determine the wining hypothesis. Nurses in an operating room will have a prior that the Dr. is performing a lifesaving operation while a crowd watching a horror movie will expect a murder.

This framework also connects perception and action (Friston, 2010). An action is an inversion of the perception model. A prediction of an action causes lower-layers to minimize the prediction error created and create the action.

**Mechanical brain theory**

One of the earliest theories connecting the creation of a higher-level self to social predictions is the Mechanical Mind theory (Barlow, 1990). Barlow explains that ‘mind’ is “the concept we use to describe the source of other people's, and our own, behavior.” This fits a definition of a higher-level ‘self’ or ‘personality’ and here these concepts will be interchangeable. Two concepts that aren’t interchangeable, as Barlow explains, are the brain, which actually controls behavior, and the mind, which is the brain's models of itself and other brains.

Barlow’s explains there are three stages in creating a high-level self. His work isn’t directly related to the predictive coding framework but here an expansion will be presented. Alice and Bob’s brains will be used as examples.

1. Creation of a ‘minimal-self’ - Alice’s brain attempts to minimize prediction error across the cortical hierarchy. The best explanation for much of the information related to movement production and sensory input is that they have the same cause. Thus, a ‘minimal-self’ model is created as a higher-layer in the cortical hierarchy binding separate information streams together explaining lower sensory input. This has been described in depth (Apps & Tsakiris, 2013; Limanowski & Blankenburg, 2013) and isn’t the focus of this work. However, it’s interesting to note that evidence exists that this minimal-self already exists at birth (Gallagher, 2005).
2. Modeling the other - When Bob is near Alice, Alice’s brain receives sensory information from Bob’s actions. To explain and predict these signals Alice’s brain can use it’s own ‘minimal-self’ model and extend it to create a model of Bob. The mechanism enabling this is suggested to be the mirror-neuron system (Frith & Frith, 2012; Kilner et al., 2007). Observations of Bob’s movements cause part of Alice’s ‘movement’ neurons to activate, creating prediction errors. Higher-levels of Alice’s cortical hierarchy, including Alice’s minimal-self model, minimize these errors. Barlow explains that good models stringently select the data they represent. “The predictive power of a model depends on its correct identification of the dominant controlling factors and their influence, not upon its completeness”. Alice’s brain simplifies Bob’s brain, modelling it as a single, coherent representation with little variance through time. In effect, Alice gives Bob a ‘personality’ or a ‘social-self’, creating a higher abstraction layer to explain the prediction errors from his behaviour. Just as a higher abstraction layer of a ‘minimal-self’ is needed to explain lower sensory data, a higher abstraction layer is needed to bind together Bob's observed actions.
3. Modelling interactions with others - for Alice to predict her interactions with Bob the ‘minimal-self’ model is not enough, actually it’s too much. Alice’s model of Bob is composed of averaging samples through time of Bob’s external behavior while Alice’s minimal-self is a less abstract more detailed representation focused on present input, which includes more sensory information. Thus, Barlow claims, Alice needs a higher-level model of herself to manage to model interactions with Bob’s high-level model. Alice’s brain needs a model of an ‘averaged through time, simplified Alice’. Barlow’s paper doesn’t discuss the mechanism how this higher-level self comes about. This is the focus of the next section. To conclude, we will note that the extensive development of the PFC is likely to be what allows humans to create higher-order models inhibiting the ‘here and now’ of sensory perceptions (Gerrans, 2007).

**A model of a model of a model**

To create a higher-level social-self, the proposed mechanism is as follows; Alice’s brain recursively uses the modelling ability it has built to model Bob, modelling Bob’s model of Alice herself. This too might use the brain’s mirror system. For instance, Bob smiling at Alice will activate the same top-down cause of a smile in Alice. Since Alice’s reason to smile is a pleasurable experience, an inference that Bob is having a pleasant experience with her ‘minimal-self’ is created. If this happens repeatedly, a higher-level prior will form and Alice’s brain will model herself as someone that Bob likes. Just as Alice needs a mirror to create a model of Alice’s face by binding perceptions from facial sensations and visual input (Apps & Tsakiris, 2013), Alice needs interactions with others mirroring her behavior to create a higher-level social-self. This recursive modelling loses much of its accuracy but the loss prevents this recursiveness from exploding to infinity. If Alice’s brain were to have an exact lossless model of Bob (and of Bob’s model of Alice), it would need the same amount of resources needed to actually be Bob.

Consequently, creation of a social-self must begin with the first social-interaction Alice has with her mother. Mother-child bonding is important even in regulating infant’s physiology (Hofner 1994). Infants aren’t born with many of the prediction abilities needed to interact with the environment. They learn these priors by synchronization with nearby attachment figures (Constantino, 2001). This is seen throughout infants’ development. Whenever Alice is in a stressful situation she will look to her mother’s reactions adopting her mother’s priors (Striano, Vaish, & Benigno, 2006). These synchronization and imitation mechanisms probably depend upon the mirror system and are also apparent from birth (Gallagher, 2005). This explains the accumulated knowledge showing that mother-child’s attachment styles have immense effects on formation of an adult personality (Constantino, 2001).

