Is a Companion a distinctive kind of relationship with a machine?

Yorick Wilks

University of Oxford

I start from the perspective of the EC COMPANIONS project, and set out its aim to model a new kind of human-computer relationship based on long-term interaction, with some tasks involved although the Companion is not inherently task-based, since there need be no stopping point to its conversation. Some demonstration of its functionality will be given but the main purpose here is an analysis of what it is people might want from such a relationship and what evidence we have for whatever we conclude. Is politeness important? Is an attempt at emotional sympathy important or achievable? Does a user want a consistent personality in a Companion or a variety of personalities? Should we be talking more in terms of a "cognitive prosthesis (or orthosis)?" — something to extract, organize, and locate the user's knowledge or personal information — rather than attitudes?

Introduction

The paper assumes that artificial Companions are on their way, and the interesting issues concern what they will be like. I am assuming two things here: first, that the robotic aspect is interesting but dispensable for this discussion. Dautenhahn has established interesting facts such as that people would prefer that robots approached them from the side rather than head on (Walters et al., 2009) and of course there will always be people who want things brought to them rather than getting up out of their chairs. But I will be concerned here with aspects of Companions such that embodiment is a secondary matter, provided they can converse with an owner and can reach out to the world via the internet for information and to establish action and control. Whether they are implemented as mobile phones, moving robots with prostheses, or just "warm furry handbags" with Wifi, is irrelevant to what I shall discuss here, though I shall often assume they can assume visual shape on a screen when necessary, but that is far short of a robot in any full sense.

Secondly, and still by way of scene setting, it is convenient to distinguish Companions from both (a) conversational internet agents that carry out specific tasks, such as the train and plane scheduling and ticket ordering speech dialogue applications back to the MIT ATIS systems (Zue et al., 1992), and also from (b) descendants of the early chatbots PARRY and ELIZA, the best of which compete annually in the Loebner competition (Loebner). These have essentially no memory or knowledge but are simple finite state response sets, although ELIZA had primitive "scripts" giving some context, and PARRY (Colby, 1971) had parameters like FEAR and ANGER that changed with the conversation and determined which reply was selected at a given point.

I take the distinguishing features of a Companion agent to be:

1) that it has no central or over-riding task and there is no point at which its conversation is complete or has to stop, although it may have some tasks it carries out in the course of conversation;
2) That it should be capable of a sustained discourse over a long-period, possibly ideally the whole life-time of its principal user;
3) It is essentially the Companion of a particular individual, its principal user, about whom it knows a great deal of personal knowledge, and whose interests it serves — it could, in principle, contain all the information associated with a whole life (in the sense of the Memories for Life consortium XXX);
4) It establishes some form of relationship with that user, if that is appropriate, which would have aspects associated with the term "emotion";
5) It is not essentially an internet agent or interface, but since it will have to have access

to the internet for information (including the whole-life information about its user) and to act in the world (as it is not a robot), we may as well assume its internet agent status, and so it should have, so far as possible, access to open internet knowledge sources.

By separating a Companion conceptually from both a task-based system and a chatbot, we immediately lose access to the two evaluation paradigms associated with those models of computer dialogue: the first in terms of task-completion (stickiness, timing, task success etc.) and the latter (usually) in terms of distinguishability from some set of human interlocutors. There is, at the moment, no clear evaluation paradigm for a Companion, even if we had one to evaluate, although there are ideas for creating one (Webb et al., 2010) and some of these have been applied to the first demonstrators from the COMPANIONS (Wilks, 2006) project itself.

Given this narrowing of focus in this paper, what questions then arise and what choices does that leave open? Here are some obvious questions that have arisen in the literature:

i) What aspects of a relationship should one aim at with a Companion, in terms of such conventional categories as emotion, politeness, affection etc.?

ii) Even if it is not a robot, in the sense of a free-moving entity, should it have a screen, and should it have a visible avatar for communication, whether human, animal or abstract?

iii) Does a Companion need a voice or could communication be by typing (such as on a mobile phone, laptop or PC)?

iv) Need it have one identifiable personality, or perhaps several, and should the user be able to choose the Companion's personality or shift between them if there are several? More generally, are the answers to these questions, and the settings and constraints they imply, dependent on the type of Companion — the domain or setting into which it is to be placed, or is there only one type of Companion subject to general constraints?

v) Does the Companion have any goals of its own, beyond carrying out a user's commands, if that is possible: should there be other overriding ethical constraints on what can be commanded, such as avoiding harm to the user, even if requested? Should there be ethical constrains *on the user* as to how the Companion can be treated?

vi) What safeguards are there for the information content of such a Companion, in the sense of controlling access to its contents for the state or a company, and how should a user best provide for its disposal in case of his/her own death or incapacity?

vii) What if anything does a Companion have to *know* to be plausible, and does it need a certain level of inference and memory capacity over the material of past conversations with the user?

Let us take these issues in turn.

