Presentation at University of Oxford Internet Institute, 26 Oct 2007 workshop on Artificial Companions in Society: Perspectives on the Present and Future

Oxford 25th–26th October, 2007

Position Paper:

Requirements for Digital Companions and Their Implications

It’s harder than you think

Aaron Sloman
http://www.cs.bham.ac.uk/~axs/

These slides will be in my ‘talks’ directory:
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#oii

The position paper for the meeting will be available here
http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-oii-2007.html

What are we trying to do?

The workshop presupposes that development of Digital Companions (DCs) will happen, and asks about ethical, psychological, social and legal consequences.

The invitation states that 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.

My claim:

The detailed requirements for DCs to meet that specification are not at all obvious, and will be found to have implications that make the design task very difficult in ways that have not been noticed, though perhaps not impossible if we analyse the problems properly.

Why?

Motivations for doing these things can vary.

Some are better than others.

Categories of goals for DC-designers

People working on digital companions may be trying to produce something that is intended to provide one or more of the following kinds of function:

Engaging functions

- TOYS:
  a toy or entertainment for occasional use (compare computer games, music CDs)
- ENGAGERS:
  something engaging that is used regularly to provide interest, rich and deep enjoyment or a feeling of companionship (compare pets, musical instruments)
- DUMMY-HUMANS (pacifiers?):
  something engaging that is regarded by the user as being like another caring, feeling individual with which 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.

Enabling functions

- HELPERS:
  something that can reliably provide help and advice that meets practical everyday needs as well as occasional unexpected problems. (Helpers for patients in therapy are a special case.)
- DEVELOPING HELPERS:
  a helper that is capable over time of developing a human-like understanding of the user's environment, needs, preferences, values, knowledge, capabilities
- CARING DEVELOPING HELPERS:
  a developing helper that grows to care about the user and really wants to help when things go wrong or risk going wrong, and wants to find out how best to achieve these goals.
  (since it will be impossible for designers to anticipate everything.)

Meta goals – Why do it?

If we ask WHY the goals are being pursued we can distinguish purposes that concern not only the user but also other humans who care about or have responsibilities to the user such as:

- User wants (or needs) it:

  the user may sincerely prefer to be helped by a DC so as not to have to impose on other humans and the others involved may respect that preference

  even if they would prefer to provide the care themselves.

- Because caring others 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. they may have children they have to look after,
  or may need to go to work to earn funds to pay for the care, or maybe 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).

NB: THERE ARE FUZZY INTERMEDIATE CASES.

Needs and wants – and special cases

The concept of “needing” something is very different from the concept of “wanting” something, and this can lead to conceptual and ethical problems regarding provision of DCs.

An individual is not always the best authority regarding what he/she needs even though what is wanted is in general something determined by the wanter.

Defining what “need” means is not simple. Some may be tempted to define it in terms of whatever makes it possible, or makes it easier, to achieve what one wants.

That analysis is not helpful in the cases of: other animals, very young children, patients who are in a coma, and some people with cognitive abnormalities, that prevent them understanding or wanting some of the things that are required for their health or well-being.

There are, however, some relatively clear cases: e.g. someone who has been injured and needs and wants help overcoming the consequences of the injury, or who has had a stroke and needs and wants help restoring some lost functionality, such as control of arm or hand movements, or balance.

It may be possible to develop some machines for which no more is claimed than that they help with very specific therapeutic procedures, e.g. providing or monitoring exercises performed by a patient. However, I suspect it would be very easy to give uninformed users exaggerated expectations regarding the capabilities of such a DC, so marketing and training would require great care.

For more on this see

“User-Robot Personality Matching and Assistive Robot Behavior Adaptation for Post-Stroke Rehabilitation Therapy”, by Adriana Tapus, Cristian Tapus, and Maja J Mataric
Intelligent Service Robotics Journal, 1(2):169-183, April 2008
`http://www-robotics.usc.edu/~tapus/publications/tapus_JISR2008.pdf`

Categories of DC use 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.

I have no interest in making machines with the "engaging" functions described previously, i.e. those described earlier as:

- TOYS:
- ENGAGERS:
- DUMMY-HUMANS (pacifiers?):

I have no objection to others doing these things – for the right motives.

I don't think I would ever want to use them myself.

People have different preferences

E.g. I have never liked most computer games and I dislike pseudo-human interfaces **intensely**.
If my bank introduced one that I could not turn off I would switch banks.

I would, however, be interested in the following but they are very difficult and way out of reach in the foreseeable future (and I'll try to explain why):

- HELPERS:
- DEVELOPING HELPERS:
- CARING DEVELOPING HELPERS:

(Remember Maggie's question: is it REALLY doing X?)

Why out of reach?

Current AI is nowhere near producing machines that meet the following:

- The need to understand the physical environment

  coffee spilt on kettle base, washing feet in shower.
- The need to understand human minds – at least as well as humans do
- The need to base deep motives on those

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

Those capabilities build on a very rich biological genetic heritage, though not necessarily a normal human body: compare humans born limbless – their brains matter more than their bodies.

