Towards the Mind of a Humanoid: Does a Cognitive Robot Need a Self? – Lessons from Neuroscience

Elena Antonova[1,2] and Chrystopher L. Nehaniv[2]

[1]Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London
De Crespigny Park, London SE5 8AF, United Kingdom

[2]Adaptive Systems Research Group and Royal Society Wolfson Biocomputation Research Laboratory
Centre for Computer Science and Informatics Research, University of Hertfordshire
College Lane, Hatfield AL10 9AB, United Kingdom

elena.antonova@kcl.ac.uk   C.L.Nehaniv@herts.ac.uk

Abstract

As we endow cognitive robots with ever more human-like capacities, these have begun to resemble constituent aspects of the 'self' in humans (e.g., putative psychological constructs such as a *narrative self*, *social self*, *somatic self* and *experiential self*). Robot's capacity for body-mapping and social learning in turn facilitate skill acquisition and development, extending cognitive architectures to include temporal horizon by using autobiographical memory (own experience) and inter-personal space by mapping the observations and predictions on the experience of others (biographic reconstruction). This 'self-projection' into the past and future as well as other's mind can facilitate scaffolded development, social interaction and planning in humanoid robots.

This temporally extended horizon and social capacities newly and increasingly available to cognitive roboticists have analogues in the function of the *Default Mode Network (DMN)* known from human neuroscience, activity of which is associated with self-referencing, including discursive narrative processes about present moment experience, 'self-projection' into past memories or future intentions, as well as the minds of others. Hyperactivity and overconnectivity of the DMN, as well as its co-activation with the brain networks related to affective and bodily states have been observed in different psychopathologies. Mindfulness practice, which entails reduction in narrative self-referential processing, has been shown to result in an attenuation of the DMN activity and its decoupling from other brain networks, resulting in more efficient brain dynamics, and associated gains in cognitive function and well-being. This suggests that there is a vast space of possibilities for orchestrating self-related processes in humanoids together with other cognitive activity, some less desirable or efficient than others. Just as for humans, relying on emergence and self-organization in humanoid scaffolded cognitive development might not always lead to the 'healthiest' and most efficient modes of cognitive dynamics. Rather, transient activations of self-related processes and their interplay dependent on and appropriate to the functional context may be better suited for the structuring of adaptive robot cognition and behaviour.

Introduction

Efforts in Artificial Intelligence in particular and, most promisingly, AI Cognitive Robotics seek to produce life-like agents by emulating the capacities seen in nature, particularly in humans (Mori, 1989; Brooks, 1986; Brooks et al., 1998; Nehaniv et al. 2013; Cangelosi and Schlesinger, 2015).[1] Various aspects of the 'self' have been introduced in synthetic agents and robots in the hope of achieving human-like intelligence and capabilities through building with constructive methods (e.g. Brooks et al., 1998; Nehaniv et al., 1999).

By endowing simple behaviour-based robots (Brooks, 1986) with extended *temporal horizon* (Nehaniv, 1999; Nehaniv et al. 2002)[2] and *second-person* capacities (Dautenhahn, 1997)[3], cognitive architectures of increasing

[1] This paper introduces and extends the ideas of Antonova and Nehaniv (2012).

[2] For example, harnessing autobiographical remembering of own sensorimotor and other 'experience' (operationally this can be taken as sensorimotor flow over an extended window of time) supports learning, prospection and expectation based on this are type of 'proto-narrative' available to robots that does not involve language. Moreover, such proto-narratives can be shared and communicated or inferred about other agents (biographic reconstruction, generalized 'story-telling' and narrative) to give robots the capacity to learn from and project their own and others' temporally extended experience into useful 'stories' or 'plans'.

[3] Concepts in social robotics of *empathic resonance* involving recognition of another agent as having a somehow similar embodiment and 'experience' to one's own and *biographic reconstruction* as recognition of the temporally extended experience of another agent are 'second person' mechanisms of immediate and broad temporal horizon, complementing other second person mechanisms such as mutual and joint attention, body mapping in imitation and social learning.

sophistication for grounded embodied cognitive development are now deployed on humanoids. Scaffolded development (Broz et al., 2012, 2016; Saunders et al. 2012, Foerster et al., 2017) and prospection (Mirza et al., 2008b) have become possible in humanoids capable of social learning of behaviours and skills in interaction with others. This can be achieved by harnessing autobiographical remembering (1^{st} person) and mapping (2^{nd} person) of the robot's own temporally extended experience to that of others (humans or robots), without building-in representational capacity (an aspect of intelligence to be explained, not pre-supposed). Adaptable body-mapping and body-schemas for control, self-repair, imitation and social learning (Alissandrakis et al., 2007a,b; Bongard et al., 2010) in addition to enhanced sensorimotor flows, with robot 'experience' and action modulated by 'affect' are now part of the toolkit of cognitive developmental systems researchers. Such capabilities are rudiments of *narrative*, *social*, *somatic* and *experiential* 'selves'.

Here we relate aspects of self-related processes as conceived in philosophy, artificial intelligence robotics and neuroscience toward a deeper understanding of the space of possibilities for how dynamic self-related processes could interplay or possibly interfere in humanoid robots.

Aspects of Self in Western Philosophy

Human experience is commonly structured by an aspect of a separate 'self' that exists independently from its environment. Ever since William James, there is a tradition of treating 'self' as a permanent feature underlying otherwise constantly changing experience. Numerous distinctions have been made in an attempt to capture aspects of self-experience (e.g. James, 1890, Neisser,1988; Strawson, 1999, Gallagher 2000, 2013). Recurring aspects include (but are not limited to) *narrative self*, *social self*, *somatic self,* and *experiential self.* These 'self'-related processes, arising in ontogeny, serve diverse and useful functions in humans. Their dynamic interplay and flexibility in shifting from one to another in a context-dependent manner is essential for adaptive (healthy) behaviour and cognition.