1. Emotion, politeness and affection

Cheepen and Monaghan (1997) presented results some thirteen years ago that customers of some automata, such as ATMs, are repelled by excessive politeness and endless repetitions of ”thank you for using our service”, because they know they are dealing with a machine and such feigned sincerity is inappropriate. This suggests that politeness is very much a matter of judgment in certain situations, just as it is with humans, where inappropriate politeness is often encountered. Wallis (Wallis et al., 2001) has reported results that many find computer conversationalists “chippy” or “cocky” and suggests that this should be avoided as it breeds hostility on the part of users; he believes this is always a major risk in human-machine interactions.

We know, since the original work of Nass (Reeves and Nass, 1996) and colleagues that people will display some level of feeling for the simplest machines, even PCs in his original experiments, and Levy (2007) has argued persuasively that the trend seems to be towards high levels of “affectionate” relationships with machines in the next decades, as realistic hardware and sophisticated speech generation make machine interlocutors increasingly lifelike. However, much of this work is about human psychology, faced with entities known to be artificial, and does not bear directly on the issue of whether Companions should attempt to detect emotion in what they hear from us, or attempt to generate it in what they say back.

The AI area of “emotion and machines” is confused and contradictory: it has established itself as more than an eccentric minority taste, but as yet has nothing concrete to show beyond some better than random algorithms for detecting “sentiment” in incoming text (e. g. Wiebe et al., 2005), but even there its success is dependent on effective content extraction techniques. This work began as “content analysis” (Krippendorff, 2004) at the Harvard psychology department many years ago and, while prose texts may offer enough length to enable a measure of sentiment to be assessed, this is not always the case with short dialogue turns. That technology rested almost entirely on the supposed sentiment value of individual words, which ignores the fact that their value is content dependent. “Cancer” may be marked as negative word but the utterance “I have found a cure for cancer” is presumably positive and detecting the appropriate response to that rests on the ability to do information extraction beyond single terms. Failure to observe this has led to many of the classic foolishnesses of chatbots such as congratulating people on the death of their relatives, and so on.

At deeper levels, there are conflicting theories of emotion for automata, not all of which are consistent and which apply only in limited ranges of discourse. So, for example, the classic theory that emotion is a response to the failure and success of the machine's plans (e. g. Marsella and Gratch, 2003) covers only those situations that are clearly plan driven and, as we noted, Companionship dialogue is not always closely related to plans and tasks. "Dimensional" theories (Cowie et al., 2001, following Wundt, 1913), display emotions along dimensions marked with opposed qualities (such as positive-negative) and normally distribute across the space emotion "primitives", such as FEAR, and these normally assigned by manual tagging, and they this rest, like the text-sentiment theories above, on pre-tagging and any learning based on them, of the sort that all learning engines perform over tag distributions (e. g. Ciravegna et al., 2004). The problem with this is that tagging for "COMPANY" or "TEMPERATURE" (in classic NLP) is a quite different task from tagging for "FEAR" and "ANGER". These latter terms are not, and probably cannot be, analyzed but rest on the commonsense intuitions of the tagger, which may vary very much from person to person — they have very low consilience between taggers.

All this makes many emotion theories look primitive in terms of developments in AI and NLP elsewhere. Appraisal Theory (Scherer et al, 2008) seeks to explain why individuals can have quite different emotional reactions to similar situations because they have appraised them differently, e. g. a death welcomed or regretted. Appraisal can also be of the performance of planned activities, in which case this theory approximates to the plan-based one mentioned above. The theory itself, like all such theories, has a large-commonsense component, and the issue for computational implementation is how, in assessing the emotional state of the Companion's user to make such concepts quantitatively evaluable. If the Companion conducts long conversations with a user about his or her life and, as in the case of the Senior Companion prototype (http://www.youtube.com/watch?v=-Xx5hgjD-Mw) which discusses photo images, then one might expect there to be ample opportunity to assess the user's appraisal of, say, a funeral or wedding by means of the application of the sentiment extraction techniques to what is said in the presence of the relevant image. In so far as a Companion can be said to have over-arching goals, such as keeping the user happy then, to that degree, it is not difficult to envisage methods (again based on estimates of the happiness, or otherwise, of the user's utterances) for self-appraisal by the Companion of its own performance and some consequent causal link to generated demonstrations of its own emotions of satisfaction or guilt.

