Jaak Panksepp, Stephen Asma,
Glennon Curran, Rami Gabriel
& Thomas Greif

The Philosophical Implications of Affective Neuroscience

*Cognitive Science Society (CogSci10)
Portland, Oregon, 12 August 2010*

Presentations: Stephen Asma, Glennon Curran, Rami Gabriel (Columbia College, Chicago, IL)
Introduction and Discussion: Jaak Panksepp (Washington State University, Pullman, WA)

A Synopsis of Affective Neuroscience — Naturalizing the Mammalian Mind

By Jaak Panksepp

Cross-species affective neuroscience is a new approach to understanding the mammalian BrainMind.[1] To achieve a coherent vision of foundational issues, the border between human and animal consciousness is intentionally blurred, especially at the primary-process level of organization (Table 1) — namely at the subcortical level — shared

Correspondence:
Jaak Panksepp, Baily Endowed Chair for Animal Well-Being Science, Department of Veterinary & Comparative Anatomy, Pharmacology and Physiology, Washington State University, Pullman, WA. 99164-6520 USA
Email: jpanksepp@vetmed.wsu.edu

[1] I employ the terms BrainMind and MindBrain interchangeably, depending on desired emphasis, capitalized and without a space to highlight the monistic view of the brain as a unified experience-generating organ with no Cartesian dualities that have traditionally hindered scientific understanding.

Journal of Consciousness Studies, **19**, No. 3–4, 2012, pp. 6–48

homologously by all mammals. This adheres to Darwin's dictum that the differences in the mental lives of animals 'is one of degree and not of kind' (Darwin, 1872/1988, p. 127). It also respects Darwin's just-preceding ontological reflection that 'There can be no doubt that the difference between the mind of the lowest man and that of the highest animal is immense', but this applies primarily to the tertiary-process level, namely the fully formed MindBrain, after it is contextualized within ever-present cultural and developmental landscapes. Most of twentieth-century behaviourism, and now behavioural neuroscience, have been devoted to characterizing the secondary-processes of learning and memory. In contrast, human cognitive science deals mostly with tertiary-processes that are largely inaccessible in animal research.

1) Tertiary Affects and Neocortical 'Awareness' Functions
i) Cognitive Executive Functions: Thoughts & Planning (frontal cortex) ii) Emotional Ruminations & Regulations (medial frontal regions) iii) 'Free Will' (higher working memory functions — ***Intention-to-Act***)
2) Secondary-Process Affective Memories (Learning via Basal Ganglia)
i) Classical Conditioning (e.g. FEAR via basolateral & central amygdala) ii) Instrumental & Operant Conditioning (SEEKING via nucleus accumbens) iii) Behavioural & Emotional Habits (largely unconscious — dorsal striatum)
3) Primary-Process, Basic-Primordial Affective States (Sub-Neocortical)
i) Sensory Affects (exteroceptive-sensory triggered pleasurable and unpleasurable/disgusting feelings) ii) Homeostatic Affects (brain-body interoceptors: hunger, thirst, etc.) iii) Emotional Affects (emotion action systems — ***Intentions-in-Actions***)

Table 1. Brain 1) Tertiary Cognitive, 2) Secondary Learning & Memory, and 3) Primary Emotional-Affective Processing Systems.

Central to the affective neuroscientific epistemic approach is the recognition that the vertebrate BrainMind is an evolved organ, the only one in the body where evolutionary progressions remain engraved at neuroanatomical, neurochemical, and functional levels. The more ancient mental functions (e.g. primary-process emotions — ancestral genetic/affective 'memories') are lower and more medial in the brain. The higher functions (e.g. cognitive functions) are situated more rostrally and laterally. The basic learning functions are nestled in-between in various basal ganglia such as amygdala and nucleus accumbens. During encephalization this progression is functionally respected, so that the developmental and epigenetic emergence of higher cognitive mind functions always remain constrained by lower genetically-dictated affective solutions to living.

The higher brain, namely neocortex, is born largely *tabula rasa*, and all functions, including vision (see Sur and Rubinstein, 2005), are programmed into equipotential brain tissues that initially resemble Random Access Memory (RAM). But after learning how to regulate affect (e.g. from finding food when hungry to controlling impulsiveness when angry, etc.), higher-order mental processes emerge and are gradually transformed into Programmable Read Only Memories (PROMs).Thus, capacities for thoughtful reflection emerge gradually in higher brain regions developmentally and epigenetically. In this hierarchical vision of self-awareness based on primal mental processes, one progresses from '*cogito ergo sum*' (a top-down RAM inspired vision) to 'I feel therefore I am' (a bottom-up ROM inspired vision), and with experience-dependent cortical programming to 'I feel, therefore I think' (reflecting higher PROM functions). Higher emergent brain functions add more flexibility as well as regulation to the earlier 'instinctual' functions that are genetically ingrained birthrights that guide learning. At the primary-process level, affective experiences — the automatic valuative functions of mental life — rule. They intrinsically anticipate survival needs (e.g. pain facilitates survival). Higher evaluations remain constrained by such primal values that continue to be experienced directly as various affective experiences that encode biologically mandated survival issues. Of course, the affect-regulated operations of a mature tertiary-process neocortex, driven to great goals by the intrinsic enthusiasm of the SEEKING system, has great deliberative abilities, even decision-making functions that surely deserve to be called free will. It is clear that our upbringing can liberate us, in part, from the more urgent tertiary-aspects of our negative passions as better regulated positive desires can promote opportunities that would not otherwise exist.

Thus, in order to understand the whole MindBrain, one has to understand the evolutionary stratifications within the central nervous system, and to recognize how functions that emerged first i) retain a substantial degree of primacy in spontaneous behaviours, ii) govern the mechanisms of learning (e.g. the unconditioned stimuli and responses behaviourists use to control animal learning are typically affective in nature), as well as iii) motivating higher (tertiary-process) reflective decision-making processes — cognitive choices that integrate affective states within the informational complexities of the world. Between the primary 'affective-regulatory' and tertiary 'affective-cognitive' processes — the first experienced directly and the other within the guiding 'light' of reflective awareness — there is a vast territory of automatized learning processes: habituation, sensitization, classical conditioning, operant conditioning, etc. — automatic brain processes that are deeply unconscious. They arise from unexperienced and unreflective mechanical operations of the brain, that help mould instinctual emotional behaviours, imbued with primal affective values, permitting organisms to fit into environments more effectively, hand-in-glove, so to speak. In this view, the brain mechanisms of 'reinforcement' are closely linked to how affective processes control learning.

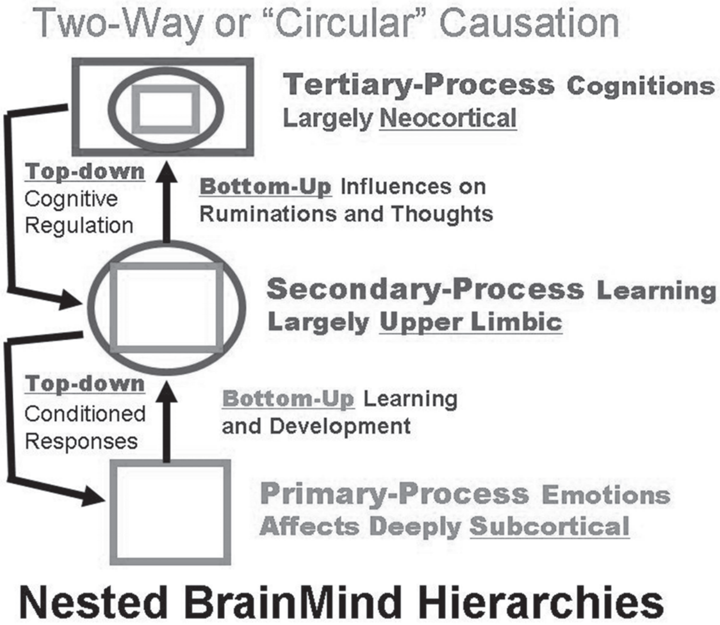

Figure 1. A conceptual summary of hierarchical bottom-up 'nested' and top-down (circular) causation that presumably operates in every primal emotional system of the brain. This schematic summarizes the hypothesis that in order for higher MindBrain functions to operate, they have to be integrated with the lower BrainMind functions, with primary-processes being depicted as squares, secondary-processing (learning and memory) as circles, and tertiary-processing (higher cognitive functions), at the top, as rectangles. Please imagine each symbol being colour-coded, to better envision the nested-hierarchies that integrate the various levels of the BrainMind (adapted from Northoff *et al.*, 2011; and Panksepp, 2011c). 'Bottom-up' control prevails in early infancy and early childhood development. Top-down control is optimized in adulthood.

These primary to tertiary gradients of mental development ultimately yield nested-hierarchies of BrainMind relationships (Figure 1), where the lower functions are re-represented within higher functions, providing multiple avenues of bottom-up and top-down relations — circular/two-way causal loops — that work as a coherent unit (for discussion, see Northoff *et al.*, 2011). For a neuroscientific understanding of mind, it is important to focus on the stratification of mental layers to better understand where to situate the various sciences of the MindBrain. The most important and most neglected foundational level, in both psychology and cognitive neuroscience, is the study of the primary experiential processes of the brain. Once we understand the affective tools for living and learning that exist at primary-process levels, we will better understand how higher mental processes operate. In short, we must illuminate the most hidden level of mental organization — the primal affective level — before we can adequately understand the rest.

Most human psychological research, including cognitive and social sciences, typically focuses on the highest levels, commonly with little recognition of the lower levels. Behavioural neuroscience tends to focus most heavily on secondary-process levels, mostly in animal models, where experiential issues are purposely neglected (and often denied), with claims of the inappropriateness of anthropomorphism and our incapacity to ever probe subjective experiences (even primal affects) empirically in animals. Only cross-species affective neuroscience explicitly acknowledges primal affective states in other animals, and seeks to understand the subcortical loci of control for the affective BrainMind. Its claims are based on abundant evidence (*vide infra*) that many of these brain functions (i.e. the primal emotional, motivational, and sensory affects) are experienced in valuative (valenced) ways — yielding a raw, unreflective affective consciousness (Panksepp, 2007). In this view, all mammals, including humans, share sets of primal affective experiences — *anoetic* tools for existence — that unconditionally guide living. This level of experience should not be called ‘awareness’ — for that would require *noetic* (knowing) and *autonoetic* (self-knowing) forms of consciousness (see Tulving, 2004; 2005; Vandekerckhove and Panksepp, 2009). Subcortical emotional networks constitute raw ‘affective experience’ — perhaps the *sine qua non* foundation for all higher forms of consciousness.

A major aim of cross-species affective neuroscience is to parse primary-process consciousness into its component networks and functions, with a special focus on emotional feelings, especially since they are of such great importance for understanding and treating psychiatric disorders (Panksepp, 2004; 2006), and potentially those social disorders that get writ large in so many individual lives and cultural fabrics. We might call them the universal 'borderline personality disorders'. Adequate evidence exists for seven primary-process emotional networks concentrated in subcortical regions of the brain — SEEKING, RAGE, FEAR, LUST, CARE, GRIEF, and PLAY — which may serve as emotional endophenotypes for psychiatric systematics (Panksepp, 2006). Identification of these systems is largely based on our ability to provoke distinct 'instinctual' behavioural actions by electrically and chemically stimulating specific regions of the brain (Panksepp, 1982; 1986; 1991; 1998a; 2005a,b; Panksepp and Biven, 2012). These primal systems are capitalized to provide a distinctive nomenclature for primal affective emotional-action networks that generate 'intentions in actions' and elaborate various distinct feelings states. Such primal states of consciousness may not be adequately captured by the vernacular emotional terms engendered by our highest, most multi-modal cortical systems that generate language, largely via learning. At present there is no direct neuroscientific evidence for a 'language instinct' in humans, albeit 'communicative urges' are reflected in homologous human and animal emotional vocalizations (Brudzynski, 2010). Human languages are coaxed into the brain, initially by the melodic intonations of motherese by which emotional communication becomes the vehicle for propositional thought (Panksepp, 2008).

