

Affect in Autonomous Artificial Systems: Interfacing Technology and Philosophy

Chryssa Sdrolia¹

Abstract. Inspired by the latest developments in affective artificial systems, this paper is concerned with investigating the insertion of emotion in artificial intelligent systems and mapping its implications on a philosophical, political, and cultural level. This goal is primarily pursued by drawing multiple parallels between the philosophies of representation and expression that have problematized the concept of ‘emotion’ as well as between the technological and philosophical uses of the loaded concepts of ‘emotion’ and ‘affect.’ What is argued is that the affective turn in artificial systems necessitates and at the same time creates an interface between technology and philosophy, which is vital for the articulation of a viable ethics within the larger framework of late post-industrial capitalism.

1. INTRODUCTION: THE CONTEXT

With networked systems having already infiltrated the patterns of everyday life and technology manifesting an ever-increasing ability to overcome its technical constraints, scientists are focusing their attention on the value of emotive expression as the condition and strategy *par excellence* for the success of human-machine interaction (Duffy 177). Efforts for the seamless insertion of machines into human ecologies are thus drawing upon an understanding of the body as an affective entity. Within the field of affective systems, this interest is especially manifest in evolutionary robotics for which the coupling of control software with the (machinic) body as a physical event of expression is vital for the creation of fully operational autonomous agents. It is known that since the late 1990s, at least, research has been revolving around the creation of emotionally grounded robotic architectures also capable of displaying real-time/real-world perception, mobile and multi-modal interaction competence (Breazeal 2003, Cañamero 2005, Picard 2003). Posing a number of practical and theoretical challenges, the affective turn in robotics and artificial intelligence, in general, is as challenging as it is vexed. The technical difficulty in synthesising emotion and emotional expression as well as the conceptual indeterminacy of ‘emotion’ vis-a-vis ‘affect’ has created the urgent and rather felicitous need for trans-disciplinarity. Cognitive psychology, neuroscience, ethology, and philosophy have been marshaled to aid the development of a computational framework that captures emotional processing and integrates it with other models of perception, behaviour, and motor control. Taking as a point of departure this productive tension, my interest in the area is thus primarily informed by the problematics posed by philosophy and cultural theory.

While scientists face their own problems in attempting to integrate emotional expression into intelligent robotic and computational models, in humanities the debates concerning the validity and implications of the concepts ‘emotion’ and ‘affect’

respectively are part of a long tradition and still raging. Yet before sketching out the philosophical theoretical approaches to the subject, it makes sense to mention in passing the historically troublesome relationship between the two fields. With the exception of a few schools to which I will shortly return, thinking alongside scientific developments and knowledge is a relatively recent trend in the humanities and one that is complicated enough in itself. As Rosi Braidotti notes, it ‘runs against a well-established tradition of criticism, if not of actual rejection of the ‘hard’ sciences in social theory’ (Braidotti, 2002 230). The mistrust has been primarily directed against the supposed ‘objectivity’ of modern human and hard sciences (physical, biological, psychological, and social). The main argument has been that these purport to offer universal scientific truths about human nature that are, in fact, often mere expressions of ethical and political commitments of a particular society. Notable currents of critical philosophy, including poststructuralism, feminism, social constructionism, and postcolonialism have undermined such claims by exhibiting how they are more often than not the outcome of contingent historical forces and not scientifically grounded truths. In Foucault’s theorizations, for instance, as systems of representation the various sciences actually produce the rules and practices they need thus engendering meaningful statements and regulating discourses in different historical periods (Foucault 196-215). Whereas I would not like to ignore or invalidate the important implications mentioned above, I would nevertheless like to turn to the fruitful convergence of current technologies with the interests of contemporary philosophy, also keeping in mind the equally dubious part many a strand of philosophical thought has played in the shaping of political reality. After all, the mutually perceived gap between the sciences and philosophy and their respective responsibilities has not nearly been that clear-cut and is rather ‘a symptom of the anxiety of contamination’ (Huysen ix) of each other’s presumed purity.