Also relevant to what a Companion should be is the "Affective Loop" (AL) paradigm (Höök, 2004) which, like most of the theories of emotion discussed, and as John Wisdom once said of philosophical discoveries, are often the "running of a platitude up a flagpole": but AL is a useful corrective to some of the claims above and is intended essentially for computational implementation. It emphasizes:

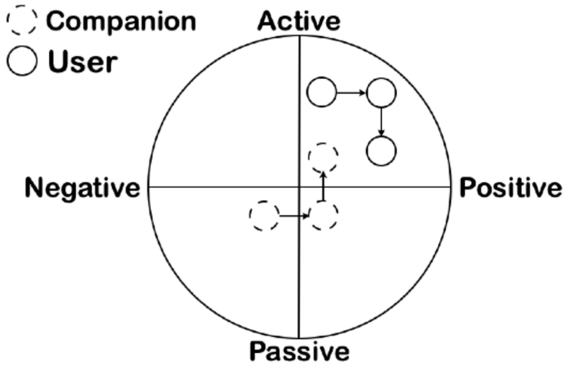
- that there is a natural "feedback loop" involved in emotional interaction between parties and which is essential to any model
- but that emotional interaction and feedback should not be thought of as a matter of information transfer.
- it is much concerned with design, and the design of multimodal interactions of the display of color and sound — it is not essentially concerned with emotional language
- it emphases the relative vacuity of emotional labels or terms, as we did above, and peoples' intuitive understanding of them.

The notion of feedback is an old one going back to cybernetic ideas and in particular to Wiener's notion that activities like walking are only possible because of constant information feedback from the "servo" muscles in contact with the ground to the brain. Wiener was emphasizing information feedback, as opposed to the "haptic" transfer from muscles, but in a computational paradigm everything must at some stage bottom out in information. Speech act theory, too, arose from considerations of human interaction that were not based on conveying information in propositions, but rather "intentional" commitments, but those again have only been implementable in computers as forms of information.

Many of Höök's examples involve multi-modal devices such as smart phones where non-verbal signals are sent to create attitudes and feelings, or to signal those of the sender. The Nabaztag rabbit toy, originally used by the COMPANIONS project as an interface (http://www.nabaztag.com/en/index.html), in its original design glowed in a number of colors to indicate the feelings of the sender (e. g. blue for "sad") and two Nabaztags and their respective senders would be a paradigm AL in Höök's sense. There are many wholly conventionalised verbal feedback loops that cannot be divorced from emotion — certainly if a respondent fails to supply the correct response, from "How do you do" and "Good morning" in English to the potentially infinite "danke, bitte, danke, bitte…" cycle of giving thanks in German.

The importance of AL is that it makes emotion central, not peripheral, to communication and relationships and does not make language behavior central to emotional communication. Everyone knows that in relationships with pets, a central relationship for many people, this is the case: strong emotions are aroused, as well as consequent actions of e. g. stroking, but there is no verbal content. There have been a number of Japanese pet robot implementations, such as wriggly seal-like creatures with dozens of servo motors to give a life-like feel, and there is no doubt that a real form of human relationship is being modeled. Companions were always designed with the pet analogy in mind, as in the phrase used early in the project of a Companion "being like a furry handbag", though language was always believed essential to the project.

In speaking of "language" and Companions, we have so far ignored speech, although that is a communication mode in which a great deal has been done to identify and, more recently, generate, emotion-bearing components (Luneski et al., 2008). Elements of the above approaches can be found in the work of Worgan and Moore (see e. g. Wilks et al., 2010), within the COMPANIONS project, where the is the same commitment to the centrality of emotion in the communication process, but in a form focusing on an integration of speech and language (rather than visual and design) technologies. The claim, not yet implemented, was conceived within the COMPANIONS project as a layer in a dialogue manager over and above local response management but one which would seek to navigate the whole conversation across a two-dimensional space onto which Companion and user are mapped using continuous values (rather than discrete values corresponding to primitive but unexplained emotional terms) but in such a way as to both respond to the a user's demonstrated emotion appropriately, but also — again, if appropriate or chosen by the user — to draw the user back to other more positive emotional areas of the two-dimensional space. It is not yet clear what the right mechanism should be for the integration of this "landscape" global emotion-based dialogue manager should be with the local dialogue management that generates responses and alters the world context: in the Senior Companion this last was sophisticated stack of networks (see Wilks et al., in press). In some sense, we are just looking for a modern and defensible interface to replace what PARRY had in simple form in 1971 when the sum of two emotion parameters determined which response to select from a stack of alternatives.



This last is a high level issue to be settled in a Companion's architecture and also, perhaps, to be under the control of the user, namely: should a Companion invariably try to cheer a user up if miserable — which is trying to "move" the user to the most naturally desirable (i. e. the top-right) quadrant of the space — or, rather, to track to the part of the space where the user is deemed to be and stay there in roughly the same emotional location — i. e. be sad with a sad user and happy with a happy one? There is no general answer to this question and, indeed, in an ideal Companion, which tracking method should be used would itself be a conversation topic e. g. "Do you want me to cheer you up or would you rather stay miserable?". In that sense, an AL is a platitude and everything depends on what kind of loop it is to be — itself a matter for negotiation.

