Position Paper for Workshop on Artificial Companions in Society:
Perspectives on the Present and Future
Oxford Internet Institute
25th–26th October, 2007
Revised: November 29, 2007

Requirements for Digital Companions
It's harder than you think[1]

Aaron Sloman
School of Computer Science,
University of Birmingham, UK
http://www.cs.bham.ac.uk/~axs/

Contents

[1]The title has changed slightly and some of my terminology has changed since my slide presentation on 26th October. In particular what I called 'Caring functions' I now call 'Engaging functions'.

1 Introduction

The workshop on Digital Companions presupposes that development of Digital Companions (DCs[2]) will happen, and asks about ethical, psychological, social and legal consequences. DCs "*will not be robots ...[but]... software agents whose function would be to get to know their owners in order to support them. Their owners could be elderly or lonely. Companions could provide them assistance via the Internet (help with contacts, travel, doctors and more) that many still find hard, but also in providing company and companionship.*

A great deal of work is currently being done on artificial agents to interact with people, whether robotic or not, and whether as long term companions or not. The target functions of these systems can be roughly divided into two categories (a) 'engaging' functions that focus on the quality of interaction between machine and user, and (b) 'enabling' functions that focus on meeting needs or achieving goals of the user, though the division is not a very sharp one. This paper focuses on the enabling functions and some of the difficulties in meeting those requirements. I suggest that the achieving the stated goals for DCs requires major advances in both our understanding of the requirements and in our knowledge of how to design and implement machines with human-like capabilities (apart from shallow physical similarities). I shall indicate some ways of thinking about how to achieve the enabling functions that relate to interdisciplinary research problems involving AI, philosophy, psychology, biology and possibly neuroscience.

In what follows I shall use the word "user" to refer to individuals for whom DCs are provided: they are people who will regularly interact directly with the DC and who are expected or intended to derive benefit from doing so. I shall use the word "carer" to refer to other individuals who are involved in the decision to acquire the DC, and who have, or feel they have, some ongoing responsibility to the user. It is not always a user's interests that are most influential in decisions to provide an artificial companion for that user.

The functions of DCs are related to the goals of various people involved in designing, buying or using them, including both users and carers. The paper will briefly contrast different sorts of goals, whose further discussion would be relevant to the ethics of production and use of DCs, but that is not the topic of this paper. Ethical issues will, however, arise in connection with some of the more sophisticated enabling functions of DCs.

1.1 Functions of DCs

Many researchers are developing digital companions and related sorts of interactive systems that are expected somehow to improve the human condition. The systems they are trying to produce can be subdivided according the functions they are aiming for. At a high level of abstraction we can distinguish 'engaging' functions from 'enabling' functions, and both can be further sub-divided as follows.

Engaging functions

- TOYS:
  toys or entertaining devices for occasional use (compare computer games, music CDs).
- ENGAGERS:
  engaging products that are intended to be used regularly to provide interest, rich and deep enjoyment or a feeling of companionship (compare pets, musical instruments).
- DUMMY-HUMANS (pacifiers?):
  engaging products that are intended to be regarded by a user as being like another caring, feeling individual with whom a long term relationship can develop – even if it is based on an illusion because the machine is capable only of shallow manifestations of humanity: e.g. learnt behaviour rules.

[2]Sometimes referred to as ACs, "Artificial Companions".

Enabling functions

- HELPERS:
  Systems that can reliably provide help and advice that meets practical everyday needs as well as occasional unexpected problems.
- DEVELOPING HELPERS:
  These are "helpers" that are capable over time of developing their understanding of the user's environment, needs, preferences, values, knowledge, capabilities, so that they extend and improve their ability to help and support the user, as opposed to being restricted by the functionality given to them by their designers. Learning helpers are needed because it will be impossible for designers to anticipate everything and also because the user's situation and the user's needs and preferences will change over time.
- CARING DEVELOPING HELPERS:
  A human user and the user's environment will provide very many opportunities for learning. The DC will have to choose what to attend to and make decisions about which observations are relevant to its function. How to control such a process is far from obvious, and it is likely to to be very difficult. One form of control would be to make the helper really want to do what is best for the user. So the most sophisticated DCs may need to be developing helpers that grow to care about the user and really want to help when things go wrong or have a significant probability of going wrong, and which want to find out how best to provide such help.[3]

  Such a helper will want to find out what the user wants, prefers, likes and dislikes; and will also come to want to act in accordance with those attitudes. It may sometimes notice conflicts between what the user wants and what is good for the user and will generally favour the long term benefit of the user,[4] making exceptions in the sorts of cases where a caring human helper would.