Gallagher (2000) proposed that various approaches to characterizing self-related processes can be divided into two main groups: the *minimal self* and the *narrative self*. The *minimal self* captures immediate momentary phenomenological subject of experience and contains, to various degrees, a sense of agency and ownership. The *narrative self* (Neisser's 'extended self') captures the sense of personal identity extended in time, constituted by the stories of the past and intentions for the future that we tell ourselves and others. The two are not necessarily mutually exclusive, but it is questioned to what extent it is possible to have a *minimal self* experience without a *narrative* one.

The *minimal self* experience can be constituted by an interplay of perceptual, somatic (bodily and homeostatic states), affective, and social (inter-personal) aspects of the present-moment experience in various degrees depending on the context. The *narrative self* in most part is supported by the episodic (autobiographical) memory processes. Both aspects of self are maintained by the process of 'self-referencing', i.e. identifying with the experience as 'I' (experiencing subject) or 'me, mine' (experiencing object). This in turn necessitates involvement of a conceptual (and perhaps language-based semantic) framework to structure one's experience.

Aspects of Self in Artificial Intelligence and Robotics

Self in Enactivist AI Approaches

Enactivist viewpoints reject the necessity of an extended 'narrative' self to structure the experience (Varela et al., 1991). They also posit the possibility of non-conceptual first-person experience that emerges from the dynamic embodied interactions with the environment and constituted by various intertwined, dependently co-originated transitory component processes. On this account, 'self' as an experience comes into existence when the relationship between top-down predictive processes and the bottom-up sensory-motor processes come into conflict. This self is not an entity extended over time, but a short-term emergent phenomenal experience brought about by self-reference (Tani, 1998).

Despite the fact that present day robots are not autopoietic (self-producing) entities, dynamical systems and evolutionary methods have been applied to analyse the robotic second-person 'experience' where action and agent-environmental coupling is given meaning for agents through Darwinian evolution (Iizuka and Ikegami, 2004; Froese and Di Paolo, 2011; Froese and Gallagher, 2012; Nehaniv, in press).

Interaction games for robots and human agents are analogous to Wittgenstein's *language games* where meaning is grounded by usage in recurring contexts (Nehaniv et al., 1999). The dynamics and information flow cut across the agent-environment distinctions, and a key success criterion comprises felicitous interaction in engagement with naïve human participants (Nehaniv et al. 2013; Foerster et al. 2017). Enactive approaches scaffolded on social learning in recurring interaction games have in this manner demonstrated the acquisition of meaningful behaviours and grounded language usage by humanoids based on 'operationalized experience' (sensorimotor-affective interaction histories / autobiographical traces) and social engagement without any explicit built-in system for representational semantics or reference (Mirza et al., 2005, 2008; Saunders et al., 2012; Broz et al., 2012; Foerster et al., 2017).

Emulating Human Features

Very often the motivation is to maximally emulate humans as much as possible in every aspect to achieve the holy grail of 'true AI' or human-like intelligence by capturing every aspect of a human in a robot, humanoid, or 'geminoid' (Ikegami and Ishiguro, 2017).

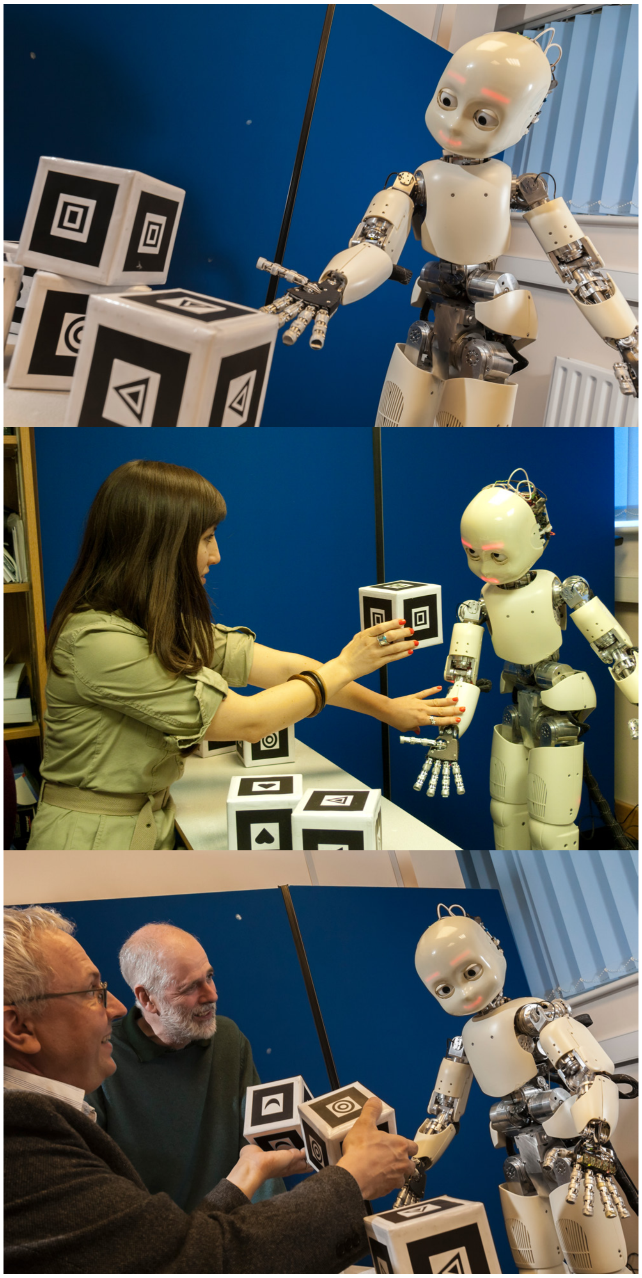

Figure 1: The iCub Humanoid DeeChee, a robotic platform at the University of Hertfordshire for embodied cognitive systems research. Cognitive architectures supporting mechanisms closely related to self processes support the humanoid’s developmental scaffolding, social learning, experiential autobiographical memory and interaction histories in embodied interaction games for bottom-up skill- and language-acquisition in interaction with human participants (e.g., Mirza et al. 2008a; Saunders et al. 2012; Broz et al. 2012; Nehaniv et al. 2013; Foerster et al. 2017).