2. What should a Companion look like?

I confess to an affection for a faceless Companion — the proverbial furry handbag, warm and light to carry, chatty but with full internet access and probably no screen. However, this may be a minority taste; after all, such a Companion could always take control of a nearby screen or a phone if it needed to show anything. If there is to be a face, the question of the "uncanny valley effect" always comes up, where it is argued that users are more uneasy the more something is very like ourselves (Mori, 1970). I personally do not feel this, indeed it cannot in principle apply to an avatar so good that one cannot be sure it is artificial, which is what I feel about the *Emily* from Manchester:

http://www.youtube.com/watch?v=UYgLFt5wfP4&feature=player_embedded#

http://www.surrealaward.com/avatar/3ddigital12.shtml

On the other hand, if the quality is not good, and in particular if the lip synch is not perfect, it may be better to go for an abstract avatar — the Companions logo was chosen with that in mind, and without a mouth at all. Non-human avatars seem to avoid some of the problems that arise with valleys and mixed feelings generally, and the best Companions demonstration video so far features Wigdog, a dog in a wig, who seems pretty popular:

http://www.youtube.com/watch?v=-Xx5hgjD-Mw

It may be worth making here a small clarification about the word "avatar" that sometimes distorts discussion in these areas: those working in computing the human-machine interface often use the word to mean any screen form, usually two-dimensional, that simulates a human being, but not any particular human being. On the other hand, in the virtual reality and game worlds, such as Second Life (http://secondlife.com/), an avatar is a manifestation of a particular human being, an alternative identity that may or may not be similar to the owner in age, sex, appearance etc. These are importantly different notions and confusion can arise when they are conflated or confused: in current COMPANIONS project demonstrations, for example, a number of avatars in the first sense are used to present the Companion's conversation on a computer or mobile phone screen. However, in the case of a long-term computer Companion that could elicit, through prolonged reminiscence, details of its owner's life and perhaps train its own voice in imitation, since research shows that more successful computer conversationalists are as like their owners as possible. One might then approach the point where a Companion could approximate to the second sense of "avatar" above, namely an avatar of its owner, which it would progressively resemble, as dogs are said to do.

3. Voice or Typing to communicate with a Companion?

At the moment the limitation on the use of voice is twofold: first, although trained ASR for a single user — such as a Companion's user — is now very good and up in the high 90%, it still introduces uncertainty into understanding an utterance that is far greater than that of spelling errors. Secondly, it is currently not possible to store sufficient ASR software locally on a mobile phone to recognize a large vocabulary in real time; access to a remote server takes additional time and can be subject to fluctuations and delays. All of which suggests that typed input — though not TTS output — from a web-based Companion may have to use typed input in the immediate future, which is no problem for most mobile phone users who have come to find typed chat perfectly natural. However, this is almost certainly only a transitory delay as mobile RAM increases rapidly and the problem should not determine research decisions — there is no doubt that voice will move back to the centre of communication once storage and access size have grown by another order of magnitude.

4. One Companion personality or several?

Some (e. g. Pulman, in Wilks, 2010) have argued that having a consistent personality is a condition on Companionhood, but one could differ and argue that, although that is true of people — multiple personalities being a classic psychosis — there is no reason why we should expect this of a Companion. Perhaps a Companion should have a personality adapted to its particular relationship to a user at a given moment: Lowe (in Wilks, 2010) has pointed out that one might want a Companion to function as, say, a gym trainer, in which case a rather harsh attitude on the part of the Companion might well be the best one. If a Companion's emotional attitude were to (figuratively) move across a two dimensional emotion space (see diagram above) imitating or correcting what it perceived to be the user's state over time (as Worgan, see above, has proposed), then that shift in attitude might well seem to be the product of different personalities, as it sometimes can with humans.

It might be better, *pace* Pulman, to give a user access to, and some control over, the display of a multiple-personality Companion, something one could think of as an "agency" of Companions, rather than a single "agent", all of which shared access to the same knowledge of the world and of the state and history of the user.

5. Ethics and goals in the Companion

The last section is very close to the question of what goals a Companion can plausibly have, beyond something very general, such as "keep the user happy and do what they ask if you can", which are goals and constraints that directly relate to the standard discussions of the ethics a robot could be considered to have, a discussion started long ago by Asimov (1975). Clearly, there will be need for a Companion to have goals to carry out specific tasks: if it is to place a restaurant table booking on the phone for a user who has just said to it "Get me a table for two tonight at Branca around 8.30" — a phone request well within the bounds of the currently achievable technology — and the Companion will first have to find the restaurant's phone number before it phones and ask about availability before choosing a reservation time. This is the standard content of goal-driven behavior, with alternatives at every stage if unexpected replies are encountered (such as the restaurant being fully booked tonight). But one does not need to consider such goals as "goals of its own" since they are inferred from what it was told and are simply assumed, as an agent or slave of the user. But a Companion that finds its user not responding after some minutes of conversation might well have to take an independent decision to call a doctor urgently, based on a stored permanent goal about danger to a user who is unable to answer but is not asleep etc.

Asimov was concerned with the ethics of the robot and its doing no harm to its users, or indeed to anyone else — even if asked to do harm explicitly. These days one might also consider the point at which ill treatment of he Companion itself might be an ethical problem for the user: again, Nass' experiments revealing feeling or sympathy even for a criticized PC suggest these will not be too far in the future.