Until empirically demonstrated otherwise, affective neuroscience suggests that the neocortex was not modularized by evolution but rather becomes specialized for diverse cognitive activities through developmental landscapes. This radical conclusion is based on dramatic findings such as the visual cortex of the mouse being developmentally constructed as opposed to being genetically dictated (e.g. Sur and Rubenstein, 2005). In contrast, subcortical emotional, homeostatic, and sensory affective functions, to all appearances, have been largely 'modularized' by evolution (Panksepp and Panksepp, 2000; 2001), although they are surely refined by experiences. For instance, initial modifications can occur via basic experiential-epigenetic mouldings which lead to sensitization and desensitization of certain systems; thus, laying a foundation for human temperaments. Basic learning and memory expand the cerebral repertoire to permit thinking. The genetically ingrained nature of the primal emotional systems is demonstrated by the ability of localized electrical stimulation of the brain (ESB) to provoke similar types of instinctual behaviour patterns, and accompanying feelings, in all animals across all mammalian species that have been studied. Such stimulations in the neocortex only impair functions, and generate no clear affects.

We know that these ancient subcortical emotional systems somehow help encode various positive and negative feeling states. This is because animals 'care' about these artificially induced central states: they treat such brain stimulations as 'rewards' and 'punishments' — namely, they arouse critically important (vital) unconditioned stimulus and response mechanisms that behavioural scientists have traditionally used to mould behaviour through regimented learning tasks. Animals turn such emotion-evoking brain stimulations on and off, depending on their affective valence. They return to places where they received 'rewarding' brain stimulations and avoid areas where they received 'punishing' brain stimulations, both electrically and neurochemically induced (Ikemoto, 2010; Panksepp, 1998a). Humans stimulated in these brain areas experience corresponding affects (Panksepp, 1985; Heath, 1996; Coenen *et al.*, 2011). This provides a cross-species dual-aspect monism strategy for understanding affects (Panksepp, 2005a,b), and allows us to consider the likelihood that whenever normal animals exhibit instinctual emotional behaviours, they probably have corresponding affective experiences (in ethology, they may be used as proxies for internal experiences). A study of these brain systems may eventually give us a causal neuroscientific understanding of what it means for the mammalian/human brain to experience distinct affective feelings. Such primal neural mechanisms of affect simply cannot be deciphered through human research.

The primary-process affective states do not require neocortical reflective capacities. Animals and humans deprived of their neocortical tissues at birth retain solid instinctual indices of experiential states (Merker, 2007; Panksepp *et al.*, 1994; Shewmon *et al.*, 1999). They remain conscious beings. Damage to the epicentres of these primal emotional networks can seriously compromise consciousness in every species that has been studied. This again highlights the primacy of these systems in the mental economies elaborated by secondary and tertiary processing of primal affective (i.e. rewarding and punishing) experiences. In other words, the emotional action systems of the brain are not psychologically vacuous. They participate in engendering affective states, and also providing organismic coherence that may provide a substratum for a core-SELF/animalian-soul, which has so far defied empirical analysis (Panksepp, 1998b; Panksepp and Northoff, 2009).

The underlying epistemology for pursuing cross-species affective neuroscience studies is an empirically-based conceptual triangulation among three critical and interdependent sources of scientific knowledge: i) studies of the brain that use causal manipulations (e.g. ESB), ii) the study of 'real' behaviours, namely the instinctual emotional-action patterns upon which mental life appears to be built, and iii) a psychological appreciation of both human and animal minds, that is congruent with the other two levels of analysis (Figure 2). This ontology and epistemology subsists on the novel cross-species emotional-state predictions that can be made, especially to human affective experiences (Panksepp, 1985; 1991; 2005a,b).

The broader implications of these perspectives for the human sciences and cultural perspectives were explored in this symposium — the repercussions for cognitive science, legal studies, and our conceptions of what it means to have a self. A main research goal of affective neuroscience is to flesh out implications of this kind of knowledge toward the development of a solid foundation for psychiatric science and clinical psychotherapeutic practices (Panksepp, 2004; 2006). Since such clinical issues were not covered at this CogSci10 symposium, I will close by briefly reflecting on some clinical implications.

According to the above analysis of the ancestry of mind, the traditional construct of the dynamic unconscious (see Berlin, 2011), introduced by Sigmund Freud, is not completely unconscious — it is not totally bereft of experiences (see Panksepp, 2011c). It feels like something to be in primary-process emotional states. We can now be confident that all mammals *experience* their emotions although most, just like newborn human infants, probably do not reflect on these *anoetic* experiences. They may not be cognitively *aware* that they are experiencing feelings, they simply experience such powers as guiding forces of their lives. It is worth emphasizing that although the basic mechanisms of learning may be deeply unconscious, shifting affective feelings (e.g. generating reinforcement effects) on which much of behavioural learning is based are not. However, if human experience is a useful guide, affects guide a considerable amount of thinking, ruminating, and decision-making (Northoff *et al.*, 2011). However, affective intensity can also be 'flooded-out' by cold-rational cognitive processes, via the ability of the neocortex to inhibit lower affective brain processes (Liotti and Panksepp, 2004).

It is possible that affective 'energization' of thinking (i.e. ruminations) that accompany all primary emotions in humans rapidly tend to be erased (become unconscious?) when the emotional arousal abates. In other words, the cognitive reflections during emotional arousals can rapidly descend into a dynamic form of subconsciousness, often remaining dormant, until the relevant primary affects are again

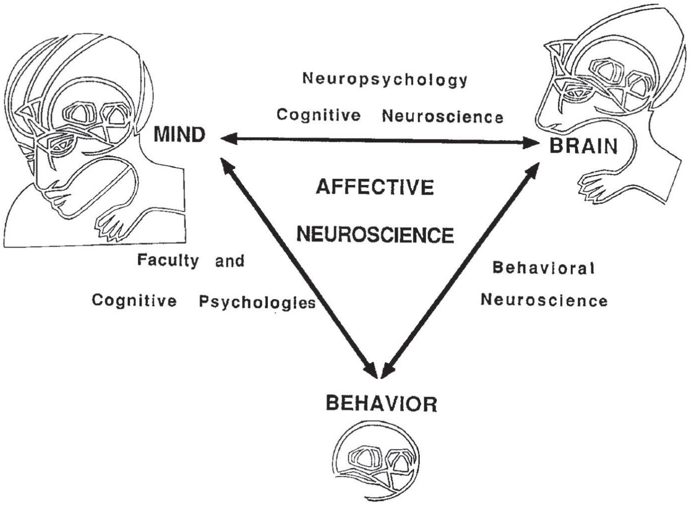

Figure 2. Progress in understanding the biological nature of affective processes can only proceed through the integration of psychological, behavioural, and neuroscientific approaches in a balanced manner (giving equal due to all). The various disciplines that bridge two of the three components are indicated. Affective neuroscience aspires to bridge all three, and the successive dissection of the logo seeks to highlight the nested hierarchies described in Figure 1: at the simplest level (bottom) we have instinctual behaviours which are primary-processes of the brain and mind (e.g. raw affects) which are evident in all other vertebrates. The complexities of learning within the brain (upper right) have been most fully addressed by behavioural neuroscience strategies. The complexities of mind (e.g. upper left) can only be addressed by self-report in humans. However, emotional affects arise from the bottom, primary-process layer of mind, and provide mechanisms which allow the middle (secondary-process, upper right) layers to mediate classical and instrumental conditioning, by yet unfathomed Laws of Affect. Currently the neurobehavioural facts strongly indicate that raw affect is built into the lowest layer of the BrainMind (as determined by brain-stimulation-induced ‘reward’ and ‘punishment’ functions), allowing instinctual emotional behaviours to be used as proxies for the presence of affective states in animals exhibiting such behaviours (a dual-aspect monism strategy), and thereby providing scientific information about homologous feelings in human minds. (Figure adapted from Panksepp, 1998a, *Affective Neuroscience*, p. 31, with permission of Oxford University Press.)

aroused. For instance, it is hard to bring the ideas that ‘naturally’ flood higher brain regions during emotional arousals back to mind, once passions have subsided. One prediction of this view is that re-evocation of primal affective states in positive therapeutic environments may rapidly help promote reconsolidation of memories and promote therapeutic change. For instance, by dealing directly with the association of ideas and ruminations that spill forth readily during real emotional arousal, lasting affective change is possible. During such evoked states, therapists can get a flavour of how raw affects are influencing tertiary cognitive processing, and thereby be better able to deal promptly and more effectively with maladaptive patterns of being. Modern memory research is clarifying that the painful edge of affectively negative memories may be dulled substantially if they are followed by intensification of positive feelings (Donovan, 2010; Nader and Einarsson, 2010), including presumably the positive affective aspects of supportive therapeutic settings. For instance, if therapists are able to evoke positive, even playful feelings, there is a reasonable chance that the subsequent memory-reconsolidation process will carry along the new positive affective contextual penumbras, thereby dulling the aches of painful memories, which are especially common in depressed people.

In sum, many of the scientific dilemmas of the twentieth century, including the Computational Theory of Mind advocated by many cognitive scientists, were created by situating all of consciousness (i.e. the capacity of have 'awareness' of experiences) just at the very top of the brain, especially the sensory-perceptual and executive regions of the brain. The instinctual-emotional action apparatus, the source of raw emotional experiences, that helps weave together a foundational form of organismic coherence (perhaps a core-SELF: Panksepp, 1998b) was provided no role in consciousness. That view prevailed, and was well-tolerated, in preference to the ever increasing empirical evidence during the second half of the twentieth century that affective consciousness (Panksepp, 2007) — perhaps the primal form of 'core-consciousness' (Panksepp, 2010) — had evolutionarily ripened into experiential states within the ancient subcortical brain networks. These foundational basic emotional and motivational urges of all mammals, which monitor vital life qualities, are the foundation of mind. If destroyed, the rest collapses (Bailey and Davis, 1942; 1943).

The tragedy of twentieth-century behaviourism — penetrating deeply into the psychological, cognitive, and social sciences — lies in a disciplinary failure to confront the deeper evolutionary psychological nature of organisms. Without an empirically justifiable vision of their affective lives, we cannot have a coherent understanding of the higher reaches of our own minds. It is now clear that modern brain imaging has seen the glimmers of basic emotions in PET and fMRI images (Damasio *et al.*, 2000; Vytal and Hamann, 2010), even though those tools, especially fMRI, are typically not well suited to visualizing the more ancient primal emotional networks coursing through upper brainstem (mesencephalic and diencephalic) regions. We should also recall that tools like fMRI detect small percentage changes of overall brain activity, with most of it remaining unseen, almost as if it were 'dark energy' (Raichle, 2010a,b; Zhang and Raichle, 2010). We will need better Hubble-type mind scopes before we can metabolically envision the more ancient recesses of brain functions that evolved much longer ago. The marginalization of affective states in the shared origins of human mental life in other organisms, when reversed by better evolutionary epistemologies (Panksepp and Panksepp, 2000; Panksepp, 2009; 2011a,b), will give us more accurate visions of our own nature.

The consequences of such a vision were explored in the following symposium: Rami Gabriel focusing on the consequences of such knowledge for cognitive science, Glennon Curran and Rami Gabriel focusing on the legal implications of the use of neuroscience data in courtroom testimony, and Stephen Asma and Thomas Greif discussing the philosophical implications for our emerging understanding of the 'core self' structures deep within the brain. They share visions of how an understanding of ancient regions of our minds may profoundly influence higher cognitive processes in humans. An appreciation of the relevance of affective neuroscience could steer the course of cognitive science toward more naturalistic visions of the foundations of human mind.

References

Bailey, P. & Davis, E.W. (1942) Effects of lesions of the periaqueductal gray matter in the cat, *Proceedings of the Society for Experimental Biology and Medicine*, **351**, pp. 305–306.

Bailey, P. & Davis, E.W. (1943) Effects of lesions of the periaqueductal gray matter on the Macaca Mulatta, *Journal of Neuropathology and Experimental Neurology*, **3**, pp. 69–72.

Berlin, H. (2011) The neural basis of the dynamic unconscious, *Neuropsychoanalysis*, in press.