These divisions are not unique to DCs, and the divisions are not very well defined: they merely provide a crude, first draft, classification, relevant to the purposes of this paper. Each category can be broken down into further sub-categories.

1.2 Motives for acquiring DCs

Decisions to apply such systems, and to choose between different combinations of functions to go into a DC will often not be made by the user, or not by the user alone. For example, government or medical agencies, members of the family, and managers of homes for such people may all be involved, either acting alone or in cooperation with one another. They may or may not take the user's desires and preferences into account. In addition to distinguishing the different kinds of *function* that can be served by a DC we can ask about the *reasons* or *motives* behind decisions to acquire one, bearing in mind that sometimes the user will not be the decision-maker.

In particular there are motives that concern only the user's desires and benefits, and motives of other humans who care about or have responsibilities to the user. Example motives for providing a DC are:

- Because the user wants the DC:
  For example, the user may sincerely prefer to be helped by a DC so as not to have to impose on other

[3] I argued in Sloman (1971) that it is possible for machines to have their own goals, preferences and values if, like humans, instead of being given only *fixed* high level goals, they start off with mechanisms for absorbing and modifying goals, preferences and values during development.

[4] Some of the issues involved in this are discussed in Will Lowe's contribution to the workshop.

humans and the others involved may respect that preference, even if they would prefer to provide the care themselves.

- Because carers have constraints:
  The others may want the DC to be available to fill gaps and provide needed help and care when human carers are unavoidably unavailable, e.g. because they have children they have to look after, or because they need to go to work to earn funds to pay for the care, or because they are badly needed elsewhere, etc.
- Because carers don't care enough:
  The others may wish to use the DC in order to enable them to avoid tasks that they find distasteful or because they have other personal preferences/priorities

Obviously these goals can in some cases be somewhat cynical: The main beneficiaries of a DC in some situations will not be the user but others connected with the user, either because of personal relationships or because of contractual relationships (e.g. the owners of nursing homes, or retirement homes). The different motives are not necessarily all sharply distinguishable. There may be fuzzy intermediate cases, including mixed motives.

2 Categories of DC use and design that interest me.

I work on robotics, not in order to produce useful machines, but because that's the best way of addressing many old philosophical problems, though it is also possible for such research to produce useful results. However, I have no interest in making or using machines with the "engaging" functions described previously, i.e. those described earlier as: toys, engagers and dummy-humans. I have no objection to others building these things – for the right motives, though I don't think I would ever want to use them myself. E.g. I have never liked most computer games and I intensely dislike pseudo-human interfaces, with smiling, nodding, expressions, moving eyes, "emotional" voices, etc., and I hope nobody ever expects me to put up with such things if I become disabled enough to need a DC. Others, of course, have different preferences.

The rest of this paper is concerned only with the problems of producing DCs with enabling functions. Really useful general-purpose DCs are very difficult and way out of reach in the foreseeable future, for reasons that do not seem to be widely understood and which I shall try to explain. The arguments are not the same as those used by opponents of AI: I am not claiming that it is impossible to produce human-like machines, or that human-like machines cannot be based on digital computers. I am claiming that the products of evolution and individual human development include many competences that have a depth and complexity that needs to be understood much better if we are to produce really good digital companions of the enabling type. It will be much harder to achieve the enabling functions than most people realise, because it will be necessary to provide some of those competences that are not yet understood. The *currently* most popular AI techniques for competent machines, making heavy use of data-mining and statistical learning, are inadequate for the task.

Moreover, the most effective DCs will fall into the category described above as "caring developing helpers", and the production of machines that are capable of caring raises some ethical issues that will be discussed briefly, below.