6. Safeguards for the information content of a Companion

Data protection, privacy, or whatever term one prefers, now captures a crucial concept in the new information society. A Companion that had learned intimate details of a user's life over months or years would certainly have contents needing protection, and many forces — commercial, security, governmental, research — might well want access to it, or even to those of all the Companions in a given society. If societies move to a clear legal state where one's personal data is one's own, with the owner or originator having rights over sale and distribution of their data — which is not at all the case at the moment in most countries — then the issue of the personal data elicited by a Companion would automatically be covered.

If we ignore the issues of governments and national security — and a Companion would clearly be useful to the police when wanting to know as much as possible about a murder suspect, so that it might then be an issue of whether talking to one's Companion constituted any kind of self-incrimination, in countries where that form of communication is protected. Some might well want one's relationship to a Companion put on some basis like that of a relationship to a priest or doctor, or even to a spouse, who cannot always be forced to give evidence in common-law countries.

More realistically, a user might well want to protect parts of his or her Companion's information, or even an organized life-story based on that, from particular individuals: e. g. "this must never be told to my children, even when I am gone". It is not hard to imagine a Companion deciding whom to divulge certain things to, selecting between classes of offspring, relations, friends, colleagues etc. There will almost certainly need to be a new set of laws covering the ownership, inheritance and destruction of Companion-objects in the future.

7. What must a Companion know?

There is no clear answer to this question: dogs make excellent Companions and know nothing. More relevantly, Colby's PARRY program, the best conversationalist of its day (Colby, 1971) and possibly since, famously "knew' nothing: John McCarthy at Stanford dismissed PARRY's skills by saying :"It doesn't even know who the US President is", forgetting as he said it that most of world's population did not know that, at least at the time. On the other hand, it is hard to relate over a long term to an interlocutor who knows little or nothing and has no memory of what it or you have said in the past. It is hard to attribute personality to an entity with no memory and little or no knowledge.

Much of what a Companion knows that is personal it should elicit in conversation from its user; yet much could also be gained from publicly available sources, just as the current Senior Companion demo goes off to Facebook, independently of a conversation, to find out who its user's friends are. Current information extraction technology (e. g. Ciravegna et al., 2004) allows a reasonable job to be made of going to Wikipedia for general information when, say, a world city is mentioned; the Companion can then glean something about that city from Wikipedia and ask a relevant question such as "Did you see the Eiffel Tower when you were in Paris?" which again gives a plausible illusion of general knowledge.

John McCarthy always maintained that the real challenge for AI was not having exotic or detailed knowledge but common-sense knowledge, what exists below our levels of consciousness, such as that dropped thing fall, and fingers go into water when pushed but not into tables: all of what Hayes once called Naïve Physics (Hayes, 1978). Some of this can be coded in the inference rules a Companion will need, such as that sisters share parents, but much of it is below the level of straightforward rules, which is what led Dreyfus (1972) and others to argue that plausible AI would need the ability to learn as we do by growing up, rather than by existing forms of machine learning or hand-coding. However, the great improvements in such learning in recent years, from speech recognition to machine translation suggests that the jury is still out on this, even if the methods that have proved successful in computers are clearly not those humans themselves use.

8. A concrete Companion paradigm: the Victorian Companion

The subsections above are mini-discussions of some of the constraints on what it is to be a Companion, the subject of a recent book collection (Wilks, 2010). The upshot of those discussions is that there are many dimensions of choice, even within an agreed definition of what a Companion is to be, and they will depend on the user's tastes and needs above all. In the section that follows, I cut though the choices and make a semi-serious proposal for a model Companion, one based on a once well-known social stereotype.

In (O'Hara, in Wilks 2010) a colleague remarks that James Boswell was a clear case of the inaccurate Companion: his account of Johnson's life is engaging but probably exaggerated, yet none of that now matters. Johnson is now *Boswell's* Johnson, by and large, and his Companionship made Johnson a social property in a way he would never have been without his Companion and biographer. This observation brings out some of the complexity of Companionship, as opposed to a mere amanuensis or recording device, and its role between the merely personal and the social.

The first Artificial Companion is, of course, Frankenstein's monster in the 19C; that creature was dripping with emotions, and much concerned with its own social life:

> *Shall each man," cried he, "find a wife for his bosom, and each beast have his mate, and I be alone? I had feelings of affection, and they were requited by detestation and scorn. Man! you may hate; but beware! your hours will pass in dread and misery, and soon the bolt will fall which must ravish from you your happiness for ever (Shelley, 1831, Ch. 20).*

This is clearly not quite the product that any modern COMPANIONS project is aiming at but, before just dismissing it as an "early failed experiment", we should take seriously the possibility, already touched on above, that things may turn out differently from what we expect and Companions, however effective, may be less loved and less loveable than we might wish. Newell has argued forcefully (e. g.in Wilks, 2010) that we must actually find out what kinds of relationship people want with Companion entities, as opposed to being technologists and just deciding a priori and then building what they believe people want.

