

ETHICAL ISSUES IN ARTIFICIAL INTELLIGENCE: DISTRIBUTED AFFECT AND THE MENTAL HEALTH OF ROBOTS

William H. Edmondson

School of Computer Science, University of Birmingham,
Edgbaston, Birmingham, B15 2TT, UK.
w.h.edmondson@bham.ac.uk

Abstract

Cognitive scientists working in the field of Human-Computer Interaction have developed a theoretical perspective on their problems and data which deserves to be more widely known. This approach – Distributed Cognition – needs elaboration and comparison with some earlier theoretical work in psychology, and also has scope for extension to cover emotion. This paper introduces the ideas of Distributed Cognition and sketches some plausible elaborations and comparisons – the restricted length here prevents a detailed account. The applicability of such work to developments in AI is explored with reference to ethical implications.

Keywords: Robots, Distributed Cognition, Distributed Affect, Ethics, Emotion.

Introduction

In 1920 Karel Capek wrote a play called in English “RUR (Rossum’s Universal Robots)” (Capek, 1920). This play did several things – it introduced into the English language the word Robot, and also the concept: an artefact that is a super-intelligent but emotion-less worker in humanoid form. It also developed the theme that these Robots had problems with mental health – they were destroyed if they behaved unexpectedly and the psychological aspect of their construction was the subject of continuous refinement (to the point where they became capable of love and perhaps procreation). The conception of Robot introduced by Capek is quite familiar to us still – individual artefacts with physical skill and strength, having a brain of encyclopaedic capacity, and natural language capability, yet isolated and without essential human traits; but nonetheless rather convincing in a mindlessly servile sort of way. Interestingly, there is also in the play a short exchange between a human (Helena) and a Robotess (Sulla) in which Helena attempts to determine through conversation if Sulla is not human.

The mental life of Robots, as envisaged by and since Capek, is a distorting mirror for human mental life, reflecting more than anything else the presumption that cognition is contained within the individual cognizers. Humans are simply perambulatory, but otherwise entire, cognizers. This view of human cognition (and thus, by implication, of Robot cognition) has been challenged, initially in relation to human interaction with artefacts.

Edwin Hutchins (Hutchins, 1996) is perhaps the originator of the Distributed Cognition approach to the study of human cognition in relation to artefacts. We will review the conception below, but the reader should keep in mind the possibility that the core notion is more extensively evidenced in human behaviour than initially recognized. For example, it is of interest that similar ideas have been discussed in a different cultural tradition; there are reasons for believing the distribution of cognition to be an interesting model for work in developmental studies, and also

for work on modelling and studying affect. If this is all plausible, or at least promising, in relation to humans, then it is relevant to the AI community working on the development of cognitively active artefacts, as well as to those working in developmental psychology.

Distributed Cognition

The core concept developed by Hutchins and taken up by others is that brains are not isolated cognizers working on sensory data from the ‘external’ environment – cognition is not *in* the head (entirely), but is rather spread out in space and time and amongst other cognizers. The examples below make this clear.

Cognition is distributed over *space* – humans are good at doing this; we leave ‘PostIt’ notes all over our offices, homes, computer casings..., with sparse data which serve to remind or prompt or record.... We use diaries and address books as cognitive extensions. In these and other ways we distribute our cognition into space.

Cognition is distributed over *people* – humans do this all the time; teamwork requires team members to understand and engage in the distributed cognition which is the demonstrated consequence of people working together. In plant control rooms, or spread out doing different tasks in the operation of complex equipment such as a naval vessel, or perhaps just as two mechanics fixing a car..., humans rely on situation awareness in order to appreciate their individual role in the larger endeavour.

Cognition is distributed over *time* – the evolution of a solution to a problem requires successive cognizers to contribute in a way which ensures that other, later, participants recognize the cognitive activity and engage in it at a ‘temporal distance’. The evolution of tools (as discussed by Hutchins, for example) is an excellent demonstration because recent tool developers manifestly read in the earlier solutions the cognitive activities of prior problem solvers, and can participate in a shared or distributed cognitive endeavour spread over perhaps decades or even longer.