2. EMOTION AND AFFECT IN THE PHILOSOPHIES OF REPRESENTATION AND EXPRESSION

Being pro-contamination, I thus endorse Gilles Deleuze’s opinion that ‘[a]rt, science, and philosophy...are caught up in mobile relations in which each is obliged to respond to each other, but by its own means’ (Deleuze xiv). My endeavour is not to criticise engineering techniques or scientific knowledge as unscientific but to pick up from that ‘beyond the scope’ point that many scientific papers naturally leave open when faced with the ethical and socio-cultural implications of their technological artifacts. The affective paradigm shift in intelligent systems and the sheer complexity of ‘affect’ can thus be seen as clearing up a promising plane of interaction, an

interface between technology and philosophy. To resume from where I left it, the import of affective systems seems to be paralleled by a quite old interest within the humanities in the conceptual grounding of emotion. A first observation that could be made at this point is the different use both disciplines make of the concepts 'emotion' and 'affect' themselves. In scientific writings these two terms usually figure interchangeably or are subcategorised with emotion overriding affect. In a number of papers, the second is presented as an aspect that connotes 'moods' or 'drives' and hence is not considered to be a fully-fledged emotion which is more properly connected to distinctive self-conscious cognitive states (Khulood & Raed 696). What is important here is that this clarity-based categorisation is contested in philosophy which locates a fundamental disparity between the two. Different schools have drawn attention to the sociopolitical uses of 'emotion' as a representational construct. The argument is that as a concept and discursive practice it actually presupposes and propagates dominant notions of subjectivity and selfhood in the interests of hegemonic ideology and power structures. Deconstruction and transcendental empiricism have, each from a different perspective, countered the classical folding of emotion into unitary accounts of the mind. Yet while deconstruction retains the term 'emotion,' transcendental empiricism with which my own sympathies lie drops it in favour of 'affect.' The conflict between emotion and affect is thus caught between philosophies of representation and philosophies of immanence/ expression, a fact which could be fruitfully connected to the debates within the scientific community as to the efficiency of representation in the construction of artificial systems and their future with or without it (Brooks 1991b, Müller 2007).

As emotion is entangled in the mysteries of consciousness, its history has been locked inside the classical histories of mind and will, and hence, to subjectivity conceived through representation. Emotion is thus seen as the territory that remains to subjects after the realisation that even not strictly 'proper' subjects have affects. Through the exposure of the supposed transparency, universality and presence of representation as a 'white mythology,' deconstructionists have pointed to the fact that '[t]he classical picture of emotion already contraindicates the idea of the subject' (Terada 7). Approaching a theory of emotion without however stating it as such, Jacques Derrida shows how emotion is embedded within and calls forth the textual structures that belong to what is known within the humanities as the 'death of the subject' – a pattern that emotion actually surpasses. He does this by challenging Plato, Rousseau and mainly Husserl's conception of 'auto-affectation' which conceives subjectivity on the basis of representational translucence (Derrida 1973 48-70). According to the Husserlian phenomenological schema, the mode of intentional, conscious self-reflexivity guides the route from affects, namely mere corporeal sensations, to meaningfully interpreted emotions that can be ascribed to cognizant subjects only. In this sense, and as an outcome of higher cognition, emotion is what happens when the subject practically represents itself to itself. Against the classical metaphysics of presence, Derrida thus undertakes the task of revealing the non-coincidence of represented meaning with a supposed fixed content.¹ The persistent *differance* he

traces within representation actually paves the way for a textual schema of emotional yet non-subjective experience. This entails the idea that because mental representations are never quite properly faithful *re*-presentations of a supposed unified subjectivity, they are also never quite equal to themselves. As such, subjectivity remains a fictional crossroads that is always deferred and never really crossed. Contrary to what Husserl and classical phenomenology would perceive as the complete effacement of emotion, this productive impossibility of the subject thus preserves emotional experience minus a central self-conscious subjectivity. To cut a long story short, emotion in poststructuralist theory 'does not demand a subject, unless an infinite abyss of transpersonal perspectives is your idea of a subject' (Terada 46).