It is no longer fashionable to explore a concept by reviewing its various senses, though it is not wholly useless either: when mentioning recently that one draft website for the COMPANIONS project had the black and pink aesthetic of a porn site, I was reminded by a colleague that the main Google-sponsored Companions site still announces "14.5 million girls await your call" and it was therefore perhaps not as inappropriate as I had first thought. Yet, for many, a Companion is still, primarily, a domestic animal, and it is interesting to note the key role pet-animals still play in the arguments on what it is, in principle, to be a Companion: especially the presence of the features of memory, recognition, attention and affection, found in dogs but rarely in snakes or newts.

I would also add that pets can play a key role in arguments about responsibility and liability, issues also raised already, and that dogs, at least under English common law, offer an example of an entity with a status between that of humans and mere wild animals: that is, *ferae naturae*, such as tigers, which the common law sees essentially as machines, and anyone who keeps one is absolutely liable for the results of its actions. Companions could well come to occupy such an intermediate moral and legal position (see Wilks & Ballim, 1990), and it would not be necessary, given the precedents with pets already available in law, to deem them either mere slaves or the possessors of rights like our own. Dogs are treated by English courts as potential possessors of "character", so that a dog can be of "known bad character", as opposed to a (better) dog acting "out of character". There is no reason to believe that these pet precedents will automatically transfer to issues concerning Companions, but it is important to note that some minimal legal framework of this sort is already in place.

More seriously, and in the spirit of a priori thoughts (and what else can we have at this technological stage of development?) about what a Companion should be, I would suggest we could profitably spend a few moments reminding ourselves of the role of the Victorian lady's Companion. Forms of this role still exist, as in a recent web posting:

> *Companion Job*
> *posted: October 5, 2007, 01:11 AM*
>
> *I Am a 47 year old lady looking seeking a position as Companion to the elderly, willing to work as per your requirements.I have been doing this work for the past 11 yrs.very reliable and respectful.*
>
> *Location: New Jersey*
> *Salary/Wage: Will discuss*
> *Education: college*
> *Status: Full-time*
> *Shift: Days and Nights*

But here the role has become more closely identified with caring and the social services than would have been the case in Victorian times, where the emphasis was on company, preferably educated company and diversion, rather than care. However, this was not always a particularly desirable or even tolerable role for a woman. Fanny Burney refers to someone's Companion as a "toad-eater" which Grose (1811) glosses as:

> *A poor female relation, and humble Companion, or reduced gentlewoman, in a great family, the standing butt, on whom all kinds of practical jokes are played off, and all ill humors' vented. This appellation is derived from a mountebank's servant, on whom all experiments used to be made in public by the doctor, his master; among which was the eating of toads, formerly supposed poisonous. Swallowing toads is here figuratively meant for swallowing or putting up with insults, as disagreeable to a person of feeling as toads to the stomach.*

But one could nevertheless, and in no scientific manner, risk a listing of features of the ideal Victorian Companion:

1. Politeness
2. Discretion
3. Knowing their place
4. Dependence
5. Emotions firmly under control
6. Modesty
7. Wit
8. Cheerfulness
9. Well-informed
10. Diverting
11. Looks are irrelevant
12. Long-term relationship if possible
13. Trustworthy
14. Limited socialization between Companions permitted off-duty.

The Victorian virtue of Discretion here brings to mind the "confidant" concept that Boden (in Wilks, 2010) explicitly rejected as being a plausible one for automated Companions:

> *Most secrets are secret from some HBs [Human Beings] but not others. If two CCs [Computer Companions] were to share their HB-users' secrets with each other, how would they know which other CCs (i. e. potentially, users) to 'trust' in this way? The HB could of course say "This is not to be told to Tommy"... but usually we regard it as obvious that our confidant (sic) knows what should not be told*

to Tommy — either to avoid upsetting Tommy, or to avoid upsetting the original HB. How is a CC to emulate that?

The HB could certainly say "Tell this to no-one" — where "no-one" includes other CCs. But would the HB always remember to do that?

How could a secret-sharing CC deal with family feuds? Some family websites have special functionalities to deal with this. E.g Robbie is never shown input posted by Billie. Could similar, or more subtle, functionalities be given to CCs?"

I think Boden brings up real difficulties in extending this notion to a computer Companion, but I do not think the problems are all where she thinks. I see no difficulty in programming the notion of explicit secrets for a Companion, or even things to be kept from specific individuals ("Never tell this to Tommy"). Companions will have less problems remembering to be discrete than people do, and I suspect there is less instinctual discretion that Boden suggests: people have to be told explicitly who to say what to in most cases, unless they are told to tell no one. In any case, much of this will be moot because Companions will normally deal only with one person — which is what makes their speech recognition problem so much easier, as we noted — they are trained for a single speaker — except when, say, making phone calls to an official, friend or restaurant, where they can try to keep the conversation to limited replies they can be sure to understand. The notion of a stored fact that must not be disclosed is simple to code, and the issue is wider in that the same fact must, to preserve the secret, not take part in inference processes either. If it is a secret that Tom is really a Russian, then the Companion should not do inferences like [IF X is of nationality Y THEN X will normally speak Y] and come out with an utterance like "I assumed Tom could speak Russian", which would rather give the game away via the reverse inference, in the hearer [IF X speaks Y THEN X may well be of nationality Y].