3 Problems of achieving the enabling functions

My claim: The detailed requirements for DCs to meet the enabling/helping specification are not at

all obvious, and have implications that make the design task very difficult in ways that have not been noticed. Building such systems will require a deep new understanding of some hitherto unexplained human competences, though perhaps they will eventually be achieved if we analyse the problems properly.

3.1 Kitchen mishaps

Many of the things that crop up will concern physical objects and physical problems. Someone I know knocked over a nearly full coffee filter close to the base of a cordless kettle. This caused the residual current device in the fuse box under the stairs to trip, removing power from many devices in the house. Fortunately she knew what to do, unplugged the kettle and quickly restored the power. However, she was not sure whether it was safe to use the kettle after draining the base, and when she tried it later the RCD tripped again, leaving her wondering whether it would ever be safe to try again, or whether she should buy a new kettle. In fact it proved possible to open the base, dry it thoroughly, then use it as before.

Should a DC be able to give helpful advice in such a situation? Would linguistic interaction suffice? How? Will cameras and visual capabilities be provided? People who work on language understanding often wrongly assume that providing 3-D visual capabilities will be easier, whereas very little progress has been made in understanding and simulating human-like 3-D vision and spatial understanding, which is far, far more than recognising things. Human vision includes a wide and deep collection of competences, and ontologies. Many researchers confuse seeing with recognising, which is wrong because you can see something without recognising it (though you will unconsciously recognise various fragments of its visible surfaces, because they reappear all over the place).

3.2 Alternatives to canned responses

The kitchen mishap was just one example among a vast array of possibilities. Of course, if the designer anticipates such accidents, the DC will be able to ask a few questions and spew out relevant canned advice, and even diagrams showing how to open and dry out the flooded base.

But suppose designers had not had that foresight: What would enable the DC to give sensible advice? If the DC knew about electricity and was able to visualise the consequences of liquid pouring over the kettle base, it might be able to use a mixture of geometric and logical reasoning creatively to reach the right conclusions. It would need to know about and be able to reason about spatial structures and the behaviour of liquids. Although Pat Hayes described the 'Naive physics' project decades ago, it has proved extremely difficult to give machines the kind of intuitive understanding required for creative problem-solving in novel physical situations. In part that is because we do not yet understand the forms of representation humans (and other animals) use for that sort of reasoning.[5]

3.3 Identifying affordances and searching for things that provide them

Understanding a human need and seeing what is and is not relevant to meeting that need may require creative recombination of prior knowledge and competences.

Suppose an elderly user finds it difficult to keep his balance in the shower when soaping his feet. He prefers taking showers to taking baths, partly because showers are cheaper. How should the DC react on hearing the problem? Should it argue for the benefits of baths? Should it send out a query to its central knowledge base asking how how people should keep their balance when washing their feet? (It might get

[5] Jackie Chappell and I have presentations on the differences between Kantian and Humean causal reasoning in natural and artificial systems: http://www.cs.bham.ac.uk/research/projects/cogaff/talks#wonac

a pointer to a school for trapeze artists.) The DC could start an investigation into local suppliers of shower seats. But what if the DC designer had not anticipated the problem? What are the requirements for the DC to be able to invent the idea of a folding seat attached to the wall of the shower, that can be temporarily lowered to enable feet to be washed safely in a sitting position? Alternatively what are the requirements for it to be able to pose a suitable query to a search engine? How will it know that safety harnesses and handrails are not good solutions?

Giving machines an understanding of physical and geometrical shapes, processes and causal interactions of kinds that occur in an ordinary house is currently far beyond the state of the art. Compare the 'Robocup@Home' challenge, still in its infancy[6] Major breakthroughs of unforeseen kinds will be required for progress to be made, especially breakthroughs in vision and understanding of 3-D spatial structures *and processes*. Some simple examples of the required competences are discussed in a draft online discussion paper on "Predicting Affordance Changes"[7]. Of course, one response to all this would be to aim for much simpler and more easily attainable competences, such as simply providing entertainment. But let's consider requirements for the harder tasks.