Thus many cognitive architectures and extensions to existing computational architectures in agents and robots seek to achieve a more human-like intelligence by implementing a particular aspect of the ‘self’. Increasingly, AI robotics designers are finding they need to include aspects of *somatic self* by introducing adaptive somatic body self-models (Bongard et al., 2006) and multisensory ego-sphere saliency (Ruesch et al., 2008); *narrative self*, including autobiographical, episodic and narrative memory processes (Nehaniv, 1999; Ho et al., 2004, 2006, 2008; Nehaniv et al., 2013; Pointeau et al., 2014; Pointeau and Dominey, 2017), and first-person remembering and interaction histories (Mirza et al. 2006, 2008a; Broz et al., 2012); *social self* by mapping to the 2nd person (imitation and social learning (Nehaniv and Dautenhahn, 2002), narratives about self/others and story-telling (Ho et al., 2004, 2008), as well as pioneering work on theory of other minds (Barnden, 2005) or perspective taking (Steels and Loetsch, 2009).

AI researchers also have sought to harness or model other aspects of self-related processes in robots and agents, e.g. 'self-awareness' (Tani 1998[4], Tani 2017), or ‘consciousness’ as a global workspace with which functional modules interact (Ramamurthy et al. 2012; Franklin et al. 2016), and mental simulation capacity (Lallee and Dominey 2013; Schillaci, Hafner, and Lara, 2016), among many others that cannot be exhaustively treated in this short article.

Faith in Emergence

These endeavours generally attempt to introduce some single aspect of self, whilst ignoring the others. The general trend is to add on new modules in an *ad hoc* way. By accreting more and more human-like modules and add-ons, it is hoped that all the ingredients for human-like cognition will eventually give rise to human-like intelligence (Brooks et al., 1998).

Of course, on reflection, most AI robotics researchers now realize that there will be interaction effects between modules and would usually know enough to expect that, and, from the viewpoint of bottom-up emergence, might even want to harness positive 'synergies' and new capabilities that might arise from these interactions. Indeed, it is has long been envisioned that human-like cognition could emerge somehow through their synergetic interactions (Minsky 1988). For example, J. K. O’Regan (2011), drawing on ideas of Hume and Dennett and trends in cognitive robotics (Brooks et al., 1998; Edsinger and Kemp, 2006), suggests that by implementing in a robot all the necessary capacities for a “cognitive self (self-cognizance, self-knowledge, knowledge of self-knowledge)”[5] and for a “societal self” as a construct ‘center of narrative gravity’ (Dennett 1991), it is possible to see the outlines of how “to build a robot with a self”.

However, the details of how such emergence from a bundle of processes, modules and ‘agents’ might work, or be appropriately orchestrated, have not been mapped out. There are vast open spaces of possibilities in this topic of enquiry.

[4] J. Tani (1998) writes “structure of the ‘self’ corresponds to the ‘open dynamic structure’ which is characterized by co-existence of stability in terms of goal-directedness and instability caused by embodiment; (2) the open dynamic structure causes the system's spontaneous transition to the unsteady phase where the ‘self’ becomes aware.

[5] These terms are O’Regan’s renaming of Bekoff and Sherman (2004)’s terms for aspects of self-cognizance (‘self-referencing’, ‘self-awareness’ and ‘self-consciousness’, respectively) which they consider in the context of evolutionary continuity of animals and humans following (Darwin, 1871; Griffin, 1976).

Scaffolded Development

More recently, *cognitive development* in long-term embodied interaction, based on such interconnected architectural components, is envisioned to arise in the course of the mutually scaffolded boot-strapping of skills (including language ability) in a developing humanoid (Mirza et al., 2006, 2008a; Vernon et al., 2007; Vernon, 2010; Cangelosi et al., 2010; Broz et al., 2012; Nehaniv et al. 2013; Broz et al., 2014; Pointeau et al., 2014; Cangelosi and Schlesinger, 2015; Lyon et al., 2017). Figure 1 shows the humanoid DeeChee in which various cognitive architectures for scaffolded development and AI processes are tested and interwoven. This is seen as paralleling the cognitive development of children (Vygotsky, 1978; Kaye, 1982), and closely linked to the development of autobiographical memory, narrative intelligence, temporal grounding for narrative processes based on temporally extended episodic experience, and the acquisition of increasingly complex skills and behaviours (Mirza et al., 2006, 2008a; Broz et al., 2012), up to rudimentary referential and non-referential language such as use of negation (Saunders et al., 2012; Foerster et al., 2017).

Self-related Processes in Neuroscience

The description and interactions of the neural networks associated with the experience of the *minimal self* are highly complex, as are the interrelations between the minimal and the narrative self-processes with many unknowns. It is therefore beyond the scope of the present paper to survey the field comprehensively. In what follows, we discuss the main developments in cognitive, clinical and contemplative neurosciences in relation to the network associated with the *narrative self* and its interactions with other networks.

