You really need to know what your Bot(s) are thinking about you

Alan FT Winfield
Bristol Robotics Laboratory

Abstract

The projected ubiquity of personal companion robots raises a range of interesting but also challenging questions. There can be little doubt that an effective artificial companion, whether embodied or not, will need to be both sensitive to the emotional state of its human partner and be able to respond sensitively. It will, in other words, need artificial theory of mind - such an artificial companion would (need to) behave as if it has feelings and as if it understands how its human partner is feeling. This essay explores the implementation and implications of artificial theory of mind, and raises concerns over the asymmetry between an artificial companion's theory of mind for its human partner and the human's theory of mind for his or her artificial companion. The essay argues that social learning (imitation) is an additional requirement of artificial companion robots, then goes on to develop the idea that an artificial companion robot will not be one robot but several. A surprising consequence of these ideas is that a family of artificial companion robots could acquire an artificial culture of its own, and the essay concludes by speculating on what this might mean for human(s) interacting with their artificial companion robots.

1. Introduction

In [Pel2007], Peltu and Wilks set out the potentially desirable features of an artificial companion in response to the question: what features are likely to make a 'good companion'? The same paper then refines this feature set by addressing the question: how human should an artificial companion (AC) attempt to be. A clear conclusion is that an effective AC will need to be endowed with the capacity to make an appropriate emotional response, that we can label *artificial empathy*, and with a built-in "...autonomous modelling capability to enable the AC to generate its own behaviour according to its perception of the interaction with the user and the environment". In the context of an artificial companion this autonomous modelling capability is, in effect, *theory of mind*.

This essay considers the implications of artificial empathy and theory of mind in *embodied* Artificial Companions, in other words *Robots*. With the caveat that meeting the potentially desirable feature set in [Pel2007] is technically well beyond the current capability of robotics, I first consider how we might implement artificial theory of mind - and hence artificial empathy - in a robot; and what this might imply in terms of the two kinds of minds (human and artificial) in the relationship between a human and his or her artificial companion.

The essay then goes on to consider the implications of embodiment and what additional constraints and/or requirements follow. In particular we consider firstly, *social learning*, and secondly the possibility that an embodied AC might in fact need not one but several bodies. Our embodied artificial companion might in reality turn out to be not one but a small *family* of autonomous robots (some mobile, some not), able to learn not only from their human charges but also from each other. The essay concludes with a cautionary discussion about what this might mean for the humans in this novel relationship.

2. How an Artificial Companion might have Theory of Mind

Theory of mind is the term given by philosophers and psychologists for the ability to form a predictive model of others [Car1996]. With theory of mind, it is supposed, we are able to understand how others might behave in particular circumstances. We can empathise because theory of mind allows us to imagine ourselves in the situation of others and hence feel as they feel. Damaged or impaired theory of mind has been conjectured as giving rise to autism in humans [Hob1990]. However, the idea of theory of mind is empirically weak - we know it should exist but we have only a poor understanding of the neurological or cognitive processes that make theory of mind. Artificial Intelligence (AI), and its embodied counterpart - robotics, provide an interesting constructionist approach to theory of mind because they allow us to ask the question "how would we build artificial theory of mind in a robot?" In the emerging field of machine consciousness, for instance, Holland and Goodman argue that if we put a computer simulation of both the world and another robot in the artificial brain of a robot, then it has the ability to ask "what if" questions about how the other robot might behave by running simulations [Hol2003]. Since computer simulation is a well-developed technology in robotics then Holland and Goodman's approach offers an apparently straightforward route to implementing a simple kind of behavioural artificial theory of mind in robots. Furthermore, the approach offers the possibility that a robot's theory of mind might improve over time. Consider two robots A and B, where A has a simulation of B in its artificial brain. A could run several simulations of B before deciding what to do next. A could then observe how B actually behaves in response and compare that observation with how it 'thought' B would behave (in simulation). The difference between the two could then be used to modify A's internal model of B. In this way A improves its artificial theory of mind, for B.

This discussion has focussed on how a robot could have an artificial theory of mind for another robot (or indeed, for itself), but in principle there is no reason to suppose that same approach couldn't be used for a robot to have (or learn) theory of mind for any other agent with which it interacts, including humans. Now one could argue that it is one thing for a robot to have theory of mind for another robot, with its relatively simple repertoire of behaviours, but quite another for a robot to have theory of mind for a human being with all of the nuanced complexity of human behaviour. But an effective digital companion surely must have some theory of mind in order to be able to demonstrate artificial empathy (or a workable analogue). It is possible therefore for a robot to be imbued with some simple theory of mind for a human being, which it can then develop and improve with experience. Models of human emotional responses are being actively explored in the field of research known in Japan as 'Kansei', see for instance [Suz1998].