The deconstructionist debunking of subjectivity in favour of emotive experience beyond intentionality is not very far from Daniel Dennett's theoretical formulations that non-subjective interpretation alone explains experience. Not unlike Derrida, his suggestions are concerned with the issue of experience that underlies most theories of emotion. In his formulation of 'heterophenomenology' and debate about 'qualia,' the notion of the self-generating, self-conscious subject is vividly undercut as the 'persuasive imagery of the Cartesian Theatre [that] keeps coming back to haunt us – laypeople and scientists alike – even after its ghostly dualism has been denounced and exorcised' (Dennett 1991 107). With this move Dennett challenges the content approach to emotion as a representation that requires a subject-overseer. Revealing the subject to be the personified correlative of qualia, he actually problematises the deeply rooted common intuition of a 'central Witness[']s] intentional stance. In his own words:

These raw materials, whether they are called 'sense data' or 'sensations' or 'raw feels' or 'phenomenal properties of experience,' are props without which a witness makes no sense. These props, held in place by various illusions, surround the ideas of a central Witness with a nearly impenetrable barrier of intuitions (Dennett 1991 322).

Dennett offers an account in which qualitative states become hostages of the human subject but which nevertheless manage to live on (Terada 110). Besides the serious debates this model has generated among engineers and computer scientists as to whether it is possible or not to construct feeling machines, his 'sub-personal' (Dennett 1981 228-229) materialist philosophy most importantly points to the fact that 'self-differential selves are dead only as [unified] subjects' (Terada 156).

The abovementioned idea is stretched even further by the school of transcendental empiricism which is closer to that materialist framework than deconstruction. Notwithstanding the valuable conceptual and political implications of deconstructionist theory, the criticism against Derrida is that he never really breaks from the endless circle of representational iterability and remains more confined in the boundaries of philosophy 'proper.' As such, emotion is caught in the state of causing only a momentary setback to the supremacy of the representing subject since it cannot but constantly repeat it. By

philosophical tradition a 'logocentrism' or 'metaphysics of presence' (sometimes known as *phallogocentrism*) which holds that speech-thought (the *logos*) is a privileged, ideal, and self-present entity, through which all discourse and meaning are derived (Derrida 1976 141-143).

¹ Deconstruction's central concern is a radical critique of the Enlightenment project and of metaphysics. It identifies in the Western

contrast, and resting on Spinoza on whom he heavily draws, Deleuze offers a way out of the impasse by mobilising the concept of 'affect' and making of his philosophy one of expression or immanence. His theory begins with a strong critique of representation as being part and parcel of the same 'state philosophy' that has haunted Western metaphysics since classical times (Deleuze 164-213). Representational thinking, Deleuze argues, constructs monstrous discursive monuments by forcing an analog of 'symmetrically structured domains' that stifle difference in favour of a slavish adherence to the ideality of transcendental Forms. As Massumi sums it up, '[t]he subject, its concepts, and also the objects in the world to which the concepts are applied have a shared, internal essence: the self-resemblance at the basis of identity' (Deleuze and Guattari 2004a xi). Contrary to the policing intentions of representational thought, the philosophy of immanence argues for affective expression by emphasising the importance of material bodily experience. In this framework, expression is not confused with the traditional subjective expressive hypothesis which is just another symptom of representation but taken to enact the material actualization of virtuality.

Bearing the flavour of theories of emergence by which it is influenced, Deleuzian theory is based on the dynamic bond between form and matter, thus tying expression to the body and engaging the two in a process of 'becoming' rather than 'being.' The force of this formulation lies precisely in the fact that it interrupts the ontological transcendental with the empirical. As Braidotti puts it, '[the] affective stratum makes it possible for Deleuze to speak of a pre-discursive moment of thinking' (Braidotti 2002 74). With a single stroke, he short-circuits interpretive subjective thought as much as he surpasses emotion. This, nevertheless, does not mean that the subject is completely done away with as common criticism against Deleuze has it; rather than discarding the subject, the focus is on its functioning as a virtual threshold through which transversal flows are constantly combined and dismantled not unlike the turbulent involutions and foldings of form and matter. The Deleuzian self is not bogged down to self-assuring ideality but is 'fascinated, always stretching to its breaking point, to the continuation of another multiplicity that works it and strains it from the inside' (Deleuze and Guattari 2004b 275). The state that becomes such a self is therefore not an emotional but an affective one. In this sense, it bears a remarkable affinity to Francisco Varela's description of auto-affection as the chiasmus of organic forces, the result of which is 'a nonsubstantial self that acts as if it were present, like a *virtual interface*.' 'The more we see the selfless nature of ourselves in various regions of the organism' Varela continues, "the more we become suspicious of our feelings of "I" as a true center" (Varela 61, original emphasis). In a similar vein, Deleuzian affect attempts to account for the multivalence and ubiquity of affective states in connection to the phenomenality of experience in a way that is surprisingly close to current developments in evolutionary affective systems: feeling-affects as the responses of affected bodies occupy the interval between affection and action. As he says, '[t]he pure self of the 'I think'...appears to be a beginning [of philosophy and thought] only because it has referred all its presuppositions back to the sensible, concrete empirical being' (Deleuze 164). Freeing the empirical being from the pretensions of the intentional subject, Deleuze thus restores thought from the status of mere