With the introduction of the neuroimaging methodology, such as the functional Magnetic Resonance Imaging (fMRI), the modular paradigm, which approached brain areas as independent processors for specific higher cognitive functions based on neurological lesion findings was superseded by the view that cognition is associated with a dynamic interplay between distributed brain areas operating in large-scale networks (review, Bressler and Menon, 2010). In the light of this development, neuroscientists and neurophilosophers have drawn parallels with Buddhism and its notion of 'no-self', with an experience of 'self' as separate independently existing entity being a useful illusion emerging through evolution (e.g. Metzinger, 2009).

The Default Mode Network and Self-referencing

With the description of what has become known as the Default Mode Network, we have a large-scale neural network whose activity and interactions with other brain networks appears to be associated with the experience of such a separate 'self'. A little discourse into the history of cognitive research using functional Magnetic Resonance Imaging (fMRI) is needed here. In the first instance of employing fMRI, cognitive neuroscientists were mainly interested in studying neural dynamics associated with higher cognitive functions by giving participants cognitive tasks requiring attention to external stimuli. These were contrasted with either control tasks or a so-called 'resting state', with either eyes closed or looking at a fixation cross. However, what happens in the brain during the so-called 'rest' was largely ignored. Until Gusnard and Raichle (2001) published a review of Positron Emission Topography (PET) research in which they have addressed directly the 'mysterious' task-independent decreases in specific areas of the brain during the resting state baseline, whether with a simple visual fixation or with eyes closed. The same task-independent decreases, predominantly in the midline structures, including medial prefrontal cortex (mPFC) and posterior cingulate (PC), were being consistently observed with fMRI when the resting state was used as a comparison baseline to a task of interest. The task-independence is the key here; that is, no matter the nature of the cognitive task, these areas were active in the absence of a task. Furthermore, as the attentional demand of the task is increased, the activity in the resting state regions is further decreased. This task-independent network was named the Default Mode Network (DMN) by Raichle and colleagues, since it is activated by default when we are not engaged in a goal-directed cognitive activity. It was further shown that not only the activity of the midline-based DMN is anti-correlated with laterally-distributed task-related networks during cognitive tasks, but the activity between these lateral and midline networks is constantly shifting and fluxing during 'rest', oscillating at a low frequency of about 0.1 Hz (Fox et al., 2005).

Raichle's initial intuition was that the DMN activity must be associated with spontaneous mentation, day-dreaming, mind-wandering, or in general terms 'self-referencing', quoting Seneca's *"The fact that the body is lying down is no reason for supposing that the mind is at peace. Rest is… far from restful*".

However, after this initial intuition Raichle and his colleagues have abandoned it, although they have still allowed for some of the DMN activity to be accounted for by spontaneous self-referential thinking. The primary reason, as argued by Raichle (2006), is that in the awake resting state, the brain accounts for 20% of the total oxygen consumption of the body, whilst the changes in brain activity associated with cognitive tasks amount to about 5% increase from the baseline activity. If this is the metabolic 'cost' of the effortful cognitive activity, the unconstrained spontaneous thinking, which is effortless, is unlikely to account for the brain's metabolic 'hunger' at rest.

However, other lines of research have confirmed the association of the DMN activity with self-referencing in general and narrative self-referencing in particular. A meta-analysis of the fMRI studies examining remembering, prospection and the theory of mind (Buckner and Carroll, 2007) has implicated the DMN as common network associated with 'self-projection', whether into the past memories, future scenarios, or another person's mind. The DMN activity has also been shown to be associated with a stimulus-independent thought, i.e. mind-wandering away from a cognitive task (e.g. Mason et al., 2007). Studies that have directly contrasted conditions requiring deliberate narrative self-referential processing with a neutral resting state whilst fixating on a cross confirmed the DMN involvement in both conditions, with the DMN activity being stronger and more wide-spread during the explicitly instructed narrative self-referential activity (e.g. Davey et al., 2016). Thus, the DMN recruitment is consistently observed during internally oriented cognition that engages narrative self-processes, whether spontaneous or deliberate, goal-directed.

The DMN and Psychopathology

The aberrations in the DMN function have been implicated in psychopathology. The DMN hyperactivity and over-connectivity has been observed in schizophrenia, depression (review, Whitfield-Gabrieli et al., 2009), ADHD (Liddle et al., 2011), autism (Kennedy et al., 2006). Siblings and parents of schizophrenia patients show reduced cognitive ability, but no difficulties in inhibiting the DMN during cognitive tasks, whereas patients appear to be unable to do so (Whitfield-Gabrieli et al., 2010). In depression, there is an increased connectivity between the regions of the DMN and the subgenual anterior cingulate (e.g. Berman et al., 2011; 2014), a region that plays an important role in modulating autonomic and visceral responses during the processing of sadness, fear, and stress. Functional connectivity between the DMN and the subgenual anterior cingulate has been shown to positively correlate with the duration of the current depressive episode (Greicius et al., 2007) and self-reported tendencies toward rumination and brooding (Berman et al., 2011). Furthermore, functional connectivity between the posterior cingulate and the subgenual anteriror cingulate significantly increases when individuals think about negative events in their life, compared with unconstrained rest (Berman et al., 2014). Considering what is known about the phenomenology of schizophrenia (specifically paranoid sub-type) and depression, these findings further confirm the association of the DMN activity with the experience of a separate 'self', which is under attack either from an external source (paranoia) or an internal critic (self-focused rumination).

