Artificial companions

YORICK WILKS
Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK

The paper concerns the closely related topics of the possibility of machines having identifiable personalities, and the possible future legal responsibilities of machines. As will become clear, I wish to explore these topics in terms which are basically those of computational linguistics; and by 'machines' here, I intend software entities, rather than robots, and in particular the sort of software agents now being encountered on the web, ranging at present from technical advisers to mere chatbots. I call them Companions. The body of the paper explores the following related aspects of a Companion in more detail: what kind and level of personality should be in a machine agent so as to be acceptable to a human user, more particularly to one who may fear technology and have no experience of it; and what levels of responsibility and legal attribution for responsibility can we expect or desire from entities like web agents in the near future?

What will an artificial Companion be like? Who will need them and how much good or harm will they do? Will they change our lives and social habits in the radical way technologies have in the past: just think of trains, phones and television? Will they force changes in the law so that things that are not people will be liable for damages; up till now, it has been the case that if a machine goes wrong, it is always the maker or the programmer, or their company, which is at fault. Above all, how many people with no knowledge of technology at all, such as the old or very young, will want to go about, or sit at home, with a Companion that may look like a furry handbag on the sofa, or a rucksack on the back, but which will keep track of their lives by conversation, and be their interface to the rather elusive mysteries we now think of as the internet or web.

One thing we can be quite sure of is that artificial Companions are coming. In a small way they have already arrived and millions of people have already met them. The Japanese toys known as Tamagochi (literally 'little eggs') were a brief craze in the West that saw sensible people worrying that they had not played with their Tamagochi for a few hours and it might have begun to pine, where pining meant sad eyes and icons on a tiny screen, and playing with it meant pushing a feed button several times! The extraordinary thing about the Tamagochi (and later the US Furby) phenomenon was that people who knew better began to feel guilt about their behaviour towards a small cheap and simple toy that could not even speak.

The brief history of those toys says a great deal about people and their ability to create and transfer their affections, despite their knowledge of what is really going on inside an object. This phenomenon is almost certainly a sign of what is to come and of how easy people will find it to identify with and care for automata that can talk and appear to remember who they are talking to. This paper is about what it will be like when those objects are available: since most of the basic technologies such as speech recognition and

DOI 10.1179/030801805X25945

simple machine reasoning and memory are already in place, this will not be long. Simple robot home helps are already available in Japan, but we only have to look at them and their hard plastic exteriors and their few tinny phrases about starting the dishwasher to realise that this is almost certainly not how a Companion should be.

The technologies needed for a Companion are very near to a real trial model; some people think that artificial intelligence (AI) is a failed project after nearly fifty years, but that is not true at all: it is simply everywhere. It is in the computers on two hundred tonne planes that land automatically in the dark and fog and which we trust with our lives; it is in chess programs like IBM's Big Blue that have beaten the world champion; and it is in the machine translation programs that offer to translate for you any page of an Italian or Japanese newspaper on the web. And where AI is certainly present, is in the computer technologies of speech and language: in those machine translation programs and in the typewriters that type from your dictation, and in the programs on the phone that recognise where you want to buy a train ticket to. But this is not a paper about computer technology any more than it is about robots; nor is it about philosophy. Companions are not at all about fooling us, because they will not pretend to be human at all.

Imagine the following scenario, which will become the principal one running through this paper. An old person sits on a sofa, and beside them is a large furry handbag, which we shall call a Senior Companion; it is easy to carry about, but much of the day it just sits there and chats. Given the experience of Tamagochi, and the easily ascertained fact that old people with pets survive far better than those without, we will expect this to be an essential lifespan and health improving object to own. Nor is it hard to see why this Companion that chats in an interesting way would become an essential possession for the growing elderly population of the EU and the US, the most rapidly growing segment of the population, but one relatively well provided with funds.

Other Companions are just as plausible as this, in particular the Junior Companion for children, that would most likely take the form of a backpack, a small and hard to remove backpack that always knew where the child was. But the Senior Companion will be our focus, not because of its obvious social relevance and benefit, possibly even at a low level of function that could easily be built with what is now available in laboratories, but because of the particular fit between what a Companion is and old people's needs.