Brudzynski, S.M. (ed.) (2010) *Handbook of Mammalian Vocalization*, Oxford: Academic Press.

Coenen, V.A., Schlaepfer, T.E., Maedler, B. & Panksepp, J. (2011) Cross-species affective functions of the medial forebrain bundle — implications for the treatment of affective pain and depression in humans, *Neuroscience & Biobehavioral Reviews*, 22 Dec 2010, [Epub ahead of print].

Damasio, A.R., Grabowski, T.J., Bechara, A., Damasio, H., Ponto, L.L.B., Parvizi, J. & Hichwa, R.D. (2000) Subcortical and cortical brain activity during the feeling of self-generated emotions, *Nature Neuroscience*, **3** (10), pp. 1049–1056.

Donovan, E. (2010) Propranolol use in the prevention and treatment of posttraumatic stress disorder in military veterans: Forgetting therapy revisited, *Perspectives in Biology & Medicine*, **53** (1), pp. 61–74.

Heath, R.G. (1996) *Exploring the Mind–Body Relationship*, Baton Rouge, LA: Moran Printing, Inc.

Ikemoto, S. (2010) Brain reward circuitry beyond the mesolimbic dopamine system: A neurobiological theory, *Neuroscience & Biobehavioral Reviews*, **35** (2), pp. 129–150.

Liotti, M. & Panksepp, J. (2004) On the neural nature of human emotions and implications for biological psychiatry, in Panksepp, J. (ed.) *Textbook of Biological Psychiatry*, pp. 33–74, New York: Wiley.

Merker, B. (2007) Consciousness without a cerebral cortex: A challenge for neuroscience and medicine, *Behavioral Brain Sciences*, **30**, pp. 63–81.

Nader, K. & Einarsson, E.O. (2010) Memory reconsolidation: An update, *Annals of the New York Academy of Sciences*, **1191**, pp. 27–41.

Northoff, G., Wiebking, C., Feinberg, T. & Panksepp, J. (2011) The 'resting-state hypothesis' of major depressive disorder — A translational subcortical-cortical framework for a system disorder, *Neuroscience & Biobehavioral Reviews*, 28 Dec 2010, [Epub ahead of print].

Panksepp, J. (1982) Toward a general psychobiological theory of emotions, *The Behavioral and Brain Sciences*, **5**, pp. 407–467.

Panksepp, J. (1985) Mood changes, in Vinken, P.J., Bruyn, G.W. & Klawans, H.L. (eds.) *Handbook of Clinical Neurology*, revised series, vol. 1 (45), Clinical Neuropsychology, pp. 271–285, Amsterdam: Elsevier.

Panksepp, J. (1986) The anatomy of emotions, in Plutchik, R. (ed.) *Emotion: Theory, Research and Experience, Vol. III. Biological Foundations of Emotions*, pp. 91–124, Orlando, FL: Academic Press.

Panksepp, J. (1991) Affective neuroscience: A conceptual framework for the neurobiological study of emotions, in Strongman, K. (ed.) *International Reviews of Emotion Research*, pp. 59–99, Chichester: Wiley.

Panksepp, J. (1998a) *Affective Neuroscience: The Foundations of Human and Animal Emotions*, New York: Oxford University Press.

Panksepp, J. (1998b) The periconscious substrates of consciousness: Affective states and the evolutionary origins of the SELF, *Journal of Consciousness Studies*, **5** (5–6), pp. 566–582.

Panksepp, J. (ed.) (2004) *A Textbook of Biological Psychiatry*, Hoboken, NJ: Wiley.

Panksepp, J. (2005a), Affective consciousness: Core emotional feelings in animals and humans, *Consciousness & Cognition*, **14**, pp. 19–69.

Panksepp, J. (2005b) On the embodied neural nature of core emotional affects, *Journal of Consciousness Studies*, **12** (8–10), pp. 158–184.

Panksepp, J. (2006) Emotional endophenotypes in evolutionary psychiatry, *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, **30**, pp. 774–784.

Panksepp, J. (2007) Affective consciousness, in Velmans, M. & Schneider, S. (eds.) *The Blackwell Companion to Consciousness*, pp. 114–129, Malden, MA: Blackwell.

Panksepp, J. (2008) The power of the word may reside in the power of affect, *Integrative Physiological and Behavioral Science*, **42**, pp. 47–55.

Panksepp, J. (2009) Core consciousness, in Bayne, T., Cleeremans, A. & Wilken, P. (eds.) *The Oxford Companion to Consciousness*, pp. 198–200, Oxford: Oxford University Press.

Panksepp, J. (2010) Evolutionary substrates of addiction: The neurochemistries of pleasure seeking and social bonding in the mammalian brain, in Kassel, J.D.

(ed.) *Substance Abuse and Emotion*, Washington, DC: American Psychological Association.

Panksepp, J. (2011a) The neurobiology of social loss in animals: Some keys to the puzzle of psychic pain in humans, in Jensen-Campbell, L.A. & MacDonald, G. (eds.) *Social pain: Neuropsychological and Health Implications of Loss and Exclusion*, pp. 11–51, Washington, DC: American Psychological Association.

Panksepp, J. (2011b, in press) The primary process affects in human development, happiness, and thriving, in Sheldon, K., Kashdan, T. & Steger, M. (eds.) *Designing the Future of Positive Psychology: Taking Stock and Moving Forward*, New York: Oxford University Press.

Panksepp, J. (2011c) Cross-species affective neuroscience decoding of the primal affective experiences of humans and related animals, *PLoS ONE*, **6**, p. e21236.

Panksepp, J., Normasell, L., Cox, J.F. & Siviy, S.M. (1994) Effect of neonatal decortication on the social play of juvenile rats, *Physiology & Behavior*, **56** (3), pp. 429–443.

Panksepp, J. & Panksepp, J.B. (2000) The seven sins of evolutionary psychology, *Evolution & Cognition*, **6**, pp. 108–131.

Panksepp, J. & Panksepp, J.B. (2001) A continuing critique of evolutionary psychology: Seven sins for seven sinners, plus or minus two, *Evolution & Cognition*, **7**, pp. 56–80.

Panksepp, J. & Northoff, G. (2009) The trans-species core self: The emergence of active cultural and neuro-ecological agents through self related processing within subcortical-cortical midline networks, *Consciousness & Cognition*, **18**, pp. 193–215.

Panksepp, J. & Biven, L. (2012) *Archaeology of Mind: Neuroevolutionary Origins of Human Emotions*, New York: Norton.

Raichle, M.E. (2010a) Two views of brain function, *Trends in Cognitive Sciences*, **14**, pp. 180–190.

Raichle, M.E. (2010b) The brain's dark energy, *Scientific American*, **302**, pp. 44–49.

Sur, M. & Rubenstein, J.L. (2005) Patterning and plasticity of the cerebral cortex, *Science*, **310**, pp. 805–810.

Shewmon, D.A., Holmes, G.L. & Byrne, P.A. (1999) Consciousness in congenital decorticate children: Developmental vegetative state as self-fulfilling prophecy, *Developmental Medicine & Child Neurology*, **41**, pp. 364–374.

Tulving, E. (2004) Episodic memory from mind to brain, *Review of Neurology*, **160**, pp. 9–23.

Tulving, E. (2005) Episodic memory and autonoesis: Uniquely human?, in Terrace, H.S. & Metcalfe, J. (eds.) *The Missing Link in Cognition: Self-Knowing Consciousness in Man and Animals*, pp. 3–56, New York: Oxford University Press.

Vandekerckhove, M. & Panksepp, J. (2009) The flow of anoetic to noetic and autonoetic consciousness: A vision of unknowing (anoetic) and knowing (noetic) consciousness in the remembrance of things past and imagined futures, *Consciousness & Cognition*, **18**, pp. 1018–1028.

Vytal, K. & Hamann, S. (2010) Neuroimaging support for discrete neural correlates of basic emotions: A voxel-based meta-analysis, *Journal of Cognitive Neuroscience*, **22**, pp. 1–22.

Zhang, D. & Raichle, M.E. (2010) Disease and the brain's dark energy, *Nature Reviews Neuroscience*, **6**, pp. 15–28.

Modularity in Cognitive Psychology and Affective Neuroscience

By Rami Gabriel

Abstract: *This paper explores modularity in the context of findings from affective neuroscience. I contrast cognitive science's formulations of the module and the transducer with Jaak Panksepp's notion of neuroaffective emotional-behavioral systems. The deeper theme of my paper is situating affective neuroscience as a biologically-based monist description of human nature.*

I. Introduction

The theory of evolution allows psychology to trace the development of the human mind and body naturalistically. It allows us to unite our knowledge of fundamental biological drives with the manner in which the mental aspects of these basic processes are instantiated in the brain. After the neuroscience revolution, we can comfortably say that in being the study of the mind, psychology is the *de facto* study of human 'nature' which is best biologized bottom-up, relying on neuropsychological homologies which cut across all mammalian species.

But there is yet no agreed-upon way to describe the roots of 'human nature'. The British Empiricists characterized the mind in terms of faculties derived from the association of ideas. This associationist psychology was largely replaced in the latter half of the twentieth century by cognitive science and its Computational Theory of the Mind (CTM). However, in the last thirty years, affective neuroscience (AN) has successfully wedded evolution with neuroscience through the identification of homological primary-process affective mechanisms in subcortical neural networks. This vision of the BrainMind allows us to engage in a dual-aspect monist understanding of the human form.[2] AN's BrainMind and the CTM of cognitive science require different epistemic approaches in that the sources of our knowledge of the nature of the mind will differ depending on the ontological framework we adopt — the neo-dualist CTM (functionalism) or an evolutionarily-based monist BrainMind.

In this paper, I discuss the consequences of AN for the cognitive sciences. My goal is to explore whether AN fits into the CTM and subsequently whether or not AN and the CTM are describing distinct levels of the mind in their respective characterizations. I will argue that

[2] The concept of BrainMind, explained in Panksepp's introduction, grounds the epistemic claims that will be made in this paper.

cognitive science's computational characterization of the mind as if it were solely comprised of tertiary level processes is a fundamental metaphysical misunderstanding that ought to be replaced, or at least strongly supplemented by the constraints of monistic biological frameworks.

II. The Computational Theory of the Mind and affective neuroscience

I begin with a brief sketch of the elements of the CTM and affective neuroscience that will be contrasted. Although there are many interpretations of the modular mind, the interpretation of the CTM that I take to be foundational is Jerry Fodor's (1983) description of basic perceptual processing via modules and transducers. Most psychologists agree that cognitive psychology describes the mind as a computational device that consists of three units: transducers, modules, and central processors. Transducers transform perceptual input, modules process this input, and central processors integrate the various processed signals from modules. This framework of conceptualizing the mind as a processor of external information gave way to the characterization of the module as *the* unit of mental processing in the cognitive sciences for the last thirty years.[3]

Although Panksepp described many facets of affective neuroscience in the introduction, there are a few elements of this approach worth highlighting for the purposes of my argument. AN is a description of the human mind that focuses on the basic affective mechanisms that evolved in subcortical regions of the mammalian brain. It describes an evolutionarily layered mind, based on a triune brain, consisting of our reptilian past (brainstem), our early mammalian kind (limbic lobe), and recent mammalian expansions of neocortex (MacLean, 1990). AN focuses on the subcortical networks rather than neocortical specialization in order to provide a more biological model of human, *qua* mammalian, behaviour. The evolutionary and biological bases of affective neuroscience, as opposed to the formal theoretical basis of cognitive psychology in computer science, have important consequences for the translatability of their respective findings. In other words, the historical foundations of each approach determine their respective epistemic ranges. In contrast to the CTM's modularity of mind, the main unit of AN is the basic affective mechanism, instan-

[3] Although not everyone agrees as to the definition of a module; for example see the discussion of this question in Barrett and Kurzban (2006).

tiated in real neural networks (see Panksepp, 1998, for detailed descriptions).