The DMN and Contemplative Neuroscience

A further line of evidence comes from the contemplative neuroscience. A study of experienced meditators (Hasenkamp et al., 2010) who were asked to focus on the breath and press a button whenever they noticed their mind has wandered found that the episodes of mind-wandering were associated with the increased activation of the DMN. This demonstrates that the DMN activity is not simply associated with an internally focused attention, but instances of narrative self-referential thought. Farb and colleagues (2007) have directly compared narrative and experiential (minimal) self-referencing in healthy participants and individuals who have undergone Mindfulness-Based Stress Reduction (MSBR), an 8-week intensive mindfulness skill training programme based on Buddhist contemplative practices (Kabat-Zinn, 1982). In both groups, the narrative self-referencing condition was associated with the DMN activation. The experiential condition yielded focal reductions in the mPFC activity in controls, whereas MBSR participants showed more marked and pervasive reductions of the same DMN region. Furthermore, functional connectivity analyses demonstrated a strong coupling between the right insula (area associated with the visceral awareness, a part of embodied self-processing) and the mPFC in controls that was uncoupled in the mindfulness group. These results demonstrate that during an experiential mode of processing when narrative self-referencing is suspended there is a fundamental dissociation in neural dynamics associated with two distinct forms of self-awareness: the 'self' across time and the present moment experience (the 'narrative self' and the 'minimal self' of Gallagher, 2000). The ability to suspend narrative self-processes associated with the activity of the DMN and engage experiential ones are enhanced by mindfulness practice. This, in part, appears to underpin brain's increased efficiency of the information processing (e.g. Pagnoni et al., 2008), relapse prevention in depression (e.g. Barnhofer et al., 2015) and general well-being associated with mindfulness (e.g. Holzel et al., 2011).

Balanced Interplay of Self-related Processes

However, it is not all or none when it comes to the DMN activity in humans. Narrative processes related to the formation of episodic and autobiographical memories and their employment in a functionally relevant context are of clear importance for coherent cognitive function and interaction. Thus, diminished DMN functional connectivity is observed in healthy aging and is associated with age-related cognitive decline (e.g. Vidal-Pineiro et al., 2014). The disruptions in the DMN connectivity with the medial temporal lobe structures, including hippocampus and parahippocampal gyrus associated with episodic memory formation, is a hallmark of the Alzheimer's Disease (e.g. Wu et al., 2011).

Hence, it is about a balanced dynamic functionally-relevant DMN activation and its co-activation with other networks in a context-dependent manner. As argued by Brewer et al. (2013) and demonstrated using fMRI neurofeedback in conjunction with subjective reports, the DMN's sustained activity, and particularly that of the PC region, when processing self-related content (e.g. sensations, memories, emotions, thoughts) may represent "getting caught up in" one's experiences rather than narrative self-referential processes *per se*. We are all familiar with what that 'feels like' in a human; it remains to be seen what it might look like in a humanoid robot.[6] According to Buddhist psychology as well as Varela's enactivist approach for AI robotics, this 'sticky' narrative self-referencing is not functionally necessary and can indeed be detrimental.

Summary and Conclusion

The balance of cognitive resources in performing tasks can be influenced by the dynamic interplay of self-related processes, the example being a decoupling of *narrative self* and *minimal*

[6] We stay open on a possibility of phenomenal experience in a robot and leave the debate on the issue of phenomenal experience per se out of this paper. We stand with Varela (1996) on the primacy of phenomenal experience in humans and advocate his remedy for 'the hard problem of consciousness' (see Bitbol and Antonova (2016) for detailed explication of the view).

self that could be enhanced by mindfulness training. In the light of this, it seems that careful assessment is needed in determining what types of self-related processes an enactive cognitive robot should embody and whether some modes of their operation and interaction might reduce rather than increase functional and 'metabolic' efficiency, as well as effective inter-personal interactions and interactions with the environment.

The discussion here has highlighted the DMN’s association with an experience of a narrative self, its implication in psychopathologies, and its interaction with other brain networks for healthy cognitive function. Similar interdependencies between self-related processes might also affect the optimal functioning of humanoid minds. Complex unforeseen interaction effects may arise between self-related processes in humanoids, including somatic, social and narrative self-processes, or other modules or mechanisms added to humanoid robots to make them more human-like in an effort to maximally emulate all aspects of human beings.

In conclusion, in humans and hence, for similar reasons, in future cognitive robots, 'self' is best conceived of as an interplay of transitory processes arising in a context-dependent and a function-specific manner. 'Self' as a continuously running process (either as a module or a distributed network) 'overseeing' the job of other processes in a top-down manner could be functionally inefficient and even detrimental. Finally, given the findings of the clinical neuroscience in relation to the functional alterations of the DMN, as well as contemplative neuroscience showing that brain processing dynamics may be enhanced when the narrative self-processes are suspended, caution must be entertained when modelling the narrative self-processes and their relative dominance within the overall interplay of self-related dynamics.

Acknowledgments. This work was supported in part by the European Commission under projects ITALK ("Integration and Transfer of Action and Language in Robots") and BIOMICS (contract numbers FP7-214668 and FP7-318202, respectively) to Prof Nehaniv, and by the King’s Together Fund award (“Towards Experiential Neuroscience Paradigm”) to Dr Antonova.

References

Antonova, E. and Nehaniv, C.L. (2012). Do Enactive Cognitive Robots Need a Self? Lessons from Neuroscience, *Foundations of Enactive Cognitive Science,* 27-28 February 2012, Cumberland Lodge, The Great Park, Windsor, U.K.

Alissandrakis, A., Nehaniv, C.L. and Dautenhahn, K. (2007a). Solving the Correspondence Problem in Robotic Imitation across Embodiments: Synchrony, Perception, and Culture in Artifacts. In Nehaniv, C.L and Dautenhahn, K., eds. *Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions,* Cambridge University Press.

Alissandrakis, A., Nehaniv, C.L., Dautenhahn, K. (2007b) Correspondence Mapping Induced State and Action Metrics for Robotic Imitation, *IEEE Trans. Systems, Man, & Cybernetics, Part B*, 37(2):299-307, April 2007.

Barnden, J.A. (2005). Metaphor, Self-Reflection, and the Nature of Mind, In Davis, D.N., ed., *Visions of Mind: Archictectures for Cognition and Affect,* Information Science Publishing Idea Group Inc., Hershey, PA.