reflection to an intensive process, the self as a threshold of flows, and the body as a plane of intensive affective states.

3. FROM AFFECT TO AFFECTIVE ARTIFICIAL SYSTEMS

For a number of reasons, including its openness to scientific knowledge, the Deleuzian framework proves to be particularly efficient in conversing with the late progress in affective computing and robotic autonomous systems. Recasting thought and the subject in an intensive loop with materiality and emphasising the body as a highly expressive medium, 'affect' serves to interface philosophy and technology. On a methodological level, it puts to the test the philosopher's universal self-indulging stance of assuming a bird's eye view of the developments of a supposedly mundane world. As such, it rids the interaction of the two fields of supremacist fantasies. As Deleuze puts it:

Philosophy obviously cannot claim the least superiority, but...creates and expounds its own concepts in relation to what it can grasp of scientific function and artistic constructions. A philosophical concept can never be confused with a scientific function or an artistic construction, but finds itself in affinity with these in this or that domain of science or style of art. Philosophy cannot be taken independently of science or art (Deleuze xiv).

On a practical level, the regime of affect takes us full circle to the developments that are specific to current scientific debates over the '[synthesis] of emotions as the primary means to create believable autonomous synthetic agents' (Velásquez 70). Outside philosophy, engineers and computer scientists are equally troubled by the efficacy and uses of representation in the construction of artificial models, including the legitimacy of the concepts 'agent,' 'emotion' and 'affect' in themselves. In line with modern cognitive science that suggests we should dispose of the image of intelligent agents as central representation processors, roboticist Rodney Brooks for instance argues for artificial cognition without representation and without agents. In one of his papers, he claims that 'in the very simple level intelligence...explicit representations and models of the world simply get in the way' (Brooks 1991b 139). Stating that central representation is the 'wrong unit of abstraction,' he is in favour of intelligent systems that interface directly to the world through perception and action. Apart from their apparent practicality, these questions become increasingly philosophical when he criticises the von Neumann model of computation and the categories of thought and reason that have dominated artificial intelligence as a field (Brooks 1991a 569). Dreyfus moves to a similar direction when he questions representational models of intelligence using Merleau-Ponty's *Phenomenology of Perception* (Dreyfus 2002).

Needless to say, the implications of such claims for the construction of affective intelligent machines are not nearly resolved, as the voices in favour of the representational approach are equally strong. Some scientists support the view that physical embodiment is neither necessary nor sufficient as a basis for affective or any AI research. Instead, they propose

the synthesis of emotion through the traditional approach, utilising symbolic rule-based agent systems in software environments (Etzioni 1993). As expected in any productive domain of research, these are countered by scientists, especially from the field of robotics, who argue in favour of a joint approach, evoking the implementation of emotional-based control architectures with physically embodied agents. In her paper on emotion understanding in autonomous robotic research, Cañamero outlines a number of different approaches regarding models, applications and theory (Cañamero 448-9). The latter include a wide range of emotion-based systems, from rule-based ones to ‘emergent emotion’ approaches influenced by evolutionary artificial life models and biologically-inspired emotion architectures (Braitenberg qtd. in Cañamero 448). Acknowledging a complex fit between an agent and its environment interactionist AI and robotics attempt to circumvent the Cartesian split precisely by assuming that material embeddedness is crucial for all let alone emotive/affective intelligence (Maes 136). Obviously, what is at issue is the tortuous relationship between emotion, affect, mind, and embodiment. Having reached the ‘beyond the scope’ point and limitations of *this* paper, however, I would like to return to the productive problematic raised by the convergence between these areas of technological and philosophical knowledge.