III. Are basic affective mechanisms modules?

In the introduction, Panksepp describes how subcortical emotional, homeostatic, and sensory affective processes have been 'modularized' by evolution. Louis Charland (1996) has written on emotions as natural kinds and the philosophical implications of AN, making a similar claim that basic affective mechanisms are modules. To explore these claims, I compare Fodor's (1983) definitive description of modules to Panksepp's basic affective mechanisms.

A number of the characteristics of Fodor's modules and Panksepp's basic affective mechanisms are similar, namely, both have narrow content, are innate, are composed of primitive evolved or unassembled processors, and are hardwired localized brain circuits. In contrast to Charland (1996) — whose paper only engaged Panksepp (1982) — I believe there are two crucial disjunctions between modules and basic affective mechanisms: basic affective mechanisms share anatomical and chemical resources whereas modules do not share computational resources, and modules are impenetrable and encapsulated whereas basic affective mechanisms are modulated by neocortical controls, bodily inputs, and diverse external factors. It is due to these major differences that it is not appropriate to describe basic affective mechanisms as modules.

These disjunctions lead me to consider whether basic affective mechanisms may be better described as transducers. Consider, the major function of basic affective mechanisms is maintaining life-supportive internal homeostasis through direct connections to action possibilities (Panksepp, 1998). This function may be prior to the cognitive computational level of modules as it manifests a physical transduction of internal and external stimuli into states that directly promote life, and hence bodily homeostasis, related to universal survival issues by means of chemical processes and simultaneous transformation of behaviour. Although affective states become represented in neural networks (which can be called modules and central processors), the central function of basic affective mechanisms, through the integration of chemical and anatomical resources, is maintenance of organismic coherence and survival in response to endogenous and exogenous stimuli; a decidedly non-computational process, except perhaps in non-linear dynamic ways (psychomotor attractor landscapes and such). In this way, basic affective mechanisms are in fact intentional in that they are about the world and represent it as a set of internal chemical values (which furthermore serve as foundational elements of the BrainMind, possessing affective phenomenal experiences, a primal form of sentience that is functional).[4]

What is notable here is that basic affective mechanisms simultaneously do too much and too little to be categorized as modules. Affective mechanisms at the subcortical level do not easily fit into the CTM because they (a) are like transducers but have the extra global function of maintaining homeostasis, (b) are intentional (via intrinsic action tendencies) without being propositional, (c) are penetrable, i.e. have a complex set of relations with newer parts of the brain, (d) are functionally not strictly computational, for example each mechanism shares neurotransmitters and brain circuits with other processes, and are thus better characterized via non-linear volumetric dynamics, (e) support and instantiate chemical states of an organism, and (f) have modulatory effects on neocortically generated processes and behaviours (in their very immediate role to facilitate survival, initially with a narrow event horizon, which is broadened by learning).

It is for these reasons that either: (i) transducer-like basic affective mechanisms are substantially more complex entities than modules conceptualized in the CTM, to the extent that they even demonstrate intrinsic intentionality (intentions-in-actions) and are penetrable all the way up and all the way down; or (ii) the CTM is the wrong kind of explanation for the subcortical functions of the mind and we need to rely rather on a biological framework to characterize it more accurately. I contend the CTM and its modular vision of the mind is not an accurate portrayal of the functioning of affective subcortical mechanisms because the CTM is essentially a psychological theory about tertiary-level mind processes which are not adequately covered by current neuroscience methodologies, and hence are envisioned as computational machinery. This approach does not pay adequate attention to the underlying neurobiological theory about the mind as consisting of evolved brain structures and functions.

IV. Why basic affective mechanisms, even as transducers, do not fit into the Computational Theory of the Mind

The first option (i) seems to transform the CTM beyond intelligibility because it takes the lowest process of the system (i.e. the transducer)

[4] Charland (1996), on the other hand, characterizes basic affective mechanisms as modules that consist of non-propositional representations that are symbolic since they can be used by modules and central processors in inferential processes.

and ennobles it with characteristics of the highest process in the system (i.e. the central processor), namely penetrability, representationality, and intentionality. This is an awful lot of power and scope for a set of unconscious chemical processes in the oldest part of the mammalian brain. In this way, AN actually inverts the traditional causal arrow since the basic affective mechanisms energize, orient, and direct the higher level processes and behaviour of the animal, towards necessary and desirable elements in the world that will help it achieve homeostasis and satisfy its most basic bodily needs and psychological drives. According to AN, modules and central processors are actually at the behest of basic affective mechanisms. Although the feedback loop between them serves to contextualize, inform, and modulate in both directions, the evolved needs of more primitive brain areas and functions are essential to the second-to-second survival of the organism in a way that neocortical processes are not (for example, see Panksepp's work on decortication of rats, reported in Panksepp, 1998). This bolsters the argument that allowing basic affective mechanisms into the transducer category, or even the module category for that matter, overly distorts the CTM and renders its characterization of the nature of the mind confused.

V. Two epistemological frameworks: The cognitive and the biological

On the other hand, the second option (ii) — that the CTM is not the appropriate language to describe subcortical affect mechanisms and that we must rather rely upon a biological epistemological framework for that layer — seems more promising for both the CTM and AN. The reason for this is that AN's BrainMind depends upon and leads to a host of biological considerations that the CTM does not. To see why, I contrast the epistemological foundations (i.e. the sources of knowledge) of the cognitive sciences with those of biology, of which AN is a branch, using Elliot Sober's (2000) summary of the main elements of the philosophical underpinnings of biology.

AN is a study of objects that are alive, whereas the cognitive sciences study objects (even non-living silicone platforms) that may have minds. AN bases its evolutionary analysis of the human mind on homology, understanding elements of the mind as evolved functions and strategies, rather than the adaptationism favoured by the cognitive sciences that emphasizes reverse engineering and logical derivation of computational circuitry. Another important contrast is between AN's characterization of basic affective mechanisms as ultimate (as well as proximate) causes versus the CTM's modules as just proximate causes of behaviour. A deeper philosophical difference between the two is that biology is an historical science whereas the cognitive sciences are nomothetic (i.e. law-like). The goal of the cognitive sciences is to construct models of the mind (i.e. weak AI) whereas the goal of AN is to reconstruct genealogical relations and isolate actual brain networks that concurrently generate adaptive behaviours and affective phenomenal experiences.

In sum, the two epistemological approaches differ in terms of their respective: logical bases, relations to evolutionary theory, ultimate goals, and area of focus. Whereas the CTM seeks 'ideal' laws of behaviour, AN seeks to penetrate what happens in neurobiological systems if certain sets of conditions apply, without specifying when/where/how often conditions are actually satisfied in Nature. Furthermore, the sources and consequences of biology are actually described in evolutionary theory (Sober, 2000), whereas the sources of the CTM are the formal language of computation.

The questions for us at this point are, does the CTM supervene on biology or do the CTM and biology *qua* AN refer to different layers of metaphysical complexity, or levels of BrainMind functioning? Alternatively, are they just different vocabularies for the same phenomena?

VI. Conclusion — affective neuroscience salvages naturalism through monism

I contend the CTM and AN essentially refer to different layers of metaphysical complexity, that cognitive psychology is in fact a special science (*cf.* the claim made by Fodor, 1974, about psychology), but at the same time it provides no clear linkages to how the human mind supervenes on neurobiology, and specifically its evolutionary origins. In contrast, these evolutionary biological origins are what is being described in AN. AN can thus be considered the meeting ground of two epistemological approaches to the human form: the psychological study of the mind and the biological. According to Panksepp,

> The core function of emotional systems is to coordinate many types of behavioral and physiological processes in the brain and body. In addition, arousals of these brain systems are accompanied by subjectively experienced feeling states that may provide efficient ways to guide and sustain behavior patterns, as well as to mediate certain types of learning. (Panksepp, 1998, p. 15)

Through the concept of the BrainMind, AN seeks to explain sentience (phenomenal consciousness) by suggesting that primary affective consciousness resides in subcortical basic affective mechanisms, the primary-level processes on which higher mental processes are built. Emotional networks are among the biological founts of affective consciousness; their function is to represent bodily homeostasis and survival via internal affective values, reflections ultimately of neurochemical homeostasis, with sufficient linkages to the world to promote intrinsic values, learning, and thinking patterns in fully developed adult minds. That is, they are intrinsically biological and intrinsically psychological. Furthermore, as discussed above, basic affective mechanisms are not adequately described via the CTM's concept of transducers or modules. Whereas neocortical cognitive processes (i.e. tertiary-level processes) may be more clearly understood via the CTM, subcortical affective mechanisms (i.e. primary-level processes) are most suitably understood via a biological epistemological framework that maintains a hold of the psychological and evolutionary causation of behaviour.

To conclude, in not fitting into the CTM, AN holds the possibility of reversing the CTM's metaphysical mistake of neo-dualism, by using the biological and evolutionary epistemological framework to span the gap of the mind–brain problem and replace computational idealism with a robust and empirically substantiated *pragmatic monism* that naturalizes the complete human form, both body and mind. This said, AN is probably working on the foundations of mind that are impenetrable to CTM approaches, and AN exhibits no tendencies for inclusive imperialism. Therefore CTM approaches may usefully try to build their structures on more realistic biological visions of organisms, where ancestral minds (primary-processes) still motivate more modern minds (tertiary-processes).

References

Barrett, H.C. & Kurzban, R. (2006) Modularity in cognition: Framing the debate, *Psychological Review*, **13** (3), pp. 628–647.

Charland, L. (1996) Feeling and representing: Computational theory and the modularity of affect, *Synthese*, **105**, pp.273–301.

Fodor, J.A. (1974) Special sciences: Or the disunity of science as a working hypothesis, *Synthese*, **28**, pp. 97–115.

Fodor, J.A. (1983) *Modularity of Mind*, Cambridge, MA: MIT Press.

MacLean, P.D. (1990) *The Triune Brain in Evolution*, New York: Plenum Press.

Panksepp J. (1982) Toward a general psychobiological theory of emotions, *Behavioral and Brain Sciences*, **5**, pp. 407–468.

Panksepp, J. (1998) *Affective Neuroscience: The Foundations of Human and Animal Emotions*, New York: Oxford University Press.

Sober, E. (2000) *Philosophy of Biology*, 2nd ed., Boulder, CO: Westview Press.

Affective Neuroscience and the Philosophy of Self

By Stephen T. Asma and Thomas Greif

Abstract: *The nature of self awareness and the origin and persistence of personal identity still loom large in contemporary philosophy of mind. Many philosophers have been wooed by the computational approach to consciousness, and they attempt to find the self amidst the phenomenon of neocortical information processing. Affective neuroscience offers another pathway to understanding the evolution and nature of self. This paper explores how affective neuroscience acts as a positive game-changer in the philosophical pursuit of self. In particular, we focus on connecting 'mammalian agency' to (a) subjective awareness, and (b) identity through time.*

I. The problem of the self

What am I? I am obviously an individual person, observable by others — a public physical organism. I can be picked out of a crowd. But to myself, I am a subject; an agent moving through the world with a rich inner life of thoughts, feelings, and memories. I am a self.

In its modern formulation, the philosophical problem of the self goes back to Descartes and David Hume. But the puzzles of self-identity are perennial (maybe even inevitable) and stretch back to the ancient Greeks and the Vedic and Upanisad literature of the Hindus. The ancient Greek playwright Epikarmos even tells a story of a man who borrows money from his neighbour, but when he is pressed to repay the loan, he reminds the courts that he is, like every other natural thing, constantly changing (what with the ceaseless exchange of matter) and can't literally be said to be the *same* guy who borrowed the money a while back.

All expedient philosophy aside, self-identity has been a longstanding puzzle. How does a subjective unity emerge out of a plurality of mental abilities? How does self-reflective awareness relate to those abilities? And how does self-identity persist (with continuity and change) over time? Jaak Panksepp's affective neuroscience brings fresh perspectives to the philosophy of self. In order to appreciate these fresh perspectives, we need to situate ourselves a bit in the modern conversation.