The interesting case Boden raises is that of Companions talking to each other, and this was presumably always a risk for Victorian ladies: that their human Companions would gossip behind their backs. For our Companions this seems a positive development that we might encourage: imagine the shy older person in a care home, too shy to approach another for a lunch together. This would be something best settled between their Companions, each knowing the tastes and habits of their owner, to whom the "date" could be presented as a fait accompli. Again, many Companion-to-Companion interactions will be between an individual's Companion and some form of "public Companion" such as one that takes restaurant bookings based on a user's tastes; or at a hospital where a hospital-Companion could triage incoming patients, who may not be articulate about their condition, on the basis of detailed knowledge of the user's medical records. When traveling, this Companion-to-Companion interaction in, say, a hospital could also combine with translation where the respective Companions worked out how to communicate across a language barrier.

In all these cases, Companion-to-Companion communication could be of obvious benefit to a user even if confidential information was at risk of disclosure: the user might have said "Never tell anyone I'm HIV positive" but in the hospital environment that constraint should obviously be overridden and the user's condition revealed. One could say at this point that secrets may be relative to a situation and that there may be nothing more complex in a Companion's guardianship of secrets than there is in explicit restrictions one could give to human hearers. The ultimate revelation of secrets by a Companion after a user's death is a wholly separate and complex subject. There are already on the market (e. g. Deathswitch: http://www.deathswitch.com/) products that save and reveal passwords and ultimate letters and secrets at death; this is undoubtedly an area with enormous possibilities as the Internet makes actual death less apparent and immediate in the electronic world than it is the real one (see also Wilks http://people.oii.ox.ac.uk/yorick/2007/01/24/death-and-the-internet/).

If the Victorian list of characteristics above is in any way plausible, it suggests an emphasis rather different from that current in much research on emotions and computers (e. g. the HUMAINE network at emotion-research.net) and their possible embodiments and deployments to a public. The emphasis in the list is on what the self-presentation and self-image of a possible, and tolerable, Companion should be; its suggestion is that overt emotion may not be what is wanted at all. I have never felt wholly comfortable with the standard Embodied Conversational Agent (ECA) approach in which if, an avatar "has" an emotion, it immediately expresses it, almost as if to prove the capacity of the screen graphics. This is exactly the sort of issue tackled by Darwin (1872) and such overtness can seem to indicate almost a lower evolutionary level than one might want to model, in that it is not a normal feature of much human interaction. The emotions of most of my preferred and frequent interlocutors, when revealed, are usually expressed in modulations of the voice and a very precise choice of words, but I realize this may be just cultural prejudice.

On the other hand, pressing the pet analogy might suggest that, if that is to be the paradigm, then overt demonstrations of emotion are desirable and sought by pet owners: dogs do not much disguise their emotions, and their positive emotions are often welcomed by owners. Language, however, does disguise emotion as much as it reveals it, and its ability to please, soothe and cause offence are tightly coupled with linguistic expertise — as opposed to the display of gestures and facial expressions — as we all know with non-native speakers of our languages who frequently offend, even though they have no desire to do so, and often have no awareness of the offence they cause. What name to call someone by, or whether or not to use vocatives like "Sir", "Mister", "Miss", "Missus" are enormously complex matters, known intuitively to native speakers but not to outsiders, who are never taught them and have nowhere to go for advice or instruction. These are not cultural matters across space only, but also time: it was pointed out long ago that in the 19C male Cambridge undergraduates would walk arm-in-arm and call each other by their last names, without giving offence, whereas in the latter part of the 20C they would use first names — since last names would have given offence — and never be seen arm-in-arm!

I personally find the lady's Companion list above an attractive one: it eschews emotion beyond the linguistic, it implies care for the mental and emotional state of the user, and I would personally find it hard to abuse any computer with the characteristics listed above. It is no accident, of course, that this list fits rather well with the aims of the Senior Companion demonstrator in the COMPANIONS project already mentioned above. But the project first produced a Health and Fitness Companion (http://www.youtube.com/watch?v=KQSiigSEYhU&feature=related) for the more active, one sharing much of the architecture with the first, and one that would require something in addition to the list above: the "personal trainer" element of weaning, coaxing and threatening which adds something quite different to that list. and something very close to the economic-game bargain of the kind discussed in some detail by Lowe (in Wilks, 2010).