Barnhofer, T., Huntenburg, J.M., Lifshitz, M., Wild, J, Antonova, E., Margulies, D.S. (2015). How mindfulness training may help to reduce vulnerability for recurrent depression: A neuroscientific perspective. Clinical Psychological Science, 4(2), 328-343.

Bekoff, M. and Sherman, P.W. (2004). Reflections on Animals Selves. *Trends in Ecology and Evolution* 19(4):176-180.

Berman, M. G., Peltier, S., Nee, D. E., Kross, E., Deldin, P. J., and Jonides, J. (2011). Depression, rumination and the default network. *Social Cognitive and Affective Neuroscience*, 6, 548–555. doi:10.1093/scan/nsq080

Berman, M. G., Misic, B., Buschkuehl, M., Kross, E., Deldin, P. J., Peltier, S., . . . Jonides, J. (2014). Does resting-state connectivity reflect depressive rumination? A tale of two analyses. NeuroImage, 103, 267–279. doi:10.1016/j .neuroimage.2014.09.027

Bitbol, M. and Antonova, E. (2016). “On the Too Often Overlooked Radicality of Neurophenomenology”, *Constructivist Foundations* 11(2):354-356.

Bressler, S. L., and Menon, V. (2010). Large-scale brain networks in cognition: emerging methods and principles. *Trends in Cognitive Sciences,* 14, 277-290.

Bongard, J., Zykov, V. and Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science* 314: 1118-1121.

Brewer, J. A., Garrison, K. A., Whitfield-Gabrieli, S. (2013). What about the “Self” is processes in the posterior cingulate cortex*? Front Hum Neurosci*, doi: 10.3389/fnhum.2013.00647.

Brooks, R. A. (1986). "A Robust Layered Control System for a Mobile Robot", *IEEE Journal of Robotics and Automation* 2(1):14–23.

Brooks, R.A., Breazeal, C., Marjanović, M., Scassellati, B., Williamson, M.M. (1998), The Cog Project: Building a Humanoid Robot. In Nehaniv, C.L. ed. *Computation for Metaphors, Analogy, and Agents,* Springer Lecture Notes in Computer Science, vol. 1562, pp. 52-87.

Broz, F., Nehaniv, C.L., Belpaeme, T., Bisio, A., Dautenhahn, K., Fadiga, L., …. and Cangelosi, A, (2014) The ITALK Project: A Developmental Robotics Approach to the Study of Individual, Social, and Linguistic Learning. *Topics in Cognitive Science* 6(3):534-544.

Broz, F., Nehaniv, C.L., Köse-Bagci, H. and Dautenhahn, K. (2012). "Interaction Histories and Short Term Memory: Enactive Development of Turn-taking Behaviors in a Childlike Humanoid Robot", arXiv:1202.5600v1 [cs.AI] , 25 February 2012.

Buckner, R. L., and Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences* 11(2), 49-57.

Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C.L., Fischer, K., Tani, J, Belpaeme, T., Sandini, G., Nori, F., Fadiga, L., Wrede, B., Rohlfing, K. Tuci, E. Dautenhahn, K., Saunders, J. and Zeschel, A. (2010). Integration of Action and Language Knowledge: A Roadmap for Developmental Robotics, *IEEE Transactions on Autonomous Mental Development* 2(3):167-195.

Cangelosi, A. and Schlesinger, M. (2015*). Developmental Robotics: From Babies to Robots.* MIT Press, Cambridge, MA.

Darwin, C. (1871/1936) *The Descent of Man and Selection in Relation to Sex*, Random House.

Davey, C.G., Pujol, J., Harrison, B.J. (2016). Mapping the self in the brain's default mode network. *NeuroImage*, 132, 390-397.

Dennett, D. C. (1991). *Consciousness Explained.* Boston, MA.

Dautenhahn, K. (1997). I could be you: The phenomenological dimension of social understanding, *Cybernetics and Systems* 28 (5):417-453.

Edsinger, A. and Kemp, C.C.. What Can I Control? A Framework for Robot Self-Discovery. *Proceedings of the Sixth International Conference on Epigenetic Robotics (EpiRob 2006),* Paris, France.

Farb, N. A., Segal, Z. V., Mayberg, H., Bean, J., McKeon, D., Fatima, Z., and Anderson, A. K. (2007). Attending to the present: mindfulness meditation reveals distinct neural modes of self-reference. *Social cognitive and affective neuroscience*, *2*(4):313-322.

Foerster, F., Nehaniv, C.L. and Saunders, J. (2017). "Robots that Say 'No': Affective Symbol Grounding and the Case of Intent Interpretations", *IEEE Transactions on Cognitive and Developmental Systems*, DOI: 10.1109/TCDS.2017.2752366

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., and Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc. National Academy of Sciences U.S.A.*, 102(27): 9673-9678.

Froese, T. and Di Paolo, E. A. (2011). Toward minimally social behavior: social psychology meets evolutionary robotics. In Kampis, G., Karsai, I. and Szathmáry, E., editors *Advances in Artificial Life: Darwin Meets von Neumann. 10th European Conference, ECAL 2009*. 426–433, Springer-Verlag, Berlin, Germany.

Froese, T. and Gallagher, S. (2012). Getting interaction theory (IT) together: integrating developmental, phenomenological, enactive, and dynamical approaches to social interaction. *Interaction Studies* 13: 436–468.

Franklin, S., Madl, T., Strain, S., Faghihi, U., Dong, D., Kugele, S., Snaider, J., Agrawal, P., Chen, S. (2016). A LIDA cognitive model tutorial. *Biologically Inspired Cognitive Architectures*, 105-130. doi: 10.1016/j.bica.2016.04.003

Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences* 4(1):14-21.