3.4 More abstract problems

Sometimes the DC will need a creative and flexible understanding of human relationships and concerns, in addition to physical matters. Suppose the user U is an intense atheist, and while trawling for information about U's siblings the DC finds that U's brother has written a blog entry supporting creative design theory, or discovers that one of U's old friends has been converted to Islam and is training to be a Mullah. How should the DC react? Compare discovering that the sibling has written a blog entry recommending a new detective novel he has read, or discovering that the old friend is taking classes in cookery. What about getting evidence suggesting that U's spouse had been had had an affair with a friend long ago? Could it reason that telling U about it might assuage guilt about an affair U had had?

Should the DC care about emotional responses that news items may produce? How will it work out when to be careful? Where will its goals come from? (More on that later.)

4 Is the solution statistical?

The current dominant approach to developing language understanders and advice givers involves mining large corpora using sophisticated statistical pattern extraction and matching. This is much easier than trying to develop a structure-based understander and reasoner, and can give superficially successful results, depending on the size and variety of the corpus and the variety of tests. But the method is inherently broken because as sentences get longer, or semantic structures get more complex, or physical situations get more complex, the probability of encountering recorded examples close to them falls very quickly. Then a helper must use deep general knowledge to solve a novel problem creatively, often using non-linguistic context to interpret many of the linguistic constructs (http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0605). Some machines can already do creative reasoning in restricted domains, e.g. planning and mathematical reasoning, but they are still very limited.

[6] http://www.ai.rug.nl/robocupathome/

[7] Available online at `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0702`

4.1 Why do statistics-based approaches work at all?

The behaviour of any intelligent system, or collection of individuals, will leave traces that may have re-usable features, and the larger the set the more re-usable items it is likely to contain – up to a point. For instance it may not provide items relevant to new technological, or cultural developments or to highly improbable but perfectly possible physical configurations and processes. So any such collection of traces will have limited uses, and going beyond those uses will require something like the power of the system that generated the original behaviours.

In humans (and some other animals), there are skills that make use of deep generative competences whose application requires relatively slow, creative, problem solving, e.g. planning routes. But practice in using such a competence can train powerful associative learning mechanisms that compile and store many partial solutions matched to specific contexts (environment and goals). As that store of partial solutions (traces of past structure-creation) grows, it covers more everyday applications of the competence, and allows fast and fluent responses.

A statistical AI system that cannot generate the data can infer those partial solutions from large amounts of data. But because the result is just a collection of partial solutions it will always have severely bounded applicability compared with humans, and will not be extendable in the way human competences are.

Moreover, if trained only on text it will have no comprehension of non-linguistic context. Dealing with novel problems and situations requires different mechanisms that support creative development of novel solutions.

If the deeper, more general, slower, competence is not available, wrong extrapolations can be made, inappropriate matches will not be recognised, new situations cannot be dealt with properly and further learning will be very limited, or at least very slow. In humans the two systems work together to provide a combination of fluency and generality. (Not just in linguistic competence, but in many other domains.)

Occasionally I meet students who manage to impress some of their tutors because they have learnt masses of shallow, brittle, superficially correct patterns that they can string together – without understanding what they are saying. They function like corpus-based AI systems: Not much good as (academic) companions.

4.2 What's needed

Before human toddlers learn to talk they have already acquired deep, reusable structural information about their environment and about how people work. They cannot talk but they can see, plan, be puzzled, want things, and act purposefully. In the first to or three years of life they are constantly learning about large numbers of affordances related to various kinds of processes involving objects and situations in the environment. They are also learning about epistemic affordances – learning to discern situations that do and do not provide good task-specific information.

They have something to communicate about. That pre-linguistic competence grows faster with the aid of language, but must be based on a prior, internal, formal 'linguistic' competence using forms of representation with structural variability and (context-sensitive) compositional semantics. We have begun to call these generalised languages, which can occur inside the head or in some public communication "Generalised languages" or "GL"s in (Sloman & Chappell, 2007) also available here: `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#glang`.

The use of GLs enables very young children to learn any human language and to develop in many cultures. DCs without a similar pre-communicative basis for their communicative competences are likely to remain shallow, brittle and dependent on pre-learnt patterns or rules for every task.