David Hume pointed out Descartes' error (but not the 'error' that Antonio Damasio focused on).[5] According to Hume, Descartes had no

[5] Antonio Damasio's book *Descartes' Error: Emotion, Reason and the Human Brain* (2005) famously argues that Descartes' great error was thinking of the mind as separate

right to think of the 'I' as a metaphysical substance. The *cogito ergo sum* does not establish the existence of metaphysical substance — it only proves the existence of momentary self in each act of thinking. But now Hume found himself in a new dilemma. If all ideas — all knowledge — originate in sense impressions (a basic Empiricist commitment), then what should we make of the self? My self cannot be found as a discrete content of consciousness — it is always the knower and never the known.[6] Hume concluded counterintuitively that I am really just a bundle of experiences (memories, emotions, cognitions, etc.) and the self is a kind of fiction.[7] Following Hume, Kant continued a more *functional* approach to the self, rather than a naïve metaphysical view. The self is the point of unity or focus of subjective perception, feeling, cognition — but the self must be presupposed or inferred in order to make sense of experience. The self is not a fiction, but it is also neither directly experienced (through the categories of understanding) nor directly encountered through intellectual intuition.

Many contemporary philosophers have continued this tradition. The self accompanies the content of experience with something like an 'awareness tone' — and this moment of self-awareness, this crystallization of subjectivity, is a 'thin subject' lacking 'ontic depth' (Strawson, 2009). This very rarefied high-level self is also exceedingly promiscuous. It flits about and colours whatever experience is currently underway. This translucent self is a movable awareness that

from the body (and therefore, the emotions). Descartes' dualism is more complicated, however, and Damasio's critique takes a somewhat uncharitable view of Cartesian mind. It is true that mind and body are metaphysically distinct, according to Descartes, but he never viewed mind as a purely rational calculator detached from emotional life. His point was that bodily affects are not a part of the subjective life until they can be read-out (as emotions or feelings) by the conscious mind. Descartes' contemporaries and next-generation philosophers like Hume, however, saw a different 'error'. For Hume and Kant the mistake was thinking that a conscious unity of experience (the *cogito*) proves the existence of a corresponding entity — an ontological self. No positive metaphysics can be derived legitimately from the *cogito*.

[6] Hume says: 'For my part, when I enter most intimately into what I call myself, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never can catch myself at any time without a perception, and never can observe anything but the perception. When my perceptions are removed for any time, as by sound sleep, so long am I insensible of myself, and may truly be said to not exist' (Treatise, I, iv, 6; Hume, 2009).

[7] Interestingly, the Buddha makes similar arguments in the *Potthapada sutta* (DN) against the metaphysical notions of *atman* and also against the notion of a separable consciousness — a *res cogitans*. In the *Mahatanhasankhaya sutta* (MN), he likens consciousness to fire, and fire exists only on the fuel it burns — never in some pure disembodied form.

emerges in different functional modes, but has no personality *per se*.[8] Where is my real self, for example, when I'm struggling with a Boolean algebra problem? In this case the self seems to 'reside' in the higher neocortical activities of mathematical thinking, but if you suddenly poke me with a pointed stick, then my self will quickly shift to the material body domain. Each new activity — indeed each new moment — brings a new self. If there is such a diaphanous self, then not much can be said about it at this point. One wonders, however, whether we may one day marry the phenomenological self-report of the self-aware subject with sophisticated brain imaging in a way that reveals some unique recursive neural reverberation. We may one day find some neural flash that serves as the material substrate for our familiar sense of translucent subjectivity. This subjectivity is probably an emergent property of various neurochemical systems, some of which reach way down into the limbic and possibly subcortical levels.

Below this arid domain of the philosopher's translucent self, however, lies the realm of self that most laypeople contemplate. Here is the self of common sense. A self that has personality — built up over time with beliefs, memories, and life history. William James and pragmatists like George Herbert Mead reminded philosophers that subjectivity is not utterly pure, but mixed and integrated with social life.[9]

Philosopher Daniel Dennett describes this more content-rich self as our 'center of narrative gravity' (Dennett, 1988). Antonio Damasio calls this our 'autobiographical self' (Damasio, 2000). And as these names suggest, this self is largely composed in the highly discursive process of neocortical reflection. Hubert Herman's psychological theory of the dialogical self draws heavily on this tradition (Herman and Kempen, 1993).

[8] My use of the term 'translucent self' is perhaps idiosyncratic, but it responds to a contemporary discussion in the phenomenology of self. Phenomenologists like Thomas Metzinger (2003) and Dan Zahavi (2005) have developed the terminology of 'transparency' and 'opacity' of the self in rather precise ways. Metzinger, for example, describes the phenomenologically transparent self as a way of describing a pre-reflective state of naïve experience (being in a world), wherein the representational (and perhaps agency) aspects are *invisible* to the subject (i.e. the self is a transparent 'window' through which the subject sees the contents of experience, and only the contents are attended to). The phenomenologically opaque self is when I am aware of my own representational processing — I attend to myself as the 'vehicle' or the 'framer' of the content of experience, as in the case of pseudo-hallucinations or lucid dreaming. My own use of the term 'translucent self' is partly to acknowledge this interesting discussion in the literature, but also to reject the dichotomous tendency of this form/content distinction. One of the implications of our 'mammalian agency' approach to the self may be that subjectivity is never purely opaque nor transparent, but somewhere in between.

[9] William James offers a compelling integration of self theories in Chapter X 'The Consciousness of Self' (James, 2007).

Language, together with frontal-lobe powers, allows us myriad ways to represent the world and represent ourselves. We make ourselves, at this level, through the stories we tell ourselves. Many of those representational processes (that govern our self identity) will be constrained by those rules of cognition that computational cognitive science seeks to isolate. And all the relativism notwithstanding, the social constructionists have also recently helped us to better appreciate the role that society can play in this narration of self identity. But while all of this is fascinating and while good work will continue at this level, Dr Panksepp's revolutionary work wants to take us lower still — into the ancient, unexplored but powerful sources of self.

II. Mammalian agency

In contrast to the neocortical, highly linguistic aspects of mind, Dr Panksepp goes down to the foundations of mammal agency. In doing so, he develops a more capacious concept of consciousness — one that includes emotions and their primitive affects — and expands our notion of mind beyond the representational and propositional versions that dominate both cognitive science and traditional philosophy. Affective neuroscience reminds us of the *body* and its *non-linguistic* forms of *meaning*. Dr Panksepp revises Descartes' *cogito*, claiming instead 'I feel, therefore I am' (Panksepp, 1998). But even deeper than this limbic consciousness he pursues the primitive SELF (Simple Ego-type Life Form) in the pre-linguistic motor, emotional-action mapping system of the ancient midbrain.

Panksepp's archaic self is a biological notion of identity. It is a concept of self based more on affectively rich *action* than rarefied intellectual *reflection*, and so it includes many other kinds of non-human animals in the club of selves. An organism trying to evade a predator, within a specific environment, is solving a multitude of challenges in real time. It does so from a specific point of view in space and time — constantly adjusting its body and modulating behaviours. A rabbit trying to evade a predator, to use Dr Panksepp's example, has little conscious sense of its own future and past (given the reality of its modest frontal lobes) but 'It is dealing with its present circumstances on a moment to moment basis. It is precisely those here-and-now states of consciousness that we must seek to understand before we can grasp how they come to be extended in time, as they are in the human mind through our frontal cortical time-extending and planning abilities' (Panksepp, 1998).

Affective neuroscience reminds us of our phylogenetic homologies with other mammals, and so our biological identity should be found near the core of the brain — not the more recent neocortex. This archaic SELF would be a basic motor-mapping system — a template for action tendencies. Despite the inclination of philosophers to think about consciousness and subjectivity in terms of perceptions (like sense data qualia), affective neuroscience reminds us that 'a level of motor coherence had to exist before there would be utility for sensory guidance' (*ibid.*). This archaic SELF would have to coordinate or integrate emotions from the periaqueductal gray (PAG) region of the brain and the primal perceptual visual, auditory, and somatosensory systems of the midbrain. The centromedial zones of the upper brainstem (especially the deep layers of the colliculi and the PAG) answer to this requirement. Moreover, Panksepp's experimental work with mammals suggests that this area is much more relevant to biological intentional identity than higher neocortical areas. Experimentally induced lesions along the PAG are much more devastating to the intentionality or seeming *agency* of the animal than lesions in the higher areas of the brain. This archaic level of self is not cognitive. It is what Dr Panksepp calls 'primary process consciousness' and it resides in the intrinsic action-readiness of the biological system.

Beyond the simple integrated motor actions of this SELF, it is also likely that this centromedial zone provides a 'coherent matrix in which a variety of sensory stimuli become hedonically valenced' (*ibid.*). In other words, the organism is establishing attraction and aversion values at the subcortical level, and so the organism's most rudimentary self-awareness, of a spatio-temporally located body in an environment, will already be coded with positive and negative affects.[10] The self is not superadded after a certain level of cognitive sophistication is achieved (a view commonly held by philosophers). Rather, the self first emerges in the pre-cognitive ability of most organisms to operate from an egocentric point of view. Way below the level of propositional beliefs, animals must solve basic motor challenges (e.g. where am I in relation to that advancing sharp claw thing? Am I moving now, or is the environment moving? Am I eating my own arm?). For mammals this low-level ability is accompanied by the archetypical survival systems, shaped by natural selection over

[10] Panksepp (2005) suggests 'there exists a subcortical viscero-somatic homunculus, laid out in motor-action coordinates, that creates a primal representation of the body (core SELF) that can be modulated by global brain emotional networks that establish affective intentions in action, which are projected onto the world as prototypical affective values, helping guide cognitive intentionality'.

geological time. These are the homological affective systems that Panksepp has isolated in the brains and behaviours of his test subjects: approach when SEEKING, escape from FEAR, attack in RAGE mode, pursue nurturance in PANIC, seek mate in LUST mode, and so on. These affects and emotions are survival skills and comprise and pervade primary and secondary consciousness — they have to be 'owned' by the organism for them to work properly. This is why Panksepp and Damasio, both fans of Spinoza's monism, are in agreement about the reality of primary or core consciousness.[11] Subjectivity resides first in the biological realm of action. It is not the disembodied Cartesian spectator.

III. Philosophical implications

Now what are some of the implications of this notion of self? First, like other forms of scientific naturalism, it demonstrates that we do not need additional metaphysical agencies (like souls, or noumenal mental realms, etc.) in order to explain personal identity or even subjectivity. Secondly, and more significantly, Panksepp's archaic self — with its primary consciousness — rescues the body and feelings from the long philosophical tradition that characterized them as purely unconscious machinery. We all know of Descartes' dualism-derived 'animal machines', but even David Chalmers (1997) seems to think that fully functioning animals with intact brains and bodies could be zombies — 'all is dark inside' with nobody home. Panksepp's approach suggests that consciousness is not superadded to otherwise functioning survival machines, nor can consciousness be abstracted out of the physio-chemical system (except I suppose in parlour-game thought experiments). Even Daniel Dennett, who is usually quite sensitive to the biological sciences, offers an example that betrays a cognitive bias about consciousness. He asks, 'What is it like to notice, while sound asleep, that your left arm has become twisted into a position in which it is putting undue strain on your left shoulder? Like nothing: it is not part of your experience. You swiftly and unconsciously shift to a more "comfortable" position…' (Dennett, 1996). And Dennett concludes that whatever 'clever' problem-solving is going on at these biological levels, it is not a part of our mental lives at all. Of

[11] To understand what it means to have raw affective feelings, Dr Panksepp suggests that 'we must entertain neuro-psychological conceptions of human and animal "souls" through concepts such as the "core self" (Damasio, 1999; Gallagher & Sheard, 1999; Panksepp, 1998a). I suspect our mental lives are critically linked to primal viscero-somatic representations of the body situated in paramedian regions of the brain, and connected to associated higher limbic areas…' (Panksepp, 1998).

course, Dennett and others have a point here. Many of our brain-based competencies (like the autonomic systems) happen below the radar — but Panksepp's approach offers the tantalizing possibility that we can get into the muddy so-called dynamic unconscious. In fact, with his notion of primary subcortical consciousness, he seems to be changing the game and eliminating the traditional notion of an unconscious.[12] After all, in Dennett's own example of the sleeping subject, I do not move my arm in *any* chaotic manner — I don't fling it into my face, or put it into a less comfortable position. My primal self solves the problem with a somewhat nuanced sense of the spatio-temporal environment and the relevant motor possibilities. We can say, as Dennett does, that this has no connection with our mental lives at all, but this only betrays an overly narrow conception of mind (e.g. neocortical computation, or what Panksepp calls tertiary consciousness). We may not have much first-person phenomenological data of this archaic self or this mind, but significant access can be gained by the kinds of experimental brain manipulations that guide affective neuroscience research.