Gallagher, S. (2013). A pattern theory of self. *Frontiers in Human Neuroscience* 7(443):.1-7.

Griffin, D.R. (1976) *The Question on Animal Awareness: Evolutionary Continuity of Mental Experience*, Rockefeller University Press.

Gusnard, D. A., and Raichle, M. E. (2001). Searching for a baseline: functional imaging and the resting human brain. *Nature Reviews Neuroscience*, *2*(10), 685-694.

Hasenkamp, W., Wilson-Mendenhall, C. D., Duncan, E., and Barsalou, L. W. (2012). Mind wandering and attention during focused meditation: A fine-grained temporal analysis of fluctuating cognitive states. *NeuroImage*, 59, 750–760.

Ho, W.C., Dautenhahn, K., Nehaniv, C.L. and te Boekhorst, R.(2004), Sharing Memories: An Experimental Investigation with Multiple Autonomous Autobiographic Agents. In F.Groen, N. Amoto, A. Bonarini, E. Yoshida, and B. Kröse (Eds.), *Intelligent Autonomous Systems 8, March 10-13, 2004*, Amsterdam, The Netherlands, IOS Press, pp. 361-370.

Ho, W.C., Dautenhahn, K., and Nehaniv, C.L., A study of episodic memory-based learning and narrative structure for autobiographic agents, *Proceedings of AISB 2006 Convention,* Vol 3, pp 26-29, Bristol, UK, 2006.

Ho, W.C., Dautenhahn, K. and Nehaniv, C.L. (2008). Computational memory architectures for autobiographic agents interacting in a complex virtual environment: a working model, *Connection Science* 20(1):21-65.

Hölzel, B.K., Lazar, S. W., Gard, T., Schuman-Olivier, Z., Vago, D. R., Ott U (2011). How does mindfulness meditation work? Proposing mechanisms of action from a conceptual and neural perspective. *Perspectives on Psychological Science* 6(6):537–559.

Iizuka, H. and Ikegami, T. (2004). Adaptability and Diversity in Simulated Turn-taking Behavior, *Artificial Life* 10(4):361-378

Ikegami, T. and Ishiguro, H. (2017) Android project. https://www.researchgate.net/project/Android-project

James, W. (1890). *The Principles of Pscyhology* (in two volumes). Henry Holt and Company, New York.

Kabat-Zinn, J. (1982). An outpatient program in behavioral medicine for chronic pain patients based on the practice of mindfulness meditation: theoretical considerations and preliminary results. General Hospital Psychiatry, 4, 33–47.

Kaye, K. (1982). *The Mental and Social Life of Babies: How Parents Create Persons*, University of Chicago Press.

Kennedy, D. P., Redcay, E., and Courchesne, E. (2006). Failing to deactivate: resting functional abnormalities in autism. *Proceedings of the National Academy of Sciences U.S.A.*, *103*(21), 8275-8280.

Lallee S and Dominey PF. Multi-modal convergence maps: From body schema and self-representation to mental imagery *Adaptive Behavior*. 21: 274-285

Liddle, E. B., Hollis, C., Batty, M.J., Groom, M. J., Totman, J. J., Liotti, M., Scerif, G., Liddle, T. (2011). Task-related default mode network modulation and inhibitory control in ADHD: effects of motivation and methylphenidate. *J. Child Psychol. Psychiatry Allied Discip.*, 52:761-771.

Lyon, C., Nehaniv, C.L., Saunders, J. … and Angelo Cangelosi, A, Embodied Language Learning and Cognitive Bootstrapping: Methods and Design Principles, *International Journal of Advanced Robotic Systems*, 13:105, 2016.

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., & Macrae, C. N. (2007). Wandering minds: the default network and stimulus-independent thought. *Science*, 315(5810), 393-395.

Metzinger, T. (2009). *The Ego Tunnel: the science of the mind and the myth of the self.* Basic Books, New York.

Minsky, M. (1988). *The Society of Mind*. Simon & Schuster, New York.

Mirza, N.A., Nehaniv, C.L., Dautenhahn, K and te Boekhorst, R. (2006). Interaction Histories: From Experience to Action and Back Again, *Proc. Fifth International Conference on Development and Learning* (ICDL - Bloomington, Indiana). IEEE Press, 2006.

Mirza, N.A., Nehaniv, C.L., Dautenhahn, K., te Boekhorst, R. (2008a). Developing Social Action Capabilities in a Humanoid Robot using an Interaction History Architecture, *Proc. 8th IEEE RAS Conference on Humanoids 2008*, IEEE Press.

Mirza, N.A., Nehaniv, C. L., Dautenhahn, K, te Boekhorst,, R. (2008b). Anticipating Future Experience using Grounded Sensorimotor Informational Relationships, *Artificial Life XI*, MIT Press.

Mori, M. (1989) *The Buddha in the Robot*, Kosei Shuppan-Sha; Original edition.

Nehaniv, C.L. (1999). Narrative for Artifacts: Transcending Context and Self. In: Sengers, P. and Mateas, M., eds., *Narrative Intelligence: Papers from the 1999 AAAI Fall Symposium FS-99-01,* American Association for Artificial Intelligence, pp. 101-104.

Nehaniv, C.L. (in press). Constructive Biology of Emotion Systems: first- and second-person methods for grounding adaptation in a biological and social world. In: Ferreira, M.I.A, Sequeira, J.S., and Ventura, R., eds. *Cognitive Architectures*, Springer Verlag.

Nehaniv, C.L. Dautenhahn, K., and Loomes, M.J. (1999), Constructive Biology and Approaches to Temporal Grounding in Post-Reactive Robotics. In McKee, G.T. and Schenker, P. Eds., *Sensor Fusion and Decentralized Control in Robotics Systems II (Proceedings of SPIE)* Vol. 3839, pp. 156-167.