Many of the situations discussed above are, at the moment, wildly speculative: that of a Companion acting as its owner's agent, on the phone or World Wide Web, perhaps holding power of attorney in case of an owner's incapacity and, with the owner's advance permission, perhaps even being a source of conversational comfort for relatives after the owner's death. Companions may not all be nice or even friendly: Companions to stop us falling asleep while driving may tell us jokes but will probably shout at us and make us do stretching exercises. Long-voyage Companions in space will be indispensable cognitive prostheses (or, more correctly, orthoses) for running a huge vessel and experiments above any beyond any personal services — Hollywood already knows all that. All these situations are at present absurd, but perhaps we should be ready for them.

Acknowledgement

This work was funded by the Companions project www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

References

1. *Colby, K. M.* “Artificial Paranoia.” Artif. Intell. 2(1) (1971), pp. 1–2
2. *Cheepen, C. and Monaghan, J.* 1997, ‘Designing Naturalness in Automated Dialogues — some problems and solutions’. In Proceedings ‘First International Workshop on Human- Computer Conversation’, Bellagio, Italy.
3. *Ciravegna, F., Chapman, S., Dingli, A. and Wilks, Y.* 2004. Learning to harvest the semantic web, in Proc. European Semantic Web Symposium (ESWS04)
4. *Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J. G.* 2001. Emotion recognition in human-computer interaction, Signal Processing Magazine, IEEE, 18(1), pp. 32–80.
5. *Höök, K.* (2004) User-Centred Design and Evaluation of Affective Interfaces, In From Brows to Trust: Evaluating Embodied Conversational Agents, Edited by Zsofia Ruttkay and Catherine Pelachaud, Kluwer's Human-Computer Interaction Series.
6. *Krippendorff, K. 2004. Content Analysis: An Introduction to Its Methodology. 2nd edition, Thousand Oaks, CA: Sage.*
7. *Levy, D.* 2007. Love and Sex with Robots: The Evolution of Human-Robot Relationships. London: Duckworth.
8. *Loebner:* http://www.loebner.net/Prizef/loebner-prize.html
9. *Luneski, A., Moore, R. K., & Bamidis, P. D.* (2008). Affective computing and collaborative networks: towards emotion-aware interaction. In L. M. Camarinha-Matos & W. Picard (Eds.), Pervasive Collaborative Networks (Vol. 283, pp. 315–322). Boston: Springer.
10. *Marsella, S. and Gratch, J.* (2003) Modeling Coping Behavior in Virtual Humans: Don't Worry, Be Happy. 2nd Int Conf on Autonomous Agents and Multiagent Systems (AAMAS), Melbourne, Australia, July 2003.
11. *Reeves, B., Nass, C.* 1996, The media equation: how people treat computers, television, and new media like real people and places, Cambridge: Cambridge University Press, 1996.
12. *Scherer, S., Schwenker, F. and Palm, G.* 2008. Emotion recognition from speech using multi-classifier systems and rbf-ensembles, in Speech, Audio, Image and Biomedical Signal Processing using Neural Networks, pp. 49–70, Springer: Berlin.
13. *Wallis, P., Mitchard, H., O'Dea, D., and Das, J.* 2001, Dialogue modelling for a conversational agent. In ‘AI-2001: Advances in Artificial Intelligence’, Stumptner, Corbett, and Brooks, (eds.), In Proceedings 14th Australian Joint Conference on Artificial Intelligence, Adelaide, Australia.
14. *Walters, M., Dautenhahn, K., te Boekhorst, R., Koay, K., Syrdal, D.*. 2009. An Empirical Framework for Human-Robot Proxemics.In Proc.AISB Convention 2009. www.aisb.org.uk/convention/aisb09/.
15. *Webb, N., Benyon, D., Hansen, P. and Mival, O.* (2010) Wizard of Oz Experiments for a Companion Dialogue System: Eliciting Companionable Conversation. Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010), Valletta, Malta. 2010.
16. *Wiebe, J., Wilson, T., and Cardie, C.* 2005. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, volume 39, issue 2–3, pp. 165–210.
17. *Wilks, Y.* 2006, ‘Artificial Companions as a New Kind of Interface to the Future Internet. Oxford Internet Institute Research report No. 13 (Oxford Internet Institute). [Online], Available at: http://www.oii.ox.ac.uk/research/publications.cfm.
18. *Wilks, Y.* (ed.) (2010) Artificial Companions in Society: scientific, economic, psychological and philosophical perspectives. John Benjamins: Amsterdam.
19. *Wilks, Y., Catizone, R., Worgan, S., Dingli, A., Moore, R. K. and Cheng, W.* (in press) A prototype system for a conversational Companion for reminiscing about images. Computer Speech and Language.
20. *Wundt, W.* 1913. Grundriss der Psychologie, A. Kroner: Berlin.
21. *Zue, V., Glass, J., Goddeau, D., Goodine, D., Hirschman, L.* 1992. The MIT ATIS system, In Proc. Workshop on speech and natural language, Harriman, New York.