4. CONCLUSION: PROBLEMATICS AND QUESTIONS

Often said to be ‘last and impassable frontier of computationalist theories of mind,’ emotion, affect and the necessity or not of their embodiment complicate the picture both in the humanities and hard sciences (de Sousa 70). Despite the common interest in the expressive body, the apparent concern that arises here is the difference in the use of conceptual tools. Thus, whereas for a certain number of scientists emotion is a means to enhance human-like robotic or artificial subjectivity and personality, for philosophers affect seems to dismantle subjectivity be it human or robotic. The paradox of the situation becomes even more complicated if we take into account the endo-disciplinary scientific and philosophical differentiations and the fact that the very concept of ‘human/ agent’ itself is under serious questioning by philosophers. At a time that robots tend to become humanoid and humans mechanoid through prostheses and their splicing with software environments, ethics have become more tentative than ever. In her account of the passage from the classical conception of human to the cybernetics-informed posthuman, Hayles draws attention to such issues; against a reading of the rising technologies as symptoms and harbingers of nihilism, she argues that the emergent discourses of distributed cognition may actually prove very helpful in shaking the solid ontological and teleological foundation of the liberal humanist subject (Hayles 281-7). Thus, from a cultural theorist point of view, the question would be whether it is possible to avoid the cliché, dangerously humanist discourses of anthropomorphism that affective systems could possibly bring along with them. If the discourse of ‘emotion,’ and ‘autonomy’ propagates the good old grand narratives, will the affective turn in robotics and artificial intelligence help us reconsider our sense of aliveness

and corporeality in ways that will avoid falling back to social constructionism, essentialism, and neo-liberal relativism?

Given the complexity of the situation and the multiply-informed interaction among so many disciplines, such vexed questions should be approached with caution for catastrophist morality is even more dangerous than the fantasised or real harms of technology. After all, ‘[t]echnology is not just the expression of the desire for mastery, but also the object of desire, curiosity and affective involvement (Braidotti 2002 215). Yet I would also like to suggest that they cannot be considered apart from the centrality of affective systems in the late capitalist culture either. In many respects, and far from being just another humanist fantasy projected on metal, the recent advances in evolutionary robotics and computer science are partly a response to the demands of an aggressive capitalism that profits on the exchange of emotions/affects at a deeper level. Among other theorists, in his analysis of the political economy of post-industrial advanced capitalism, Massumi points to the commodification and management of anxieties, affective states and ambient fears (Massumi 187). After the transition from the personal computer era to the ubiquitous computing era, the ‘social robot’ seems to be the latest taste in the global market for the so-called ‘personal robot.’ This trend is a curiously apt example of the schizophrenic double-bind of capitalism that Deleuze and Guattari identify within contemporary culture (Deleuze & Guattari 2004a 242-260). Within this ‘seamless’ consumerist context that the paradox of ‘user-controlled’ robotic ‘autonomy’ becomes even more obvious: to meet the perceived desires of the market, robots must be autonomous/self-organising and emotionally limited/user-controlled at the same time. As such, they are part of a post-industrial complex that reasserts individualism and personalised commodities as the unquestionable standard while feeding off the breaching of individual sanctity by pushing commercial profit-making to the innermost boundaries of subjectivity (Braidotti 2006 11). In the end, it is not only a matter of whether computers and robots will be capable of bodily affective expression but of how users invest them emotionally, as well. If the possibility of constructing an emotionally intelligent (or affective?) agent that will throw up metaphysical punchlines is an elusive goal for many years to come, these areas are of crucial importance and certainly preclude any attempts to think technology outside the cultural and philosophical terrain and vice versa. As Bishop puts it, ‘[i]t is clear that the purpose of...computations is contingent on their *social use*. In Heideggerian terms, computing machinery doesn’t exist in the world until it is put to some use’ (Bishop 7). Reconsidering the symbiotic relation between the human and the technological is therefore vital and an ethics of mutual interrelation can be best achieved through the interfacing of technology and philosophy.

REFERENCES

- Bishop, Mark. ‘Why Computers Can’t Feel Pain.’ *International Conference on Computers & Philosophy*. Laval, France, 2006. 1-8.
- Braidotti, Rosi. *Metamorphoses. Towards a Materialist Theory of Becoming*. Cambridge: Polity, 2002.