Current statistical programming methods will always be limited, fragile and unreliable (except for very restricted applications).

NOTE: if we make machines that can come to care, then we need to take account of their desires and preferences: anything with desires should have rights: treating them as slaves would be highly immoral.

(As noted in the Epilogue to *The Computer Revolution in Philosophy* (1978), now online here: `http://www.cs.bham.ac.uk/research/projects/cogaff/crp/`)

Example: A kitchen mishap

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 central fuse box to trip, removing power from many devices in the house.

She knew what to do: unplugged the base, reset the RCD, and quickly restored power.

But 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?

Many people who do not work on vision wrongly assume that providing 3-D visual capabilities will be relatively easy (e.g. compared with understanding language).
In fact 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.

Canned responses – and intelligent responses

- That was just one example among a vast array of possibilities that will vary from culture to culture, from household to household within a culture and from time to time in any household, as the people and things in the house change.
- 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 – which involves both Humean and Kantian causal understanding.
    See these papers and presentations on understanding causation with Jackie Chappell:
    `http://www.cs.bham.ac.uk/research/projects/cogaff/talks#wonac`

Understanding affordances

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 a pointer to a school for trapeze artists.)

The DC could start an investigation into local suppliers of shower seats: but that requires working out that a seat in the shower would solve the problem, despite the fact that showers typically do not and should not have seats.

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 lowered temporarily 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?

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 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?

The state of the art in commonsense physics

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:

`http://www.ai.rug.nl/robocupathome/`

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, including causal interactions

Of course, one response to all this would be to aim for much simpler and more easily attainable competences, such as providing entertainment.

(Not one of my interests.)

But let’s consider requirements for the harder tasks.

Is the solution statistical?

The current dominant approach to developing language understanders, advice givers, and even some engaging interfaces, involves mining large corpora using sophisticated statistical pattern extraction and matching, to find re-usable patterns.

This is 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 rapidly to near zero.

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

See

`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0605`

Spatial prepositions as higher order functions:
And implications of Grice's theory for evolution of language.

`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0601`

Orthogonal Recombinable Competences Acquired by Altricial Species
(Blankets, string, and plywood)

Admittedly, some machines can already do creative reasoning in restricted domains, e.g. planning, mathematical reasoning, but they are very limited.

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.**

How humans use statistical traces

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.

Frequent use of such competences trains powerful 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 in more contexts and tasks.

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.

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.

(Many jokes depend on that.)

If the deeper, more general, slower, competence is not available when stored patterns are inadequate, 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, and probably some other animals, the two systems work together to provide a combination of fluency and generality. (Not just in linguistic competence, but in many other domains.)

Where does the human power come from?

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: 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.

This enables them 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), DCs 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.

There are several slide presentations related to this on my “talks” web site, e.g.
`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#glang` (and others)

What mechanisms?

Very little is understood about these processes, whether by neuroscientists, developmental psychologists or AI researchers.

Major new advances are needed in our understanding of information-processing mechanisms.

Some pointers to requirements for future solutions are in these online presentations:

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#compmod07`
(Mostly about 3-D vision – especially seeing processes.)

`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, because it either blindly applies inappropriate statistical generalisations, or because it finds no applicable pattern, and freezes.

Can we take the risk?

Would you trust your one of them to look after your child?
Or your elderly parent?
Or you?

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.

(We are walking on a small subset of the problems in the EU CoSy robotic project.
`http://www.cs.bham.ac.uk/research/projects/cosy/PlayMate-start.html`)

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?

See also
`http://www.cs.bham.ac.uk/research/projects/cogaff/crp/epilogue.html`

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 they 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.

It is possible that some of the digital helpers with very narrowly defined functions, e.g. robots designed to support specific forms of therapy, such as therapy involving exercises, are not problematic in this way because their advertised competences are very precisely defined.

Some related online papers and presentations

To be expanded:

`http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0703`
Computational Cognitive Epigenetics (With J. Chappell in BBS soon.)

`http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-aaai-representation.pdf`
Diversity of Developmental Trajectories in Natural and Artificial Intelligence
(For AAAI Fall symposium 2007)

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

THANK YOU!

For the importance of virtual machines and supervenience see

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#bielefeld`

For ideas about how machines or animals can use symbols to refer to unobservable entities see

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#models`

Introduction to key ideas of semantic models, implicit definitions and symbol tethering

For an argument that internal generalised languages (GLs) preceded use of external languages for communication, both in evolution and in development, see

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#glang`

What evolved first: Languages for communicating, or languages for thinking
(Generalised Languages: GLs) ?

Additional papers and presentations

`http://www.cs.bham.ac.uk/research/projects/cogaff/talks/`
`http://www.cs.bham.ac.uk/research/projects/cosy/papers/`