When deploying a power tool to fasten two pieces of wood with a screw, one can be overwhelmed by the sense of knowing how the originator of the ‘cross-head’ screw came to have the idea – the business end of the power-tool slips out of conventional slot-head screws, but cannot when cross head screws are used. Likewise, in effective teamwork one comes to ‘know what the other team members are thinking’. In general this sense of ‘shared’ cognition can be referred to as situation awareness – and it can be quite simple (knowing that a deaf person needs to see your well-lit faced if they are to lip-read) or quite complex (understanding what a student doesn’t understand in order to help them come to understand it!). This doesn’t imply identical thoughts, but awareness of the other’s mental world in sufficient detail for it to feel familiar.

Mind-reading

The successful exploitation or establishment of clear scenarios where distributed cognition is an unavoidable interpretation of behaviour creates in participants a sort of ‘mind-reading’ sensation. This ‘mind-reading’ aspect of Distributed Cognition is key to understanding the concept. What is intriguing is that this concept surfaces elsewhere in cognitive science but in an apparently unrelated domain: developmental studies. For many years scientists looking at the cognitive development of children have worked with the notion ‘theory of mind’, which is short-hand for a child’s ability to understand what is going on in the mind of another person (Carruthers and

Smith 1996). A child is said to have a theory of mind when it can conceive of how another person views a situation or event, when that viewpoint is not the same as the child's (because of some intervening event witnessed by the child but not known to the other person).

A necessary elaboration of the notion of Distributed Cognition, in my view, concerns development of the facility within children and the subsequent personal distribution of cognition over time. The way a person thinks about a particular problem, and its solution, say, will change with that individual's experiences. The solution may itself change and thus the cognizer is not an unchanging processor any more that it is an apparatus without situational anchor. The 'mind-reading' aspect of Distributed Cognition in the context of individual development means not just that the child comes to be able to think about others in an objective way, it also means that the child comes to be able to review their own mental state – to read their own mind as it were – and this introspection leaves its own traces ensuring that even within an individual cognition is distributed over time.

Vygotsky's *Thought and Language* (1934, 1986) deals with the relationship between language and thought. Kozulin's introductory essay (Kozulin, 1986) reviews Vygotsky's work and its context, and presents an account of the developing child becoming an individual through acquisition and use of psychological tools such as language. This account would appear to be a pre-echo of modern work on 'theory of mind' in the sense that individuation is a necessary precursor for recognition of difference of perspectives in different actors, noting further that such recognition is also necessary from a social perspective, and constitutes an elaboration of social awareness.

In summary thus far – the notion of Distributed Cognition appears relevant to new and 'old' work on the development of cognition in children, and plausibly to the further development of cognition in adults (for example, the training of teamwork skills – this aspect is not discussed here further).

Distributed Affect

The conjecture to be discussed at this point is that human emotional experience and behaviour is susceptible to characterization using the same four-way approach to distribution of cognition. Briefly, it seems that affect can be distributed over space (e.g. mementoes, graveyards and stones...); over people (e.g. families, ceremonies, audiences for theatre, music...); over time (e.g. literature, religion...); and during individual development (e.g. attachment, separation, individuation). (See, for example, Parkin, 1999). Of course, Distributed Cognition and Distributed Affect may be combined in some situations – an example here might be the cognitive and emotional teamwork in a small ensemble of musicians such as a string quartet.

Clearly the literature which needs to be surveyed to reach any sort of appreciation of the plausibility or otherwise of this perspective is not closely related to work in AI or Human-Computer Interaction. It is important even to 'stray' as far afield as, say, psychoanalytic theory, if the distribution of cognition and affect is to be properly understood. For example, it seems clear (Gosling, 1968) that the psychoanalytic notion of *Transference* is very much the same as the concept introduced here as Distributed Affect, although the latter is perhaps more wide ranging. However, it is possible without such a survey to recognize some of the implications of the idea.

Implications

Arising from the above comments and conjectures there are three classes of implications, all with ethical dimensions. Firstly, there are those concerning the scientists and clinicians working on developmental psychology and affect. Secondly, there are implications for humans who develop complex emotional relationships with artefacts. Thirdly, the artefacts themselves (the Robots) may need to include in their design emotional understanding/processing in order to be able to understand the behaviour of humans (at least) around them.

i) If it is the case that a new theoretical approach – Distributed Affect – seems relevant, and especially if it is the case that the new approach is in some way implementable or testable experimentally, then it is important for those who do such implementations to interact with the more conventional specialists to promote what is necessarily a joint, though interdisciplinary, intellectual endeavour. The issue has an ethical dimension in the sense that understanding the emotional worlds of humans (and other species) is important but cannot involve experiments likely to cause harm. Living organisms have irreversible experiences; computational models and implementations can be reversible and thus used to aid understanding of extreme situations and events as experienced by organisms. This can already be done with, say, metabolic models. The ethical issues here are clear enough.

