Artificial Emotions

Philip K. Dick's novel, Do Androids Dream of Electric Sheep—the basis for the narrative of the film, Blade Runner—considers the status of empathy as a cognitive function. The distinction between humans and androids in the book is adjudicated along the lines of respective capacities for empathy. The question of whether empathy is a potential characteristic of androids has been a trope of popular fiction and cinema for nearly a century, Dick's novel is significant in that it explores the ways in which empathy can be cultivated or expressed. Can empathy become a competency if it is not an intrinsic feature of a cognitive inheritance? Many early science fiction works treat the emotional life of androids as matters of emergence: when a system becomes sufficiently sophisticated, the properties humans describe as "emotions" or "feelings" may appear of their own accord. From a narrative perspective, this mysterious emergence may be compelling, but in the case of the frontiers of robotics, particularly A.I. science, the question is increasingly treated as one of engineering.

The concept and content of theory of mind has long been the province of philosophy, but now, as engineering becomes an epistemic horizon of its own, the production or recapitulation of cognitive functions in digital form necessitates an understanding of the minds that may putatively emerge as a result of A.I. research. To produce a mind, one must understand the features of that mind that one is seeking to produce. This may be, conceptually, easier with regard to computational functions and perception —aspects of cognition digital technologies can fairly easily replicate—with regard to emotional or empathic features of cognition matters are more difficult to resolve. A.I. engineers, however, appear unwilling to trust the aesthetic principle of waiting for a sui generis emotional life to emerge and pursue projects with robust theories of mind, and emotional response in train. Ironically, however, it is in the most lapidary elements of A.I. that the closest phenomena to emergent emotional states yet observed in A.I. have been recorded.

Scientists working on Google's DeepMind project have sought to facilitate the learning process of robots by simulating the properties of sleeping in the neural networks they have produced. In the interest of facilitating planning functions and strategic decision making, the scientists have been training the networks by using vintage video games. As robots become more acquainted with the strategic requirements and choice matrices such games produce, they also become more conscious of the valences of outcomes. Negative outcomes appear to be capable of producing forms of perseverating in the networks during their dream phases in which the networks examine negative valence outcomes from the video games over and over again. This activity could be understood as a form of a A.A. or artificial anxiety. Though matters have not reached the point where neural networks might feel anxiety over the wellbeing of their Tamagotchi or neuroses over whether Siri is really into them or just being a professional, nevertheless, the realm of the kind of perseverative behaviours exhibited by these neural networks suggest that perhaps one of the first examples of potential emergent A.I. emotional states would likely be a variation on worry rather than an imperial megalomania or a sense of Enlightenment vintage public spirited empathy.

This tension between artificial emotions and technology could be understood as a kind of mirroring of the very Enlightenment belief in systems from which computer science emerged. Instead of creating machines devoid of emotion, it has turned out to be the case that emotions remain important to human beings when it comes to the machines humans create. Ironically, it is often that in terms of appearance, the least humanoid machines are often those that solicit the most emotional engagement. Robots with only the vaguest human appearances solicit greater levels of affection, for example, in Japan, than robots that are most obviously “human”. One need only think of the strange affection R2D2 commands in the Star Wars films relative his more humanoid and austere companion C3PO. The aesthetic affects of mediation and strange-making or alienation are well known as tropes of theatre as in Brecht or Beckett, but in a world increasingly mediated by technology, where the biosphere and the technosphere begin and end are increasingly difficult points to delineate. Nevertheless, the status of alienation as at the very least a teleological expedient has applications outside of the realm of art. Technological mediation has often served as a means of distancing human emotions from complex or unpleasant tasks in which functions like empathy might interfere. One may think of surgical procedures in which all but the relevant region of the body is covered up to ensure that the anxieties of empathy do not interfere with the mechanical task at hand. Empathy and embodiment can, literally cut more than one way.

Whether or not empathic robots will ever dream of electric sheep, VR tech is permitting human beings to cultivate a further mirroring: the capacity to exist for extended periods of time in virtual environments. This has implications on multiple levels. Immersed in a virtual ecology, human emotions are recontextualised by second level processes: humans applying theory of mind to tech design create simulated environments for humans to experience genuine emotion from a first person perspective. Here the ebb and flow of “realism” and “strangeness” becomes a matter of fidelity: a cognitively acceptable reality emerges as a byproduct of sufficient fidelity and technical finish. The human being encased in a VR helmet or wearing VR glasses is both totally estranged from their surroundings, existing in parallel as a cognitive being and a material being in two different spatial dimensions of experience. Experience is richer for being less “real”, less materialised. One may think of Bifo’s notion of the erotics of a generalised intellect in this context. The fate of such an erogenous zone is peculiar: human subjects experience, for example, empathy more deeply but in greater isolation. The joining of these two realms maybe one of the results of an AI culture predicated on dynamic emotional relations between humans and machines. Androids may bond with their human companions by reminiscing about the difficulties of mastering vintage Atari games. If this is a strange proposition, it is perhaps only so because at present, our own minds are as mysterious to us as those we seek to create with technology.