Common sense tells us that no matter what we read in the way of official encouragement, a large proportion of today's old people are effectively excluded from information technology, the web, the internet and some mobile phones because 'they cannot learn to cope with the buttons'. This can be because of their generation or because of losses of skill with age: there are talking books in abundance now, but many otherwise intelligent old people cannot manipulate a tape recorder or a mobile phone, which has too many small controls for them with unwanted functionalities. All this is pretty obvious and well known and yet there is little thought as to how our growing body of old people can have access to at least some of the benefits of information technology without the ability to operate a PC or even a cellphone.

After all, old people's needs are real, not just to have someone to talk to, but to deal with correspondence from public bodies, such as councils and utility companies demanding payment; with the need to set up times to be visited by nurses or relatives by phone; with how to be sure they have taken the pills, when keeping any kind of diary may have

become difficult; as well as with deciding what foods to order, even when a delivery service is available via the net but is impossibly difficult in practice for them to make use of. In all these situations one can see how a Companion that could talk and understand and also gain access to the web, to email and a mobile phone could become an essential mental prosthesis for an old person, one that any responsible society would have to support. But there are also aspects of this which are beyond just getting information, such as having the newspapers blown up on the TV screen till the print is big enough to be read, and dealing with affairs, like paying bills from a bank account.

It is reliably reported that many old people spend much of their day sorting and looking over photographs of themselves and their families, along with places they have lived and visited. This will obviously increase as time goes on and everyone begins to have access to digitised photos and videos throughout their lives. One can see this as an attempt to establish the narrative of one's life; what drives the most literate segment of the population to write autobiographies (for the children) even when, objectively speaking, they may have lived lives with little to report. But think of what will be needed if a huge volume of material is to be sorted. A related notion of 'memories for life' has been developed and recently been declared a major challenge for future UK computing, and we shall return to this whole matter again below.

One can see what we are discussing, then, as democratising the art of autobiography, which was about Chaps, as Benchley readers will remember, whereas the art of Geography was about Maps. And this will mean far more than simply providing ways in which people with limited manipulative skills can massage photos and videos into some kind of order on a big glossy screen: it will require some guiding intelligence to provide and amplify a narrative that imposes a time order. Lives have a natural time order, but this is sometimes very difficult to impose and to recover for the liver; even those with no noticeable problems from aging find it very hard to be sure in what order two major life events actually happened: 'I know I married Lily before Susan but in which marriage did my father die?'

The frivolous example is to illustrate, and we shall see something of how it is actually done later on, how an artificial agent might assist in bringing events whether in text or pictures, into some single coherent order or at least some partial orders so that one knew some things were before others, even if there were some events (e.g. Teddy's and Joan's marriages) that could not be ordered with certainty. This is the kind of thing today's computers can be surprisingly good at, but it is a very complex and abstract notion, that of the time ordering of events, which can be simple (I know James was born before Ronnie) or only partial in some situations (I know my brother's children were born after my marriage and before my wife died, but in what order did they come?). These examples may seem odd or contrived, but they do represent real problems at the border of memory and reasoning for many people, especially in older age.

Another reason why this notion of ordering life narratives is so important, for a Senior Companion and otherwise, is that it is also a way of separating different but similar lives from each other on the web, and these two notions are closely related. This is jumping ahead a little, but many people know the experience of searching on the web for, say, 'George Bush' in Texas and finding there are about twenty-five of them who merit some attention from Google. Since two of them have been US Presidents, they cannot be distinguished from each other by name and job alone, and one must then use life events, dates of birth and so on to separate them. To put the whole thing simply, distinguishing

closely related or confusingly named people on the web requires something like a coherent lifeline of some of the events in their lives, which is the same notion we have been discussing for imposing coherence on the, possibly muddled, whole life memories of old people.

We have talked of Companions as specialised computers that could talk and assist needy groups of people such as the old. That help will almost certainly involve helping to organise their lives and memories but also interacting on their behalf with the electronic world outside. That may be as simple as using the web to find Tesco's home delivery service for groceries. More interestingly, it may involve using the web to find out what has happened to their old school friends and workmates, something millions already use the web for. But, as we have seen, we shall need some notion of time lines and the coherence of lives on the web to sort out the right friends and schoolmates from the thousands of other people with the same names.

So, the reasoning technologies we shall need to organise the life of the Companion's owner may turn out to be the very same technologies – we have not really shown this yet, just said it – required to locate other individuals and select them out from all the personal information about the world's population that fills up the web, since the web is now not just for describing the famous but covers potentially everyone. Two of my friends and colleagues who are professors of computer science have some difficulty on the web distinguishing and maintaining a difference between themselves and in one case a famous pornography supplier in Dallas, and in another a reasonably well known disc jockey in Houston.