Consider now the relationship between a human and his or her artificial companion robot. If we accept the argument above then the robot can have a theory of mind for the human and, furthermore, that theory of mind can improve over time. But what kind of theory of mind does the human have for its robot companion? There is an *asymmetry* here, because the robot's theory of mind has been carefully engineered with, as a starting point, a model of human behaviour and psychology (with some level of fidelity), and the ability to learn from experience. In contrast the human has no theory of mind for its robot companion. Of course humans have an extraordinary propensity to anthropomorphise so will quickly assume some theory of mind for their robot companions. But that theory of mind is almost certain to be wrong. The question then is: does it matter if we have the wrong theory of mind for our AC robots? I believe that it does. Because our AC robot has been designed with both intentionality, or at least to behave as if it has intentionality, and with artificial empathy, then when it acts or reacts in a way that we didn't expect, we will be much more disconcerted than we are when machines without agency behave in

ways we don't expect. My point is this: if your car or washing machine behaves in a way you don't expect, then you fix it without worrying about whether you might hurt it's feelings; if your AC robot behaves unexpectedly do you (a) fix it, (b) interact with it in an effort to get it to change it's behaviour (ask it nicely to adjust it's theory of mind), or (c) modify your expectations (adjust your theory of mind for it and avoid hurting it's artificial feelings).

3. Artificial Companion robots will need to be able to learn socially, by imitation

In section 2 above I argued that an AC robot with theory of mind might be able to improve its theory of mind by, in effect, learning from experience. Our AC robot would be classified as a Popperian Creature within Dennett's Tower of Generate-and-Test, because it has the ability to internalise and test (in it's simulated model of other robots or humans) possible actions before trying them out [Den1995]. With the ability to modify it's internal model our AC robot can individually learn from its own experience.

I now argue that an AC robot needs to be able to learn socially (i.e. by imitation), as well as individually, for the following reasons. Firstly, an embodied AC will be embodied precisely because it needs to physically interact with its human charges and its environment. That physical interaction is likely to be by means of manipulators (arms and hands with fine motor skills) coupled with mobility (legs or wheels), and an AC robot is highly likely to be humanoid because it needs to share human living or work spaces and, quite possibly, use human tools and appliances. Given that human spaces are chaotic and constantly changing our AC robot cannot possibly be pre-programmed with all possible physical interactions – instead it will have to learn those interactions. By far the easiest and most natural way to learn many such interactions will be by observing and imitating a human. In other words a human should be able to *teach* an AC robot how to perform a task, by simply performing the task.

The second reason that I believe the ability to learn socially (by imitation) to be an essential feature of an AC robot goes far beyond utility, or naturalness. Dautenhahn argued that “social intelligence for artefacts ... may be a necessary prerequisite for those scenarios in which autonomous robots are integrated into human societies, interacting and communicating both with humans and with each other” [Dau1995]. Researchers within the emerging interdisciplinary field of imitation and social learning in robotics, humans and animals [Neh2007] suggest that imitation and theory of mind are strongly interdependent. In that volume, for instance, Williams writes of the imitative origins of mind reading [Wil2007]. I believe that this body of work strongly suggests that an AC robot will be a more useful and effective companion if it's theory of mind incorporates social learning.

4. An Artificial Companion robot will not be one robot but several

When we think of digital companion robots we think of one robot as a companion to one human, or perhaps one household. The reality is likely to be more complex. Because all the functions of our robot are unlikely to be embodied in a single machine, it's likely that instead of a single complex multi-functional companion robot we will have several simpler single- or few-function specialist robots. For instance the robot that vacuums the floor is highly unlikely to be the same robot that provides healthcare monitoring and assistance; the 'butler' robot of the movie *Bicentennial Man* is, I believe, highly unlikely. Of course it's then not only likely but inevitable that the several robots that, collectively, comprise our digital companion are networked and thus wirelessly able to share information to coordinate their actions or alert each other to the needs of

their human(s).

This raises the important question: is it better to design such a system as, either, several robots each able to perform its specialist function autonomously and independently of the others but also able to communicate with each other if necessary, or, a single centralised artificial intelligence (perhaps with the theory of mind discussed above) able to control and coordinate a number of specialist but 'dumb' physical robots? The first approach we can think of as a 'society' of embodied robots cooperating to meet our needs, the second as a single digital companion with a 'distributed' body.

I argue that the first approach: a society of cooperating robots is both more likely and much more desirable. More likely for pragmatic reasons: specialist function robots are significantly easier to design than complex multi-functional ones. More desirable for several reasons, firstly reliability, if one robot fails the others will be unaffected in their ability to continue their specialist functions; secondly flexibility, because you only need the set of robots that meet your particular needs. More importantly, I would argue that a society – or perhaps *family* might be a better word – of robots that collectively meet your companionship needs will be more socially acceptable than a single AI with multiple robot bodies. I believe the latter approach could be deeply unsettling because the multiple robot bodies of the centralised AI will appear to speak with the same voice, and that voice will be perceived as a disembodied all-seeing centralised AI (cf. HAL in the movie 2001). In contrast the family of robots approach will be comforting since each autonomous robot will have a different 'personality' – like a family of pets. Why will they each have a different personality? Because, although each robot may well be programmed with the same 'initialised' theory of mind, each will quickly acquire a different learned experience of interacting with its human charges and the world, through its unique body, and those learned differences will be manifest as different personalities.