One of the most interesting implications of this biological notion of self-identity is that it answers some of the traditional scepticism about the self. From the Buddha's criticisms of *atman*, through Hume's bundle theory, and up to today's postmodern rejection of an essential core, these sceptical traditions have adopted the decentred subject. But, if Panksepp is right, then the fracturing of the subject is overestimated, and the embrace of a decentred self is premature. Yes, the diaphanous self is momentary and cannot be directly observed inside experience, but the so-called 'binding problem' of apperception may be more imaginary than real. Panksepp's SELF gives us a way in which the fleeting and ontically thin 'I' keeps getting referred back to the biological 'I'. The centromedial zones of the mind-brain produce a primitive self that *persists over time*, because it is a 'central processor' of immediate survival inputs and outputs for an organism that *is* extended in space and time. The fleeting I of the *cogito* may be reborn during every change in emotional or perceptual or cognitive content, but much of the content of our experience (perhaps all of it) will first be organized by the centromedial zones of the midbrain and the core

[12] 'The traditional answer has been that one does not have any mental experiences until certain kinds of information interact with — are "read out" by — higher neocortical mechanisms that elaborate our awareness of the world. Many still believe that affects are not experienced in the lower reaches of the brain — that all brain functions below the neocortex are experientially implicit and unconscious. Within such anthropocentric worldviews, emotional feelings cannot be understood until we figure out how the higher regions of the brain generate awareness of the world' (Panksepp, 2005).

affect systems. So, one of the implications of Panksepp's work is showing how the higher rarefied subjectivities of self may find constant tether to our very specific animal identity. Hume, who didn't have the benefit of living after Darwin, went looking for the self in the wrong part of the psyche — namely in the representational mind.[13] Strangely enough, contemporary philosophers are still looking in the wrong, albeit well-lit place. Panksepp, however, is finally delivering on the Darwinian promissory note: a subject, an 'I', that is truly born out of the struggle for survival. The binding problem is not a problem because subjectivity is always content-laden with the unified life of the spatio-temporally located organism and its evolved archetypical dispositions. In this way, I think Panksepp's and Damasio's solutions are somewhat similar, though Panksepp seems more explicit in locating such identity in the ancient brain.

Damasio's most recent book, *Self Comes to Mind*, attempts to clarify the similarities and differences with Panksepp's long-held theory of a primary self. Circumspect about his own work, Damasio explains that his previous accounts of the self were focused too high up in brain processing, and now he recognizes a brainstem based 'primordial' or 'proto self'. This proto self, according to Damasio, corresponds more with Panksepp's primary SELF, but Damasio wants to locate it even lower down the brainstem (nucleus tractus solitarius) than Panksepp suggested (periaqueductal gray). 'In the perspective of evolution,' Damasio says, 'and in the perspective of one's life history, the knower came in steps: the protoself and its primordial feelings; the action-driven core self; and finally the autobiographical self, which incorporates social and spiritual dimensions' (Damasio, 2010, p. 10).

Panksepp and now increasingly Damasio want to locate a fundamental self deep in the real-time processing of mammal brains, but what remains contentious and empirically unverified is whether that locus is grounded more in the motor structures (Panksepp) or in the sensory structures (Damasio).[14] It's hard to see how this disagreement will be resolved. Brainstem processes produce felt body states in the organism, and these primitive sensations of pain and pleasure are intimately integrated with the action-orientation of the motor systems.

[13] The irony is that Hume uniquely grasped the overwhelmingly passional/emotional nature of human beings — demoting reason down to lowly 'rationalizer' rather than imperious controller. But then, prior to Darwin, Hume had no real way to connect (logically, let alone chronologically) the limbic life with the rational. Subsequently, the affects slowly submerged into a swamp of philosophical incognita.

[14] See footnote 17 of Chapter One (2010) for Damasio's clearest articulation of his difference with Panksepp regarding self.

The points of agreement between Damasio and Panksepp are many, however, and perhaps the most important is the way both affective scientists marshal impressive data to demonstrate that the self is not just a product of the cortex. Panksepp's work with de-cortication of rats is well known (Panksepp, 1994), but in *Self Comes to Mind* Damasio strengthens the argument significantly by taking us into the rich emotional life (grounded in the proto self) of children born without a functioning cerebral cortex. Damasio argues that these children demonstrate low-level agency and basic levels of emotional integration.

Damasio's brand of affective science, which also tries to get all the way up into the higher levels of cognitive life, may have more to offer philosophers who are interested in the uniquely complex subjectivity of human mind, perhaps to an extreme extent where affectively experienced life, as William James speculated, is concentrated in the neocortex. Self-identity over time is woven together, according to Damasio, in the 'autobiographical self' which at first sounds like the discursive representational narrations of higher neocortical processes.[15] There's no doubt that big-brained *Homo sapiens* can spin elaborate coherence out of disparate experience, using memory, discursive rationality, and intentional projections. But combinations of non-linguistic perceptions, like visually based image schemas, together with engraved feeling dispositions may be all that is necessary to begin some rudimentary *autobiography* of self. Animals with very impoverished symbolic and conceptual skills may nonetheless have the ability to sense (literally) their own personal history and then comport themselves into the near future (again, drawing on their affective entrenchments, rather than cognitive reflections).[16] Nonetheless, many philosophers are more captivated with the truly symbolic manipulations of

[15] In the first chapter of Damasio's *The Feeling of What Happens* (2000) he seems to suggest that the promiscuous self (the ontically thin 'I') is something that accompanies the real-time here-and-now experiences of many non-human animals. But these core-self subjects must be woven together into a coherent record of the organism's life history. For humans this weaving will be heavily cognitive, volitional, and reflective, but for other mammals it will be more deterministic neurological engraving. Damasio recognizes, in Chapter Six, that autobiographical composition of the self can be non-linguistic and image based. His latest work further strengthens the idea that certain kinds of self are pre-linguistic.

[16] It's my view that this might be an interesting meeting place between affective neuroscience and the metaphor-based epistemology of philosophers like Mark Johnson (2007) and George Lakoff (Lakoff and Johnson, 1980). Representational cognition is obviously very sophisticated when compared with sensual problem solving in lower animals, but I doubt that it emerged as a *sui generis*. The progenitor of propositional conceptual knowledge must be bodily knowledge, which in turn must be more image-based, affect-based, and spatially, temporally relative to our particular evolution. The metaphorical root (what Johnson calls 'aesthetic' and I would call 'affective') of cognition is just one more reason why the computational model of mind is unsatisfactory.

the human autobiographical self. For these philosophers, Damasio will be more intriguing than Panksepp.

Ultimately, sceptics about the self have been right to scoff at the idea of a mysterious transcendental homunculus that sits like a spectator in a Cartesian theatre. But they were wrong to dispense with agency. It is often said of bundle theorists, whether the Buddha or David Hume, that they want to characterize *thinking* without a *thinker* — or they want to get the thoughts to *think themselves*. These are laudable moves as philosophers try to account for the invisibility of the self, but perhaps these counterintuitive moves are the unfortunate product of doing one's philosophy in the neocortical paradigm of representation and perception. Go lower into the biological agency of affective consciousness and the idea of a self that collects, unifies, and weights content makes more sense (even while it remains largely invisible to tertiary-consciousness). However, if one recognizes how profoundly the primal euphoric SEEKING urges of the brain, foundational for all addictions — from drugs and sex to rock-and-roll — influence higher mental activities, even tertiary-consciousness may come to recognize the ancestral affective powers of the subcortical mind, as it energizes and brings joy and anticipatory fervour (and at times an addictive urgency) to many of the daily activities of life.

References

Buddha 'Potthapada sutta' (1995) *The Long Discourses of the Buddha: A Translation of the Digha Nikaya*, 2nd ed., Walshe, M. (trans.), Somerville, MA: Wisdom Publications.

Buddha 'Mahatanhasankhaya sutta' (1995) *The Middle Length Discourses of the Buddha: A Translation of the Majjhima Nikaya*, 2nd ed., Bhikkhu Bodhi, Bhikkhu Nanamoli (trans.), Somerville, MA: Wisdom Publications.

Chalmers, D. (1997) *The Conscious Mind: In Search of a Fundamental Theory*, Oxford: Oxford University Press.

Damasio, A. (2000) *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*, Orlando, FL: Mariner Books.

Damasio, A. (2005) *Descartes' Error: Emotion, Reason and the Human Brain*, London: Penguin.

Damasio, A. (2010) *Self Comes to Mind: Constructing the Conscious Brain*, New York: Pantheon Books.

Dennett, D. (1988) Why we are all novelists, *Times Literary Supplement*, 16–22 September.

Dennett, D. (1996) *Kinds of Minds*, New York: Basic Books.

Herman, H. & Kempen, H. (1993) *The Dialogical Self: Meaning as Movement*, Waltham, MA: Academic Press.

Hume, D. (2009) *A Treatise of Human Nature*, London: Merchant Books.

James, W. (2007) *The Principles of Psychology, Vol. I*, New York: Cosimo Classics.

Johnson, M. (2007) *The Meaning of the Body*, Chicago, IL: University of Chicago Press.

Lakoff, G. & Johnson, M. (1980) *Metaphors We Live By*, Chicago, IL: University of Chicago Press.

Metzinger, T. (2003) Phenomenal transparency and cognitive self-reference, *Phenomenology and the Cognitive Sciences*, **2**, pp. 353–393.

Panksepp, J. (1994) Effects of neonatal decortication on the social play of juvenile rats, *Physiology and Behavior*, **56** (3).

Panksepp, J. (1998) *Affective Neuroscience: The Foundations of Human and Animal Emotions*, Oxford: Oxford University Press.

Panksepp, J. (2005) On the embodied neural nature of core affects, *Journal of Consciousness Studies*, **12** (8–10), pp. 158–184.

Strawson, G. (2009) *Selves: An Essay in Revisionary Metaphysics*, Oxford: Oxford University Press.

Zahavi, D. (2005) *Subjectivity and Selfhood*, Cambridge, MA: MIT Press.

Affective Neuroscience and Law

By Glennon Curran and Rami Gabriel

Abstract: *The paradigm of cognitive neuroscience is currently being used to evaluate the legal relevance of neuroscience, and in particular brain imaging, in courts of law. However, an affective neuroscience perspective would suggest that the conclusions being drawn by applications of cognitive neuroscience to law are oversimplified. Panksepp's descriptions of affective subcortical processes that can motivate behaviour without deliberation call into question the evidentiary value of localized neocortical mechanisms that are increasingly thought to play foundational roles in the actions and mental states at issue in a criminal charge.*

I. Introduction

John Locke's influence on legal philosophy is apparent in the view that reason, above all, characterizes human nature. In the context of crime, a transgression against the law is often understood as an actor's willing departure from reason. This notion is embodied in the common law maxim '*actus non facit reum nisi mens sit rea*', meaning 'an act does not make one guilty unless the mind is guilty'. Sir William Blackstone stated that, 'to constitute a crime against human laws, there must be, first, a vicious will' (Blackstone, 1765–69). Legal scholar Roscoe Pound elaborated on these concepts by explaining that a criminal justice system that punishes the vicious will 'postulates a free agent confronted with a choice between doing right and doing wrong and choosing freely to do wrong' (Pound, 1927). But the mind and the will have long been murky concepts in the creation of a legal philosophy that connects the criminal culpability of an actor to the act committed.