Nehaniv, C.L. and Dautenhahn, K. (2002). The Correspondence Problem. *Imitation in Animals and Artifacts,* pp. 41-62, Bradford Books/MIT Press, Cambridge, MA.

Nehaniv, C.L., Förster, F., Saunders, J., Broz, F., Antonova, E., Köse, H., Lyon, C., Lehmann, H., Sato, Y. and Dautenhahn, K. (2013). Interaction and Experience in Enactive Intelligence and Humanoid Robotics, *IEEE Symposium on Artificial Life (IEEE ALIFE),* Singapore, 15-19 April 2013, pp. 148-155, IEEE Press.

Nehaniv, C.L., Polani, D., Dautenhahn, K., te Boekhorst, R. and Cañamero, L. (2002) Meaningful Information, Sensor Evolution, and the Temporal Horizon of Embodied Organisms. In *Artificial Life VIII*, MIT Press, pp. 345-349.

Neisser, U. (1988). Five Kinds of Self-Knowledge. *Philosophical Psychology* 1(1):35-59.

O'Regan, J. K. (2011). *Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness*, Oxford University Press.

Pagnoni, G., Cekic, M., Guo, Y. (2008). "Thinking about not-thinking": neural correlates of conceptual processing during Zen meditation. *PLoS One*, *3*(9), e3083.

Pointeau, G. and Ford Dominey, P. (2017). The Role of Autobiographical Memory in the Development of a Robot Self. *Frontiers in Neurorobotics* 11:27. doi: 10.3389/fnbot.2017.00027

Pointeau G, Petit M, Dominey PF. (2014). Successive developmental levels of autobiographical memory for learning through social interaction *IEEE Transactions on Autonomous Mental Development.* 6(3): 200-212.

Raichle, M. E. (2006). The Brain's Dark Energy. *Science*, 314:1249-1250.

Ramamurthy, U., Franklin, S., and Agrawal, P. (2012). Self-system in a model of cognition. *International Journal of Machine Consciousness*, 4(2):325-333.

Ruesch, J., Lopes, M., Bernardino, A., Hörnstein, J., Santos-Victor, J., Pfeifer, R. (2008). Multimodal Saliency-Based Bottom-Up Attention: A Framework for the Humanoid Robot iCub, *IEEE - International Conference on Robotics and Automation (ICRA'08),* Pasadena,California, USA.

Saunders, J., Nehaniv, C.L., Dautenhahn, K. and Alissandrakis, A. (2007). Self-Imitation and Environmental Scaffolding for Robot Teaching, *International Journal of Advanced Robotics Systems*, 4(1):109-124.

Saunders, J., Lehmann, H., Foerster, F. and Nehaniv, C. L. (2012). Robot Acquisition of Lexical Meaning: Moving Towards the Two-word Stage. In: *Proc. 2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL 2012),* IEEE. [DOI: 10.1109/DevLrn.2012.6400588]

Schillaci, G., Hafner, V.V., Lara, B. editors (2016). Re-Enacting Sensorimotor Experience for Cognition, special issue of *Frontiers in Robotics and AI,* Volume 3.

Steels, L. and Loetsch, (2009) M. Perspective Alignment in Spatial Language. In Coventry, K.R., Tenbrink, T., and Bateman, J., eds. *Spatial Language and Dialogue*, Oxford University Press.

Strawson, G. (1999) The self and the SESMET. In Gallagher, S. and Shear, J., eds. *Models of the Self*, pp. 483–518, Imprint Academic

Strawson, G. (2018). The Consciousness Deniers. *New York Review of Books*, 13 March 2018.
http://www.nybooks.com/daily/2018/03/13/the-consciousness-deniers/

Tani, J. (1998). An interpretation of the 'self' from the dynamical systems perspective: a constructivist approach, *J. Consciousness Studies* 5(5-6):

Tani, J. (2017). *Exploring Robotic Minds: Actions Symbols, and Conscousness as Self-Organizing, Dynamic Phenomena*, Oxford University Press.

Varela, F.J. (1996). Neurophenomenology: A Methodological Remedy for the Hard Problem, *J. Consciousness Studies*, 3(4):330-349

Varela, F. J., Thompson, E. and Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience.* MIT Press, Cambridge, MA.

Vernon, D. (2010). Enaction as a Conceptual Framework for Developmental Cognitive Robotics, *Paladyn: Journal of Behavioral Robotics*, 1(2):89-98.

Vernon, D., Metta, G. and Sandini, G. (2007). A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents, *IEEE Transactions on Evolutionary Computation* 11(2):151-180.

Vidal-Pineiro D., Valls-Pedret C., Fernandez-Cabello S., Arenaza-Urquijo E. M., Sala-Llonch R., Solana E., et al. (2014). Decreased default mode network connectivity correlates with age-associated structural and cognitive changes. *Front. Aging Neurosci.* 6, 256. 10.3389/fnagi.2014.00256

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes.* Harvard University Press.

Whitfield-Gabrieli, S., and Ford, J. M. (2012). Default mode network activity and connectivity in psychopathology. *Annual Review of Clinical Psychology*, *8*, 49-76.

Whitfield-Gabrieli, S., Thermenos, H. W., Milanovic, S., Tsuang, M. T., Faraone, S. V., McCarley, R. W., ... and Wojcik, J. (2009). Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia. *Proceedings of the National Academy of Sciences*, *106*(4), 1279-1284.

Wu X, Li R, Fleisher AS, Reiman EM, Guan X, Zhang Y, Chen K, Yao L. Altered default mode network connectivity in Alzheimer's disease—a resting functional MRI and Bayesian network study. *Hum Brain Mapp*. 2011;32:1868–1881.