5. A family of Artificial Companion Robots could acquire an Artificial Culture of its own

This essay has argued that an AC robot will need theory of mind, and the ability to learn both individually through experience, and socially by imitation. And, furthermore, that an AC robot will not be one robot, but several wirelessly networked, each with specialised physical functions or capabilities. It follows that each of the robots in this 'family' that collectively comprises the AC will, subject to their physical differences and limitations, be able to imitate and hence socially learn from each other, as well as their human companions. If we adopt the terminology of memetics, and accept the notion that memes¹ copied by imitation from one robot to another may be subject to the three evolutionary operators of variation, selection and heredity, then it follows that new patterns of socially learned behaviour could emerge and (memetically) evolve within a family of AC robots [Bla1999]. These patterns of behaviour would represent an emerging *artificial culture*.

Of course memetic transmission could (and indeed will) occur between humans and robots but, because this will be mediated largely by humans teaching their AC robots it is, I would suggest, unlikely to result in a free-running emerging human-robot culture. The asymmetry discussed in section 2 would hinder any emerging human-robot culture. More likely, I believe, is that a robot-robot culture will emerge and evolve within the AC robots; the fact that the robots which collectively comprise the AC share the same basic AI and are – in effect telepathically – linked will allow the artificial culture to evolve very quickly. This emerging robot-robot artificial culture will of

1 A 'meme' is defined as a unit of cultural transmission.

course be inscrutable to humans; it will be, in essence, an alien or exo-culture. As Blackmore writes in [Bla2003] if AC robots' "...perceptual and categorisation systems were different from ours then their language might be extremely difficult for us to learn. We would find it hard to understand what they were talking about because they would parse the world in different ways and talk about completely different things from us".

6. Concluding discussion

What might the arguments presented in this paper suggest about how we approach the future design, or practice, of robots as artificial companions? This essay has, in essence, raised questions and concerns about the *asymmetry* of kinds of minds: human and artificial, both at an individual level in section 2, and - significantly amplified - at the cultural level, in section 5.

We can conclude that at an individual level it's likely that a relationship between a human and his or her artificial companion will need to be one of compromise. Thus, even though our AC robot is an artefact, we will need to be prepared to learn its ways, just as it will have to learn our ways. This, I believe, is significantly and properly different to the common expectation of robots as servants, or even slaves.

At a cultural level the problem is potentially more challenging. We could mitigate the possibility of an emerging and evolving robot culture inscrutable and therefore disconcerting to humans (but note again the asymmetry: human culture is understandable to our AC robots because we have designed them that way), by engineering the AC robots perceptual and categorisation systems to be just like ours [Bla2003]. But this will be difficult, perhaps even impossible. Our artificial companions will see and experience the world very differently to us, not least because their sensory systems will be different to ours - even if we were to tune our AC robots' sensors to the same range of colours and sounds that we experience. We might need instead to embrace and celebrate a new kind of robot culture complementing and enriching our own. A dress rehearsal perhaps, for possible future contact with Alien cultures?

References

[Pel2007] Peltu M and Wilks Y, Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues, OII Forum Discussion Paper, Oxford Internet Institute Discussion Paper No 9, November 2007.

[Car1996] Carruthers P and Smith PK (eds), Theories of Theories of Mind, Cambridge University Press, 1996.

[Hob1990] Hobson RP, On the origins of self and the case of Autism, Development and Psychopathology, 2, 163-81, 1990.

[Hol2003] Holland O and Goodman R, Robots with Internal Models, in Machine Consciousness, ed. O Holland, Imprint Academic, 2003.

[Suz1998] Suzuki K, Camurri A, Ferrentino P, Hashimoto S, Intelligent agent system for human-robot interaction through artificial emotion, in Proc IEE conf. on Systems, Man. and Cybernetics, San Diego, 1998.

[Den1995] Dennett D, Darwin's Dangerous Idea, Simon and Schuster, 1995.

[Dau1995] Dautenhahn K, Getting to Know Each Other - Artificial Social Intelligence for Autonomous Robots, Robotics and Autonomous Systems, 16, 333-356, 1995.

[Neh2007] Nehaniv C and Dautenhahn K (eds), Imitation and Social Learning in Robots, Humans and Animals, Cambridge University Press, 2007.

[Wil2007] Williams J, Copying Strategies by people with Autistic Spectrum Disorder: why only imitation leads to social cognitive development, in Imitation and Social Learning in Robots Humans and Animals, eds. Nehaniv C and Dautenhahn K, Cambridge University Press, 2007.

[Bla1999] Blackmore S, The Meme Machine, Oxford University Press, 1999.

[Bla2003] Blackmore S, Consciousness in Meme Machines, in Machine Consciousness, ed. O Holland, Imprint Academic, 2003.