Perhaps, like humans (and some other altricial species), they can escape these limitations if they start with a partly 'genetically' determined collection of meta-competences that continually drive the acquisition of new competences building on previous knowledge and previous competences: a process that continues throughout life. The biologically general mechanisms that enable humans to grow up in a very wide variety of environments, are part of what enable us to learn about, think about, and deal with novel situations throughout life. Very little is understood about these processes, whether by neuroscientists, developmental psychologists or AI researchers, and major new advances are needed in our understanding of information-processing mechanisms. Some pointers towards future solutions are in these online presentations:
`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#compmod07`
(Mostly about 3-D vision)
`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#wonac07`
(On understanding causation)
`http://www.cs.bham.ac.uk/research/projects/cosy/photos/crane/`
(On seeing a child's toys.)

A DC lacking similar mechanisms and a similar deep understanding of our environment may cope over a wide range of circumstances that it has been trained or programmed to cope with and then fail catastrophically in some novel situation. Can we take the risk? Would you trust your child with one?

5 Can it be done?

Producing a DC of the desired type may not be impossible, but is much harder than most people realise and cannot be achieved by currently available learning mechanisms. (Unless there is something available that I don't know about). Solving the problems will include:
(a) Learning more about the forms of representation and the knowledge, competences and meta-competences present in prelinguistic children who can interact in rich and productive ways with many aspects of their physical and social environment, thereby continually learning more about the environment, including substantively extending their ontologies. Since some of the competences are shared with other animals they cannot *depend* on human language, though human language depends on them. However we know very little about those mechanisms and are still far from being able to implement them.

(b) When we know what component competences and forms of representation are required, and what sorts of biological and artificial mechanisms can support them, we shall also have to devise a *self-extending architecture* which combines them all and allows them to interact with each other, and with the environment in many different ways, including ways that produce growth and development of the whole system, and also including sources of motivation that are appropriate for a system that can take initiatives in social interactions. No suggestions I have seen for architectures for intelligent agents, come close to requirements for this. (Minsky's *Emotion machine*, takes some important steps.)

I suggest the only reliable way to meet these objectives is to understand and replicate, and later on to build on some of the generic capabilities of a typical young human child, including the ability to want to help.

6 Rights of intelligent machines

If providing effective companionship requires intelligent machines to be able to develop their own goals, values, preferences, attachments etc., including really *wanting* to help and please their owners, then if some of them develop in ways we don't intend, will they not have the right to have their desires considered, in the same way our children do if they develop in ways their parents don't intend? I discussed this briefly in

the epilogue to (Sloman, 1978), available here: http://www.cs.bham.ac.uk/research/projects/cogaff/crp/epilogue.html
I have also argued that Asimov's laws of robotics are immoral, because their formulation is excessively hostile to robots.

7 Risks of premature advertising

I worry that most of the people likely to be interested in this kind of workshop will want to start designing intelligent and supportive interfaces without waiting for the above problems to be solved, and I think that will achieve little of lasting value because the machines will be too shallow and brittle, and potentially even dangerous – though they may handle large numbers of special cases impressively. If naive users start testing them, and stumble across catastrophic failures that could give the whole field a very bad name.

Some related online papers and presentations

http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0703
: Computational Cognitive Epigenetics

http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0704
: Diversity of Developmental Trajectories in Natural and Artificial Intelligence

http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#cafe04
: Do machines, natural or artificial, really need emotions?

References

Jablonka, E., & Lamb, M. J. (2005). *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Cambridge MA: MIT Press.

Sloman, A. (1971). Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence. In *Proc 2nd ijcai* (pp. 209–226). London: William Kaufmann. (http://www.cs.bham.ac.uk/research/cogaff/04.html#200407)

Sloman, A. (1978). *The computer revolution in philosophy*. Hassocks, Sussex: Harvester Press (and Humanities Press). (http://www.cs.bham.ac.uk/research/cogaff/crp)

Sloman, A., & Chappell, J. (2007). Computational Cognitive Epigenetics (Commentary on (Jablonka & Lamb, 2005)). *Behavioral and Brain Sciences*, *30*(4). (http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0703)