These problems will soon become not just quirky but the norm for everyone, and what I shall want to argue below is that the kind of computer agency we shall need in a Companion, and possibly a Companion that deals with the web for us if we are old or maybe just lazy, is in fact closely related to the kind of agency we shall need to deal with the web in any case as it becomes more complex. To put this very simply: the web will become unusable for non-experts unless we have human-like agents to manage its complexity for us. The web itself must develop more human-like characteristics at its peripheries to survive as a usable resource and technology: merely locating individuals on the web when a majority of the EU and US populations have a web presence will become far more difficult and time consuming than it is now. If this is right, Companions will be needed by everyone, not simply the old, the young and the otherwise handicapped. It is going to be impossible to conceive of the web without some kind of a human face.

The notion of a Companion developed so far is anything but superhuman; it is vital to stress this because some of the public rhetoric about what companionable computers will be like has come from films such as *2001*, whose computer HAL is superhuman in knowledge and reasoning. He is a very dangerous Companion, and prepared to be deceptive to get what he wants, which may be not at all what we want. Seymour Papert at MIT always argued that it was a total misconception that AI would ever try to model the superhuman, and that its mission just like AI pioneer John McCarthy's emphasis on the importance of common sense reasoning was to capture the shorthand of reasoning, the tricks that people actually use to cope with everyday life. Only then would we understand the machines we have built and trained and avoid them becoming too clever or too dangerous. This same issue was very much behind Asimov's Laws of Robotics, which we shall discuss below, and which set out high level principles that no robot should ever break so as to bring harm to humans.

The difficulty here is fairly obvious: if a robot were clever enough it would find a way of justifying (to itself) an unpleasant outcome for someone, perfectly consistently with acceptable overall principles. I say 'clever enough', but doing that has also been a distinctively human characteristic throughout history: one thinks of all those burned for the good of their own souls and all those sacrificed so that others might live. In the latter case, we are probably grateful for those lost in what were really medical experiments, such as the early heart transplants, even though they were never called that.

It will not be possible to ignore these questions when presenting Companions in more detail, in particular the issue of where responsibility and blame may lie when a Companion acts as a person's agent and something goes wrong. At the moment, Anglo-American law has no real notion of any responsible entity except a human, if we exclude Acts of God in insurance policies. The only possible exception here is dogs, which occupy a special place in Anglo-Saxon law and seem to have certain rights and attributions of character separate from their owners. If one keeps a tiger, one is totally responsible for whatever damage it does, because it is *ferae naturae*, a wild beast. Dogs, however, seem to occupy a strange middle ground as responsible agents, and an owner may not be responsible unless the dog is known to be of bad character. We shall return to this later and argue that we may have here a narrow window through which we may begin to introduce notions of responsible machine agency, different from that of the owners and manufacturers of machines at the present time.

It is easy to see the need for something like this: suppose a Companion told your grandmother that it was warm outside, and that when she went out into the freezing garden on the basis of this news she caught a chill and became very ill. In such a circumstance one might well want to blame someone or something and would not be happy to be told that Companions could not accept blame, or that if one read the small print on the Companion box one would see that the company had declined all blame and required a signature on a document to that effect. All this may seem fanciful, and even acceptable if one's grandmother recovered and the company gave the Companion a small tweak so it never happened again.

This last story makes no sense at the moment, and indeed the Companion might point out with reason, when the maintenance doctor called round, that it had read the outside temperature electronically and could show that it was a moderate reading and the blame should fall on the building maintenance staff, if anywhere. These issues will return later but what is obvious already is that Companions will have to be prepared to show exactly why they said the things they said and offered the advice they did.

A Companion's memory of what it has said and done may be important but will be used only rarely one hopes; though it may be necessary for it to repeat its advice at intervals with a recalcitrant user ('You still haven't taken your pills. Come on, take them now and I'll tell you a joke you haven't heard before'). What may become a very important feature of Senior Companions is putting coherence into the memories of their owners: this was mentioned above as a way of organising memories and sorting fragments of text and old photographs, but there is another aspect to this which is not so much of relevance for the user as for their relatives later on.