Over time the courts were forced to expand the notion of mental culpability beyond a rigid rationalist understanding of choice to account for circumstances when an actor's diminished capacity limits his choices. The legal purpose for requiring mental culpability during commission of a crime has also changed with the evolving philosophical underpinnings of the criminal justice system. Traditionally, it was important to identify a culpable mental state because the objective of the criminal justice system was to punish evil-doing (Bonnie *et al.*, 2004, p. 171, quoting Sayre, 1932). However, as Francis Bowes Sayre has argued, 'the mental element requisite for criminality… is coming to mean, not so much a mind bent on evil-doing as an intent to do that which unduly endangers social or public interests' (*ibid.*). Martin Gardner has documented and pointed out that the historical confusion surrounding the purpose of culpable mental states lies in conflicting philosophical traditions that underpin the criminal justice system: retributivism (punishing the deviant actor) and utilitarianism (achieving the greatest common good by ridding society of those who endanger the public interest, but also rehabilitating those who might re-enter society) (see Gardner, 1993).

This history of criminal responsibility includes numerous terms to denote mental culpability, as well as philosophical lenses through which these terms may be viewed. This history is inconsistent, confusing, and at times, downright arbitrary. Terms like 'mind', 'mental state', and 'will' often defy definition and are difficult to apply. The conflicting philosophies of retributivism and utilitarianism make the task of interpreting and establishing the reasoning behind the use of these concepts even more inconsistent. However, the psychological sciences have recently become a tool to clarify the uncertainty surrounding the issue of criminal responsibility for the purposes of legal thinking. The mind, the will, and all of the shadowy implications contained therein, are being illuminated by modern science. By way of example, some scholars as early as the 1920s used the findings of behaviourism to re-evaluate the role of mental states in a criminal charge (Malan, 1922). Clearly the particular application of behaviourism to law never caught on. But, this is an early example of the relationship between trends in the sciences and trends in legal understandings of criminal responsibility. The shift from behaviourism to neuroscience now signals a deeper scientific understanding of the mind, with renewed potential to aid our understanding of criminal responsibility. More specifically, in the past thirty years, neuroscience itself is experiencing a shift towards the greater need to understand the affective and social foundations of the human mind. This affective turn, which shifts our attention to the deep structures of the human brain, forces us to come to terms with the brain's numerous integrated levels when trying to identify the causal factors of human action. A greater understanding of the deep affective causes of human action may further refine our understanding of criminal responsibility.

This paper will provide an example of how the affective turn can help shed new light on criminal responsibility, as well as the theoretical underpinnings of our criminal justice system. Affective neuroscience clarifies the oversimplification of criminal responsibility demonstrated by the premature applications of neuroscience to law by, for instance, cognitive neuroscience — and as a result, lead us to the conclusion that law and neuroscience are not yet capable of full integration. Furthermore, affective neuroscience shows us the way forward in understanding legally relevant aspects and sources of human action because it focuses well below the surface of the brain to evolutionarily older regions, namely midbrain and limbic structures. This paper will conclude with some musings on how findings from affective neuroscience might inform the philosophical underpinnings of our criminal justice system, in terms of retributivism and utilitarianism.

II. Connecting neuroscience and law through 'diminished capacity'

The relation between neuroscience and the law is a burgeoning area of research because neurological evidence is slowly but surely being utilized by lawyers and offered to courts in a variety of contexts. In many cases, judges, juries, and lawyers are not wholly competent to assess the credibility of evidence derived from neuroscience. Stanford Law Professor, Hank Greely, stated the problem succinctly:

> Neuroscientists have been conducting path breaking research using neuroimaging technology… but there are a lot of open questions about how the findings will be applied in the context of existing law and no guide posts for judges and juries who will have to weigh this complicated neuroscientific evidence when making decisions about guilt, innocence, or liability.[17]

Greely is part of 'The Law and Neuroscience Project', which received a three-year $10 million grant from the MacArthur Foundation — an

[17] Reported in an article entitled 'Stanford Law School to Advance "Neurolaw" as Part of $10 Million Grant' by the Stanford Law School News Center on 31 October 2007.

organization whose honorary chair is former United States Supreme Court Justice Sandra Day O'Connor. Based out of the University of California, Santa Barbara, 'The Law and Neuroscience Project' consists of a collection of scientists, philosophers, and legal experts conducting research that proposes to serve as, in Greely's words, 'guide posts' for legal professionals in regards to the use of neurological evidence at trial.

A goal of this project is to explore whether neuroscience may be harnessed to identify the causal mechanisms of actions and mental states elemental to criminal charges. Analyses of such mechanisms may be legally relevant to a cause of action under the law of evidence. Very recently, trial courts at the state and federal level have ruled on attempts to admit evidence like brain electrical activity mapping, brain fingerprinting, and brain electrical oscillation imaging (Belcher and Sinnott-Armstrong, 2010). These cases show that neuroscience has been most readily applicable in the context of evidence offered by a defendant in a criminal prosecution to reduce criminal responsibility. As such, neuroscientific evidence is being offered as proof that a defendant could not have had the requisite mental state during commission of a crime.

The law generally understands a crime as a compound expression of behaviour and mental volition meeting the elements of a specific statutory or common law definition. The majority of criminal charges contain two important components known as *actus reus* and *mens rea*. *Actus reus* refers to a particular action that law and society deem culpable, while *mens rea* refers to a particular culpable mental state accompanying the commission of the *actus reus*. For example, the State of Illinois recognizes that 'a person commits battery if he intentionally or knowingly without legal justification and by any means (1) causes bodily harm to an individual or (2) makes physical contact of an insulting or provoking nature with an individual'.[18] A person meets the *actus reus* portion of a battery as defined in Illinois if he/she 'causes bodily harm' or 'makes physical contact of an insulting nature' with an individual. However, the person must have committed one of the proscribed acts with the requisite *mens rea* — 'intentionally or knowingly'. Typically, a criminal defendant cannot be guilty of a charge unless he/she committed a proscribed action with a proscribed mental state.[19]

[18] See the Illinois battery statute at 720 ILCS 5/12-3.

[19] There are instances where the law recognizes criminal culpability in circumstances lacking an *actus reus* in whole or in part (i.e. inchoate crimes such as attempt) as well as

A person cannot be convicted of criminal battery if he did not have the capacity to act knowingly or intentionally. An attorney may argue that a defendant had diminished capacity preventing him from forming the requisite *mens rea* due to intoxication, mental retardation, or minority of age. Legal precedent also exists for the proposition that neurological abnormalities can cause diminished capacity. For many years, law has recognized that a person undergoing a seizure who injures another person may lack critical aspects of the *mens rea* and the *actus reus* (like the voluntariness requirement).[20] As the methods of neuroscience achieve greater sophistication, the analysis of functional and causal neurological mechanisms is being used to make more far-reaching arguments concerning diminished capacity. In this paper, we compare affective neuroscience and cognitive neuroscience explanations of criminal behaviour; we focus specifically on the concept of intention as our test example.

III. Affective neuroscience and cognitive neuroscience as distinct analytical approaches to the generation of legal evidence.

Much of the extant research in the field of neuroscience and law applies neuroscience to the law through the paradigms of cognitive neuroscience — localizing causal and functional mechanisms of behaviour and mental states within neocortical representational and perceptual systems. Neuroimaging is relied upon to identify dysfunctions in neocortical mechanisms.[21] For example, it has been argued that the prefrontal cortex may play causal and functional roles in moral reasoning, the sensation of regret, and regulation of impulse control (Sapolsky, 2004). As we will elaborate below, it has also been argued that an 'intention' mechanism may be localized in the neocortex (Aharoni *et al.*, 2008). While there is work suggesting that such high level concepts as moral reasoning and intentionality cannot be reduced so simply, there is undeniably a strong trend towards purely neocortical localizations.

circumstances in which a culpable *mens rea* is not required (i.e. actions that are prohibited *per se* like statutory rape).

[20] The 'voluntariness requirement' is an analytical component of *actus reus*. Generally, criminal actions must be committed voluntarily. For example, a person who strikes another because someone is forcing him/her to do so at gunpoint would have committed the act prohibited by the Illinois Battery statute, but would not have done so voluntarily.

[21] 'Inferences about brain activity are typically made by designing experiments that contrast the MR (magnetic resonance) signal measured during two different tasks. Ideally, the tasks differ in one respect, and the location and magnitude of the difference in measured signal is attributed to brain activity involved in the difference in task performance' (Sinnott-Armstrong *et al.*, 2008, p. 361).

In contrast, affective neuroscience focuses on the evolutionary origins of causal and functional mechanisms in the affective brain. Affective neuroscience provides evidence that there are a set of basic affective mechanisms in subcortical regions that underlie a diverse range of behaviours and internal experiences. Furthermore, these foundational circuits are integrated into the more recent levels of brain development through interactive and emergent dynamics. In *Affective Neuroscience* (1998), Dr Jaak Panksepp identifies the neuro-anatomy and associated neuro-chemical governance of intrinsic psycho-behavioural control systems of subcortical emotional networks. Specifically, he maps systems for SEEKING, RAGE, FEAR, LUST, PANIC, PLAY, and CARE. Panksepp demonstrates that lesions in higher areas of the brain do not diminish responses from lower areas, while damage to lower areas of the brain compromise the functions of higher areas (*ibid.*, p. 196). Accordingly, we believe the emotional systems of the human mind are relevant factors in the neurological analysis of criminal responsibility.

While subcortical considerations of emotions and behaviour are not entirely dispositive of legal issues, they complicate the conclusions being drawn by applications of cognitive neuroscience to law. This can be shown by comparing how each discipline may approach a neurological defence to the legal element of criminal intent incorporated in a battery charge. Affective neuroscience can explain attack behaviour (i.e. battery) in a fundamentally different way than cognitive neuroscience. To be consistent with an affective understanding of the brain, applications of neuroscience and law should consider both the foundational role of subcortical regions and their dynamic interaction with cognitive mechanisms, which, only when taken together, may provide a comprehensive vehicle for legal applications of neuroscience.

Cognitive neuroscience employs a predominantly neocortical analysis to identify a defendant's diminished capacity to form intent. This methodology utilizes brain imaging techniques to show physical abnormalities in areas of the neocortex thought to serve causal or functional roles in the expression of intentional behaviour. Aharoni *et al.* (2008) gives credence to a body of imaging research purporting to localize an 'intention' mechanism in the pre-supplementary motor area (pre-SMA) of the neocortex. The authors state:

> Thus, there are two important points to glean from the neural framework outlined here: (1) There is an emerging case to be made that the pre-SMA reflects a neural basis of intention and that it displays the functional connectivity necessary for cognitive influence on intention formation and thereby on the execution of action; (2) when the neural

> areas responsible for intention are dysfunctional, an imbalance in competition between various automatic action plans allows complex actions to be performed in the absence of intention. (*Ibid.*, p. 152)

The authors explain that a criminal defendant can defend against the element of intent by offering fMRI images of the pre-SMA 'to demonstrate that the defendant's brain is dysfunctional during attempts at planned action, preventing him from reliably forming intentions but leaving intact the ability to perform the prohibited action' (*ibid.*, p. 151).

In contrast, affective neuroscience, although not necessarily contradicting the above scenario, can also approach a defence to an apparently intentional attack by referring to subcortical circuitry foundational for direct affective expressions of attack. Dr Panksepp highlights a series of experiments that use direct electrical stimulation of the brain (ESB) to evoke unconditioned anger responses in mammals. His research provides evidence for subcortical circuits (the RAGE network) stemming from the amygdalae to the hypothalamus and the periaqueductal gray (PAG); this circuit controls the expression of attack behaviours across mammals (Panksepp, 1998, p. 196). Moreover, these expressions are controlled hierarchically; the response from lower brain regions (PAG) is not dependent on higher brain areas (e.g. amygdala), for example, and perhaps even the pre-SMA. Subcortical processes are necessary for the expression of rage, and in some circumstances can be sufficient to explain attack behaviours without cognitive intention — for example, in instances where output from the RAGE circuits overpowers or floods cognitive controls. Likewise, there is evidence that damage to frontal cortical regions, where the pre-SMA is situated, can disinhibit RAGE circuitry.