As Alice grows-up, she interacts with others and begins modeling them. She also begins modeling them modeling her. This second recursive loop reflects back into Alice’s own simplified representation of herself. Within the predictive coding framework, this could be described as internalizing a prediction from an outside agent. Once this prediction is in Alice’s brain it will create an inversion of the modeling process and lower parts of her cortical-hierarchy will change to minimize the prediction error created by this prediction. This fits with the documented Pygmalion affect showing for example that teacher’s expectations influence student performance (Rosenthal and Jacobson, 1968, Tauber, 1998).

**Hunter-gatherer Vs. modern-human**

Predictive mechanisms are shaped by the environment, which in primates case, became more social (Brown & Brüne, 2012). Brain’s structure reflects this. For instance social-network size and social-status in Macaques correlate to amount of gray-matter and connectivity of various areas of the cortex (Sallet, 2011). The social environment of humans has changed drastically in 200,000 years. An Alice living today will encounter more people in one day of shopping than prehistoric Alice encountered throughout her life.

Early humans lived in small close-knit hunter-gatherers groups requiring Alice to accurately predict only a few others. These others would also have an accurate prediction of this early Alice projecting their complex understanding of her back into her own predictions of herself.

These predictions probably included movement patterns needed for joint coordinated hunting as in apes (Boesch, 2005) and were likely based on increased physical play (Gray, 2009). Biological and anthropological evidence show these societies were not sexually monogamous (Ryan & Jetha, 2010) so Alice would predict many other’s sexual behaviors. Prehistoric Alice’s emotional interactions were probably different from today too as modern society mostly values goal-oriented rationality, correlated with the DLPFC’s ‘cold’ executive-network. It has been theorized (Jaynes, 1976) that in primitive cultures the social-self was less constrained and included sub-agents manifesting as auditory hallucinations guiding behaviour. Possession of sub-agents and hallucinations as part of religious ritual normally occur even today in some tribes (Richardson, 1979), while this behavior is considered a mental-illness in modern society.

Minimization of prediction error can explain these changes. One way to reduce prediction error is to actively intervene and change the environment, making it more predictable (Clark, 2013). Changing the environment on large-scales would require cooperation on large-scales. Farming, dams, bridges and cellphones require cooperation on a scale unique to humans. Hence, a feedback loop began, more control over the environment created larger social groups that allowed for even more control over the environment. Some suggest the unique aspects of human cognition is driven by social cooperation (Moll & Tomasello, 2007). Indeed, these large-scale cooperations probably shaped brain structures and other aspects of cognition, for instance increasing the importance of language. Language is a highly compressed symbolic representation of complex data-structures needed for large-scale cooperation. Language also helps simplify representations of others even more. Instead of prehistoric Alice representing Bob’s movements, sexuality, and emotional-state, today’s Alice simply models Bob as ‘interesting’.

As humanity moved from an in-depth prediction of a small number of close social connections to predicting masses of shallow social interactions, new mechanisms developed to minimize prediction error in this new social niche. Social-norms, increased influence of language, mass-media and even the flourish of chain-stores can be seen as attempts to minimize prediction error within an ever-growing social-sphere. However, along the way the details have been lost. For instance, only professional improvisational dancers and musicians can predict each other’s movement (Noy, Dekel, & Alon, 2011).

As predictions of others became shallower, focused on ‘left-brain’ language and logic abilities, so did Alice’s model of her own brain. Today’s Alice probably has a very poor model of her own movements, sexuality and perhaps even her emotional state.

Within the predictive coding framework depression is described as an inability to change priors (Corlett, Frith, & Fletcher, 2009) and an over-stable state with less free-energy (Carhart-Harris et al., 2014). An over-simplified, high-level social-self, explaining away too much of the lower-layers of the brain’s hierarchies might just be one of the causes of this 21st century plague which is correlated to lack of social interaction (Nezlek, 2000). Research regarding solitary-confinement has proven the disastrous effects of lack of social interaction (Grassian & Hill, 1993).

In summary, attempting to minimize prediction error from social interactions is what brought about the creation of a single higher-level social-self. The mechanisms enabling this are probably dependent on the brain’s mirror system. As shown, changes in the social structure are likely to have limited the ‘social-self’, creating an increasing gap between the ‘minimal-self’ and the ‘social-self’. Friston explains that humans cannot minimize ‘free-energy’ by simply sitting in a dark room because in the evolutionary processes they went through a dark room isn’t a survivable niche environment (Friston, Thornton, & Clark, 2012). Following the increasing limitation in Alice’s social interactions, one might ask what type of niche environment the human race is heading to, and whether it isn’t a dark room, at least in the social domain.