- _____. *Transpositions: On Nomadic Ethics*. Cambridge: Polity, 2006.
- Breazeal, Cynthia. 'Emotion and Sociable Humanoid Robots.' *Human-Computer Studies* 59 (2003): 119–155.
- Brooks, Rodney. 'Intelligence without Reason.' *Proceedings of 12th International Joint Conference on Artificial Intelligence (IJCAI), Computers and Thought*. Ed. John Myopoulos Sidney: Morgan Kaufmann, 1991a. 569-95.
- _____. 'Intelligence without Representation.' *Artificial Intelligence* 47 (1991b): 139-59.
- Cañamero, Lola. 'Emotion Understanding from the Perspective of Autonomous Robots Research.' *Neural Networks* 18.4 (2005): 445-455.
- Deleuze, Gilles. *Difference and Repetition*. Trans. Paul Patton. London & New York: Continuum, 2004.
- Deleuze, Gilles & Guattari, Felix. *Anti-Oedipus: Capitalism and Schizophrenia* Trans. Robert Hurley, Mark Seem & Helen R. Lane. London & New York: Continuum, 2004a.
- _____. *A Thousand Plateaus: Capitalism and Schizophrenia*. Trans. Brian Massumi. London & New York: Continuum, 2004b.
- Dennett, Daniel. 'Why You Can't Make a Computer That Feels Pain.' *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge: MIT Press, 1981. 190-229.
- _____. *Consciousness Explained*. Boston: Little, Brown, 1991.
- Derrida, Jacques. *Speech and Phenomena and Other Essays on Husserl's Theory of Signs*. Trans. David Allison. Evanston: Northwestern University Press, 1973.
- _____. *Of Grammatology*. Trans. Gayatri Chakravorty Spivak. Baltimore: Johns Hopkins University Press, 1976.
- de Sousa, Ronald. *The Rationality of Emotion*. Cambridge: MIT Press, 1987.
- Dreyfus, Hubert L. 'Intelligence without Representation. Merleau-Ponty's Critique of Mental Representation: The Relevance of Phenomenology to Scientific Explanation.' *Phenomenology and the Cognitive Sciences* 1.4 (2002): 367-83.
- Duffy, Brian. 'Anthropomorphism and the Social Robot.' *Robotics and Autonomous Systems* 42 (2003): 177–190.
- Etzioni, Oren. 'Intelligence without Robots: A Reply to Brooks.' *AI Magazine* 14.4 (1993): 7-13.
- Foucault, Michel. *The Archaeology of Knowledge*. London and New York: Routledge, 1989.
- Hayles, N. Katherine. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago: UCP, 1999.
- Huysen, Andreas. 'The Vamp and the Machine: Fritz Lang's Metropolis.' *After the Great Divide. Modernism, Mass Culture and Postmodernism*. Bloomington and Indianapolis: Indiana University Press, 1986. 65-81.
- Khulood, Abu Maria & Raed, Abu Zitar. 'Emotional Agents: A Modeling and an Application.' *Information and Software Technology* 49.7 (2007): 695-716.
- Maes, Pattie. 'Modeling Adaptive Autonomous Agents.' *Artificial Life* 1.1-2 (1994): 135 – 162.
- Massumi, Brian. *A User's Guide to Capitalism and Schizophrenia*. Boston: MIT Press, 1992.
- Müller, Vincent. 'Is There a Future for AI without Representation?' *Minds and Machines* 17.1 (2007): 101-115.
- Picard, Rosalind. 'What Does It Mean for a Computer to Have Emotions?' *Emotions in Humans and Artifacts*. Eds. Robert Trapp, Paolo Petta and Sabine Payr. Cambridge, Massachusetts: The MIT Press, 2003. 213-235.
- Terada, Rei. *Feeling in Theory. Emotion after the 'Death of the Subject'*. Chambridge, Massachusetts, and London: Harvard University Press, 2001.
- Varela, Francisco. *Ethical Know-How: Action, Wisdom, and Cognition*. Stanford: Stanford University Press, 1999.
- Velásquez, Juan. 'When Robots Weep: Emotional Memories and Decision-Making.' *Proceedings of the Fifteenth National/ Tenth Conference on Artificial Intelligence/ Innovative Applications of Artificial Intelligence*, 1998. 70-75.