

Research Statement

Stephen Petersen

steve@stephenpetersen.net

November 13, 2006

My research centers on specifying the nature of *good thinking* in a way intended to please philosophers and cognitive scientists alike. What is it for a physical creature—biological or artificial—to think better? This question immediately involves issues in my three areas of specialization: normative epistemology, philosophy of mind, and cognitive science.

An illustrative case is my dissertation, which developed a computationally amenable approach to internal epistemology. I first argued that the guiding principle behind “internal” epistemic norms is that of epistemic self-improvement. I then showed why this role in natural creatures must take the form of a function to learn new cognitive dispositions. The epistemic norm that results has a natural computational model that is well-suited for implementation in artificial neural networks. (This builds on work by Paul Thagard on coherence, such as Thagard (2000).)

In the future, one major area of research will be to expand on this project in a few directions. For example:

- *Normative epistemology*: The functional approach to internal epistemology, especially contrasted with Alvin Plantinga’s work (see Plantinga, 1993), can shed light on the relation between internal and external epistemic norms.
- *Philosophy of mind*: My previous work suggests that functions to learn can help solve a puzzle of Ruth Garrett Millikan’s (see Millikan, 1989, p. 99) about how representations with different “directions of it” arise in higher creatures.
- *Cognitive science*: The computational model provides a natural framework for understanding propositional attitudes and emotions. (Nerb (2004), in fact, independently developed a very similar model to mine expressly for the purpose of a good model for emotions.) I have outlined this in my brief 2004 paper, but would like to expand on it.

In ways like these I am now capitalizing upon the groundwork provided by my dissertation.

My guiding focus has also motivated me to invest research time in other topics that are not directly addressed by my dissertation, but are closely-related projects with promise. For example:

- *Functions, and the teleological approach to mental content.* My current work relies at bottom on some such picture of mental content, but there is much to be done in the field, such as the problem of “vertical indeterminacy” and content.
- *Inference to the best explanation.* IBE is a powerful tool for understanding the reasoning of natural creatures. But what counts as an explanation, what makes an explanation “best”, and (of special interest to me) how and why might a natural creature calculate it? This is closely related to the next topic,
- *Formalizing simplicity.* There have been several attempts to specify simplicity and theory choice formally—Kevin Kelly’s formal learning approach, Forster and Sober’s use of Akaike’s theorem, Harman and Kulkarni’s VC-Dimensionality, and Bayesian or information-theoretic approaches such as Minimum Message Length. This seems to me an important and fruitful area for philosophers and cognitive scientists to interact.

Somewhat further afield, I also think my research has potential impact in ethics—such as for grounding an axiology in rational desires, and laying a foundation for robot ethics. I have already explored this latter in a forthcoming publication, and would like to continue such work in the future.

Naturally I do not plan to do this research alone—I look forward to the cooperation of undergraduates, philosophical colleagues, and the greater cognitive science community.

References

- Hudlicka, E. and Cañamero, L., editors (2004). *Architectures for Modeling Emotion: Cross-Disciplinary Foundations*. The American Association for Artificial Intelligence, AAAI Press.
- Millikan, R. G. (1989). Biosemantics. In *White Queen Psychology and Other Essays for Alice*, pages 83–101. MIT Press.
- Nerb, J. (2004). From desire coherence and belief coherence to emotions: A constraint satisfaction model. In Hudlicka and Cañamero (2004), pages 96–103.
- Petersen, S. (2004). Functions, creatures, learning, emotion. In Hudlicka and Cañamero (2004), pages 112–113.
- Plantinga, A. (1993). *Warrant and Proper Function*. Oxford University Press.
- Thagard, P. (2000). *Coherence in Thought and Action*. MIT Press.