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

Yorick Wilks
Florida Institute of Human and Machine Cognition
ywilks@ihmc.us

Abstract

I start from a perspective close to that 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 a Companion should not be *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?

1. Introduction

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 plausible 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;
4) It establishes some form of relationship with that user, if that is appropriate, which would have aspects associated with the term "emotion", and shared initiative is essential;
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—which could be public data like Facebook, or life information built up by the Companion over long periods of interaction with the user) and to act in the world, e.g. to reserve at a restaurant or call a doctor. But a Companion *need not* be a robot to act in the world in this way, and we may as well assume its internet agent status, with access to open internet knowledge sources.

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

i) 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

Proceedings of the 2010 Workshop on Companionable Dialogue Systems, ACL 2010, pages 13–18,
Uppsala, Sweden, 15 July 2010.

risk in human-machine interactions.

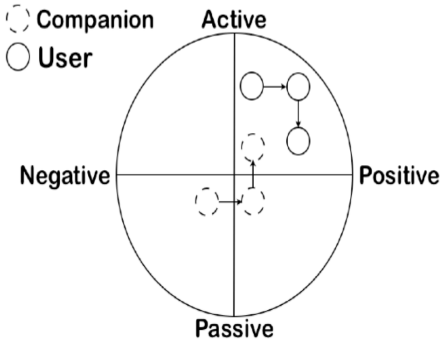
We know, since the original work of Nass (Reeves and Nass, 1996) and colleagues that people will display some level of feeling for the simplest machines, even PCs in his original experiments, and Levy (2007) has argued persuasively that the trend seems to be towards high levels of "affectionate" relationships 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 utterance 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. All such assignments of tags rest, like the text-sentiment theories above, on human pre-tagging. 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, 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.

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 figure below, from REFERENCE REMOVED), where there is the same commitment to the cen-

trality of emotion in the communication process, but in a form focusing on an integration of speech and language (rather than visual and design) technologies. Their argument is for 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?".



ii) What should a Companion look like?

A faceless Companion is a plausible candidate for Companionhood: the proverbial furry handbag, warm and light to carry, chatty but with full internet access. 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). But many observers do not feel this, and, indeed it cannot in principle apply to an avatar so good that one cannot be sure it is artificial, as many feel about the *Emily* from Manchester (Emily 2009).

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 REMOVED demonstration video so far features REMOVED.

iii) Voice or Typing to communicate with a Companion?

At the moment the limitation on the use of voice is two-fold: 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 a web-based Companion may have to use typed input in the immediate future—though using TTS output—

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.

iv) 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.

v) Ethics and goals in the Companion

The issue 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.

vi) 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.

vii) 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 performance 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.

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.

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. One could, 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 func-

tionalities to deal with this. E.g Robbie is never shown input posted by Billie. Could similar, or more subtle, functionalities be given to CCs?"

Boden brings up real difficulties in extending this notion to a computer Companion, but the problems are not 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 people have less instinctive discretion than Boden believes: they have to be told explicitly who to say what to, or not, 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 except when, say, making phone calls to an official, friend or restaurant, where they can try to keep the conversation to limited replies that they can be sure to understand. The notion of a stored fact that must not be disclosed is relatively simple to code. Nonetheless, the Lady's Companion analogy foresees that Companions will, in time, gossip among themselves behind their owners' backs.

I would argue that the "Lady's Companion" list above an attractive and plausible one: it assumes emotion will be largely linguistic in expression, 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. 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.

Acknowledgement:

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

References

Colby, K.M. "Artificial Paranoia." Artif. Intell. 2(1) (1971), pp. 1-2

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.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, JG. 2001. Emotion recognition in human-computer interaction, Signal Processing Magazine, IEEE, 18(1), pp. 32–80.

Emily,2009.http://www.youtube.com/watch?v=UYgLFt5wfP4&feature=player_embedded#

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

Krippendorff, K. 2004. Content Analysis: An Introduction to Its Methodology. 2nd edition, Thousand Oaks, CA: Sage.

Levy, D. 2007. Love and Sex with Robots: The Evolution of Human–Robot Relationships. London: Duckworth.

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.

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.

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.

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.

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.

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.

Wilks, Y. (ed.) (2010) Artificial Companions in Society: scientific, economic, psychological and philosophical perspectives. John Benjamins: Amsterdam.

Wundt, W., 1913. Grundriss der Psychologie, A. Kroner: Berlin.

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.