**References**

Apps, M. a J., & Tsakiris, M. (2013). The free-energy self: A predictive coding account of self-recognition. *Neuroscience and Biobehavioral Reviews*, *41*, 85–97. http://doi.org/10.1016/j.neubiorev.2013.01.029

Barlow, H. (1990). The Mechanical Mind. Annu. Rev. Neurosci. 13:15-24

Bastos, A. M., Usrey, W. M., Adams, R. a., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical Microcircuits for Predictive Coding. *Neuron*, *76*(4), 695–711. http://doi.org/10.1016/j.neuron.2012.10.038

Boesch, C. (2005). Joint cooperative hunting among wild chimpanzees: Taking natural observations seriously. *Behavioral and Brain Sciences*, *28*(05), 692–693. http://doi.org/10.1017/S0140525X05230121

Brown, E. C., & Brüne, M. (2012). The role of prediction in social neuroscience. *Frontiers in Human Neuroscience*, *6*(May), 1–19. http://doi.org/10.3389/fnhum.2012.00147

Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., … Nutt, D. (2014). The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in Human Neuroscience*, *8*(February), 20. http://doi.org/10.3389/fnhum.2014.00020

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181–204. http://doi.org/10.1017/S0140525X12000477

Constantino, J. N. (2001). A General Theory of Love. *American Journal of Psychiatry*, *158*(12), 2107–2107. http://doi.org/10.1176/appi.ajp.158.12.2107

Cook, N. D., Carvalho, G. B., & Damasio, A. (2014). From membrane excitability to metazoan psychology. *Trends in Neurosciences*, *37*(12), 698–705. http://doi.org/10.1016/j.tins.2014.07.011

Corlett, P. R., Frith, C. D., & Fletcher, P. C. (2009). From drugs to deprivation: A Bayesian framework for understanding models of psychosis. *Psychopharmacology*, *206*(4), 515–530. http://doi.org/10.1007/s00213-009-1561-0

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138. http://doi.org/10.1038/nrn2787

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1521), 1211–1221. http://doi.org/10.1098/rstb.2008.0300

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3*(MAY), 1–7. http://doi.org/10.3389/fpsyg.2012.00130

Frith, C. D., & Frith, U. (2012). Mechanisms of Social Cognition. *Annual Review of Psychology*, *63*(1), 287–313. http://doi.org/10.1146/annurev-psych-120710-100449

Gallagher, S. (2005). How the Body Shapes the Mind. *Leonardo*, *20*(July 2015), 284. http://doi.org/10.1093/0199271941.001.0001

Gerrans, P. (2007). Mechanisms of madness: Evolutionary psychiatry without evolutionary psychology. *Biology and Philosophy*, *22*(1), 35–56. http://doi.org/10.1007/s10539-006-9025-y

Grassian, S., & Hill, C. (1993). Psychiatric Effects of Solitary Confinemen, *22*(617).

Gray, P. (2009). Play as a Foundation for Hunter- Gatherer Social Existence. *American Journal of Play*, (Spring), 476–522. http://doi.org/10.1300/J082v41n02\_07

Hofer, M. A. (1994). Hidden regulators in attachment, separation, and loss. Monographs of the Society for Research in Child Development, 59(2/3), 192-207.

Jaynes, J. (1976). The Origin of Consciousness in the Breakdown of the Bicameral Mind. Houghton Mifflin

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing*, *8*(3), 159–166. http://doi.org/10.1007/s10339-007-0170-2

Limanowski, J., & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in Human Neuroscience*, *7*(September), 547. http://doi.org/10.3389/fnhum.2013.00547

Moll, H., & Tomasello, M. (2007). Cooperation and human cognition: the Vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1480), 639–648. http://doi.org/10.1098/rstb.2006.2000

Nezlek, John B., Hampton, Christianne P, Shean, Glenn D. (2000). Clinical depression and day-to-day social interaction in a community sample. Journal of Abnormal Psychology,109(1), 11-19. doi:10.1037/0021-843X.109.1.11

Noy, L., Dekel, E., & Alon, U. (2011). The mirror game as a paradigm for studying the dynamics of two people improvising motion together. *Proceedings of the National Academy of Sciences*, *108*(52), 20947–20952. http://doi.org/10.1073/pnas.1108155108

Richardson, T. (1979.). *Impro*: Improvisation and the Theatre. Faber and Faber

Rosenthal, R, and L. Jacobsen. (1968) . Pygmalion in the classroom: teacher expectation and pupils’ intellectual development. New York: Holt, Rinehart and Winston,.

Ryan C, Jetha C. (2010). Sex at dawn: how we mate, how we stray, and what it means for modern relationships. New York:Harper Perennial.

Striano, T., Vaish, A., & Benigno, J. P. (2006). The meaning of infants’ looks: Information seeking and comfort seeking? *British Journal of Developmental Psychology*, *24*(3), 615–630. http://doi.org/10.1348/026151005x67566

Synthesis, T. N. (2011). References and Notes 1., (November), 697–701.

Tauber, Robert T. (1998) Good or Bad , What Teachers Expect from Students They Generally Get ! ERIC Digest ., 1–7.