ii) My second concern is that the human users of Robots will develop feelings for Robots, and will subject themselves and their metal friends to all sorts of emotional energies and loads. Humans will need to extend their emotional development to embrace these strange affectual challenges. Indeed it might be argued by anthropologists that this is already ‘a problem’ for humans in relation to artefacts such as cars or homes. We know already that people form very complex relationships with their pets, and that indeed there is some reciprocity in such attachments, but that will not necessarily help us understand our interactions with Robots. The problem is not so much that humans make such attachments, it is that we do not understand very well what is going on, and what might need to be done in terms of human development to equip people to cope emotionally with Robots. This too is an ethical issue.

iii) Thirdly, if it is the case that scientists working on AI expect to build Robots or ‘intelligences’ then these creations should probably be equipped with more than just Distributed Cognition (but see below). Robots in a sea of affect will fare badly if they haven’t got a clue what is being attributed to them, or being thought/felt of/about them, and so forth. They might, in an important sense, need to be affective in order to bear any affective load; they might have to be emotional to maintain their mental health. This is an ethical issue in the sense that emotionally immature Robots will present humans with strange behaviours and these might raise ethical concerns. Additionally, of course, the Robots themselves may present the ethical challenge – is it unethical to construct Robots with emotions, or unethical to build them without emotions?

The third issue above is actually rather complex, both inherently and in its dependence on the other two issues. Consider the first point – the building of models of emotion, as expressions of our understanding of human emotion, and as ways of advancing that understanding through empirical evaluation. Distributed Affect sounds plausible but that doesn’t make it valid. It could

be that workable models can be made, and that they suffice for a while, until a better theory comes along. But are the models to be confined to the laboratory? If the models exist, and the need for their use in Robots is thought overwhelming, then they will be deployed in Robots. Is it ethical to deploy inevitably immature models? Is it safe, acceptable, ethical, or merely pragmatic to argue that the empirical evaluation should be being conducted via the Robot implementations, and not just in the laboratory?

In relation to the second point, one ethical concern that might arise is that of controlling the emotional models in the Robots to ensure that they do not induce unreasonable or unacceptable behaviour in their human ‘users’, or encourage such behaviour in, or via, their ‘owners’, ‘masters’, or ‘controllers’. One need only think of the Stanford ‘Prisoners and Guards’ experiment (Haney et al, 1973) to recognise the possibilities for intentional misuse of Robots. However, it is likely that some misuse – in the sense of inappropriate interactions with humans – is potentiated by the mere existence of Robots.

In relation to the third point then, it seems clear that although some sort of emotional modelling within a Robot is desirable if it is to be able to understand human behaviour, and thus to maintain its own mental health, the argument is by no means closed. Give Robots the capability to interact emotionally with other Robots as well as humans, and the outcomes cannot readily be predicted; is this sufficiently concerning ethically to warrant any moratorium on the development of such machines? If so, how would that be achieved; if not is society ready to live with the consequences?

References

Capek, K. 1920. *RUR (Rossum’s Universal Robots)*. Translated by Paul Selver. Doubleday 1923, and also OUP 1923. See also <http://capek.misto.cz/english/> for details of the origin of the word Robot.

Carruthers, P., and Smith, P.K. 1996. *Theories of Theories of Mind*. Cambridge University Press.

Gosling, R. 1968. What is Transference? In *The Psychoanalytic Approach*, edited by J.D. Sutherland. Published for the Institute of Psychoanalysis, by Baillière, Tindall and Cassell Ltd.

Haney, C., Banks, C., and Zimbardo, P. 1973. A Study of Prisoners and Guards in a Simulated Prison. Office of Naval Research, *Naval Research Reviews*, September 1973. Washington, D.C.

Hutchins, E. 1996. *Cognition in the Wild*. MIT Press.

Kozulin, A. 1986. Vygotsky in Context. Introductory essay to the 1986 edition of Vygotsky’s *Thought and Language*. MIT Press.

Parkin, D. 1999. Mementoes as Transitional Objects in Human Displacement. *Journal of Material Culture* Vol.4(3):303-320.

Vygotsky, L. 1934/1986. *Thought and Language*. Translated/edited by A. Kozulin. MIT Press 1986.