Affective neuroscience research also provides evidence that learned behaviour and the organism's environment are catalysts for the causal and functional dynamics that arise from the interaction of subcortical affective systems and neocortical cognitive mechanisms. Dr Panksepp states: 'affective feelings help animals to better identify events in the world that are either biologically useful or harmful and to generate adaptive responses to many life challenging circumstances. In addition to responding to emergency situations, mild arousal of these brain systems presumably helps generate characteristic moods and coaxes animals to perform their everyday activities in characteristic ways' (Panksepp, 1998, p. 26). When comparing the two approaches, it becomes clear that cognitive neuroscience is oversimplifying the causes of aggressive behaviours. Affective neuroscience suggests neocortical localization does not sufficiently account for the causal effects of an individual's environment, nor does it account for subcortical endophenotypic tendencies that shape an individual's behaviour.

In the above example, the respective explanations of non-intentional attack by cognitive neuroscience and affective neuroscience are consistent in so far as their conclusions presume either the dysfunction of neocortical mechanisms in the former or the bypass of those mechanisms in the latter. However, the dysfunctional pre-SMA analysis of cognitive neuroscience is oversimplified in so far as it does not consider the integration of intrinsic subcortical behavioural tendencies into the neocortical localization of the presumed 'intention' mechanism itself, which has been derived largely by correlative brain imaging rather than direct causal manipulation of brain systems. Also, the presumptive localization of an 'intention' mechanism in higher brain regions does not take into account the RAGE circuit's foundational role in the expression of attack behaviours. Dr Panksepp explains, 'although cognitive and affective processes can be independently conceptualized, it comes as no surprise that emotions powerfully modify cortical appraisal and memory processes and vice versa' (*ibid.*, p. 26). In instances when the subcortical substrates are not overpowering higher level mechanisms, they are still foundational — interacting with neocortical mechanisms in a way that is not sufficiently appreciated in extant cognitive neuroscientific localizations of intentionality. Since the RAGE circuit plays a foundational role in both neocortically-mediated and non-neocortically-mediated expressions of violence, a purely neocortical analysis of intentionality is necessary but not sufficient. It is doubtful that something like 'intention' can be completely localized in a cognitive mechanism in light of evidence of similar behavioural tendencies in other mammals and the complicated dynamics between evolutionary levels of an evolutionarily layered brain (MacLean, 1990). The above discussion of attack behaviours suggests that applying the localization of neocortical mechanisms to law runs the risk of insufficiently taking into account affective and environmental complexities that can exert robust influences on legally culpable behaviours.

IV. Questioning the relevance of evidence generated through the paradigm of cognitive neuroscience

Since purely neocortical models of actions and mental states radically underdetermine human behaviour, the extant evidence generated in cognitive neuroscience applications to law is incomplete and hence arguably irrelevant. To elucidate this claim, we will introduce the following two components of legally relevant evidence: materiality and probative value. Evidence is material if it is being offered to help prove a proposition that is a matter in issue within a given case (Broun *et al.*, 2006). Probative value is the tendency of that evidence to establish the proposition that it is offered to prove (*ibid.*). These common-law components of evidence are integrated into the Federal Rules of Evidence (FRE) — the applicable code of evidence in the federal court system. Federal Rule of Evidence 401 incorporates the concepts of materiality and probative value in its definition of relevant evidence. It states:

> Relevant evidence means evidence having any tendency to make the existence of any fact of consequence to the determination of the action more probable or less probable than it would be without the evidence.[22]

Certainly, evidence bearing on a defendant's potential incapacity to form the requisite *mens rea* is material, i.e. the evidence purports to address a 'fact of consequence to the determination of the action' since a specific *mens rea* is an element of most crimes. The more illuminating analysis is whether the evidence is probative, i.e. whether any functional image of a defendant's uninjured brain has 'any tendency' to make the existence of a specified *mens rea* more probable or less probable than it would be without the image.

A cognitive neuroscience approach argues that an fMRI image of physical/functional abnormalities in the pre-SMA has a tendency to make the existence of intent less probable (Sinnott-Armstrong *et al.*, 2008). However, since a neocortical model of intentionality is, as we argued above, oversimplified, judges should question whether this brain localization evidence actually has 'any tendency' to establish a lack of intent. Since two separate neuroscientific models, in this case cognitive neuroscience and affective neuroscience, are capable of explaining the same phenomenon at two distinct levels of the brain, a judge should question which model is more accurate — at least until the two models are synchronized, or until one model proves to be empirically decisive. Even if a judge finds that the fMRI image has some tendency to make the existence of intent less probable, i.e. has probative value, the fMRI can still be said to be only minimally probative of the issue, again, because of its oversimplification of intention

[22] The language of Federal Rules of Evidence 401 and 403 (discussed *infra*) have been amended for stylistic purposes (effective 1 December 2011). The amendments are not intended to be substantive.

in light of affective neuroscience. Thus, although relevant, the evidence may be excluded under FRE 403, which states:

> Although relevant, evidence may be excluded if its probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury, or for considerations of undue delay, waste of time, or needless presentation of cumulative evidence.

FRE 403 is a balancing test: if relevant evidence is probative, but one of the listed dangers substantially outweighs that probative value, then the evidence should be excluded.

Furthermore, there is evidence that neuroscientific evidence poses significant dangers of confusing the issues and misleading the jury (Sinnott-Armstrong *et al.*, 2008, pp. 367–9). Studies show that test subjects rated bad explanations more satisfactory when accompanied by irrelevant brain information (Weisberg *et al.*, 2008). Another study showed test subjects rated articles with bad arguments as making more sense when accompanied by neural images (McCabe and Castel, 2008). Brain images and neuroscience jargon pose legitimate dangers considered by FRE 403. But, does that danger substantially outweigh the probative value, and are jurors capable of understanding the nuances of the relations between the subcortical and neocortical control of behaviour?

In ‘Brain Images as Legal Evidence’, Sinnott-Armstrong and his collaborators (2008) admit that the probative value of brain images is currently minimal, because: (1) functional normality is dubious in light of individual differences; (2) false alarms are numerous because of low base rates; (3) criminal behaviour is unlikely even with functional abnormalities; (4) correlations cannot show that abnormalities cause particular criminal acts; and (5) even causation by a brain abnormality does not prove any lack of control that would remove criminal responsibility (*ibid.*, p. 367). However, the authors argue that ‘the problems we listed can be overcome’.

The assessment of probative value undertaken in Sinnott-Armstrong *et al.* (2008) is largely restricted to the following issues: differences between individual brains, and the relationship between brain abnormalities and causation of behaviour. On the other hand, our analysis takes another tack in demonstrating how affective neuroscience suggests the methodology of fMRI itself is minimally probative because, as a tool purporting to identify causal and functional mechanisms in the neocortex, its results are oversimplified. Since neuroimaging, the primary methodology — not simply the data — of the extant literature in cognitive neuroscience and law, does not sufficiently probe subcortical processes, it can easily be argued the resulting images are minimally probative. Balancing this minimal probative value against the dangers of confusing the issues and misleading the jury should lead to the exclusion of this particular brand of neuroscientific evidence. Until affective considerations of behaviour and mental states are considered there is good reason to believe that the problem of minimal probative value cannot be overcome. As it stands, we have outlined a strong argument that neocortical images should be ruled irrelevant under FRE 401, or nevertheless excluded under FRE 403.

V. Conclusion

Affective neuroscience provides evidence that a broad array of intrinsic behavioural tendencies are organized and aroused by subcortical structures, but that learned behaviour and the environment can modulate these presets. Such findings should have a substantial impact on the integration of neurological evidence into law. We suggest this impact is two-fold. First, as discussed above, neuroscience is not yet capable of full integration into the legal system until the inconsistencies between affective and cognitive causal factors of behaviour can be harmonized. Second, as discussed below, the relationship between affective systems, learned behaviour, and the environment also should motivate us to take a critical view of the philosophical underpinnings of the criminal justice system.

An application of neuroscience and law that takes into account affective neuroscience localizes causative structures of criminal behaviour in the whole brain, but understands that the development of those causative structures, and the way they inevitably express behaviour, are formed by the cultural environment and learned behaviours. For example, Panksepp shows that the causal subcortical substrates for RAGE are predominantly activated when an individual faces restraints on freedom of action, or restraints on access to resources (Panksepp, 1998, p. 187). Affective neuroscience demonstrates that the subcortical RAGE circuit is more likely to be active in persons whose environment presents more restraints on access to resources. It follows that a person raised in an impoverished sector of society may be more likely to demonstrate RAGE as a behavioural means to confront a restraint on resources.

From a whole brain point of view, the nature of agency is much more complex than if we simply considered the tertiary-process level of mind. If affective neuroscience is to be taken as seriously as cognitive neuroscience has been taken, we may be approaching a neurological explanation of criminal behaviour that depends as much or more on the developmental brain transformations caused by societal structures than it does on deviant agency. This broadening of the explanation of criminal responsibility has consequences for the retrubutivist and utilitarian philosophies that underpin the criminal justice system in the United States. Specifically, from the perspective of retributivism, the deviant agent that committed the criminal act cannot bear full responsibility since a causative factor of his culpable action is the structure of society itself. A person should not be held fully responsible and punishable for his actions when it can be shown that the form of social organization within which the person lives amplifies antisocial behaviour. The moral responsibility for the act does not fall squarely on the actor, but, in part, on the society that knows its particular form of organization amplifies that type of behaviour. From the utilitarian perspective of rehabilitation, the greater good thought to be provided by the criminal justice system is undermined because, in reality, it may merely be reinforcing a form of social organization that knowingly amplifies antisocial behaviour. Thus, a greater good would be reached through a criminal justice system that works to challenge the ability of social structures to amplify antisocial behaviours, and not one that simply reinforces them.

References

Aharoni, E., Funk, C., Sinnott-Armstrong, W. & Gazzaniga, M. (2008) Can neurological evidence help courts assess criminal responsibility? Lessons from law and neuroscience, *Annals of the New York Academy of Sciences*, **1124**, pp. 145–160.

Belcher, A. & Sinnott-Armstrong, W. (2010) Neurolaw, *WIREs Cognitive Science*, **1**, pp. 18–22.

Blackstone, W. (1765–69) *Commentaries on the Laws of England*, Book 4, Chapter 2, Oxford: Clarendon Press.

Bonnie, R.J., Coughlin, A.M., Jeffries, Jr., J.C. & Low, P.W. (2004) *Criminal Law*, New York: Foundation Press.

Broun, K.S., Dix, G.E., Imwinkelried, E.J., Kaye, D.H., Mosteller, R.P., Roberts, E.F., Strong J.W. & Swift, E. (2006) *McCormick on Evidence*, Eagan, MN: Thompson/West.

Gardner, M.R. (1993) The mens rea enigma: Observations on the role of motive in the criminal law past and present, *Utah Law Review*, **635**, p. 640.

MacLean, P.D. (1990) *The Triune Brain in Evolution: Role of Paleocerebral Functions*, New York: Springer.

Malan, G.H.T. (1922) The behavioristic basis of the science of law, *8 American Bar Association Journal*, p. 737.

McCabe, D.P. & Castel, A.D. (2008) Seeing is believing: The effect of brain images on judgments of scientific reasoning, *Cognition*, **107** (1), pp. 343–352.

Panksepp, J. (1998) *Affective Neuroscience: The Foundations of Human and Animal Emotions*, New York: Oxford University Press.

Pound, R. (1927) *Introduction to Sayre's Cases on Criminal Law*, Eagan, MN: Lawyers Co-operative

Sapolsky, R.M. (2004) The frontal cortex and the criminal justice system, *Philosophical Transactions of the Royal Society London, B*, **359** (1451), pp. 1787–1796.

Sayre, F.B. (1932) Mens rea, 45, *Harvard Law Review*, **974**.

Sinnott-Armstrong, W., Roskies, A., Brown, T. & Murphy, E. (2008) Brain images as legal evidence, *EPISTEME*, pp. 359–373.

Weisberg, D.S., Keil, F.C., Goodstein, J., Rawson, E. & Gray, J.R. (2008) The seductive allure of neuroscience explanations, *Journal of Cognitive Neuroscience*, **20**, pp. 470–477.

Paper received September 2010; revised March 2011.