One can reasonably suppose that years of talking with an elderly user, and helping them organise their memories, will mean that a Companion also has access to a life story of the

user that is has built up from those hours of conversation. The user might not want to hear any of this played back, but it could be used as a body of facts and assumptions to help organise the user's life as well as text and images. But what should become of this biographical account that the Companion has, and which could exist in several forms – for example a document the Companion could print or a collage of things said by the user and Companion put together coherently from recorded pieces, or even a single autobiographical account in the Companion's version of the user's own voice? This is not at all far-fetched: there are speech generation packages now available that can be trained to imitate a particular person's voice very accurately with plenty of training time, and time is exactly what the Companion and user would have.

I would suggest that memoirs like these, produced over a long period by the Companion, are exactly what the user's relatives will want after the user is gone, something more helpful and less gruesome than the tiny video clips one can now find and run at the touch of a button on some Italian tombstones. The need for these is now greater than it was, as people live farther from their parents when old and see them less. Many people have wished they had spent more time discussing a parent's earlier memories before their deaths – how one's parents had met and fallen in love, for example – but then suddenly it is too late to ask, or the answer is unobtainable because of shyness on the part of parent or child. Production of this kind of limited memoir is an important role a Companion might come to play in society; experience may well show that old people will reveal memories and anecdotes to a Companion they would perhaps not feel able to tell their close relatives. Indeed, there is a long tradition in artificial intelligence of arguing that people may sometimes prefer machines to people in certain roles: AI pioneer Donald Michie always claimed that drivers preferred traffic lights (or 'robots' as they are called in some parts of the world) to policemen on traffic duty.

The possibility of a Companion constructing and storing a biography for its owner raises in one form the major issue of identity: is the Companion to be distinguished from the owner whose life it knows so well and whose voice it will almost certainly be able to imitate. And this may be just another version of the old joke about pets getting like their owners, since we have chosen to describe the Senior Companion as a kind of pet. Other forms of identity will also be touched on below, in particular the identity of the Companion as an agent for the owner, and its similarity and distinctness from the owner, while on the other hand functioning as a web agent in the world of electronic information. We have not yet properly introduced the notion of a web agent. It is, very roughly, an intelligent package of software that one can encounter on the internet and which will give expert advice. It is now expected that web agents will have to make deals and transactions of all sorts with each other and learn to trust each other, as they already do in a rudimentary sense in the world of banking where agents in their thousands clinch money transactions between financial institutions. The kind visible to a user, but still rudimentary, are those which will search the whole web to find the cheapest source of, say, a particular camera.

All this is for later and elsewhere, but these two issues of identity in an artificial world will also draw upon other ways in which identity has become an issue for internet users and which are relatively well known. The first, known to all newspaper readers, is that of chat room users who pretend to be what they are not; normally this is quite innocent pretence, and little more than hiding behind a pseudonym during conversations and

sometimes pretending to be a different kind of person, often of the opposite sex. In that sense, the Victorian game of sex-pretending, on which Turing based his famous imitation game for computers, has come back as a widely played reality. The problems only arise, and they are very real, when impressionable people, usually children, are lured into meetings with people who have been encountered under false pretences.

The other issue of identity, which is a standard problem for web searchers, is finding too many people under the same name with a Google search and trying to find the right one, or even how many there are, a topic we already touched on above. It is a much researched problem at the moment how to sort out exactly how many people there are in the pages retrieved by Google for the same name. One can see the pressing interest here, in that many scientists now rate how famous they are by how many Google hits their name gets compared to their competitors. But how can they be sure all those retrieved are really themselves?

The George Bush example mentioned above suggests that the best way to tell must require something like a rule that looks at dates, on the assumption that two people with the same name are very unlikely to have the same date of birth. And other rules will almost certainly deal with aspects of people such as their occupations; which will, however, come up against unusual but very real cases like John Vanbrugh, the eighteenth century playwright, and his separation from the eighteenth century architect of Blenheim Palace of the same name who, amazingly enough, were one and the same person, hard though it would be for most rules (and people) to accept.

There is a further kind of complication in that, even if we could sort out muddled identities of these kinds, given enough information, in some cases people do not agree on how many objects or people there are under discussion, so that it becomes a matter of conflict or, as you might say, individual belief, how many people are being talked about. In the case of Vanbrugh we can imagine a strained conversation between a person sure they were two similarly named individuals and someone else who knew they were not. It is, as one could put it, very difficult but just possible for people to communicate who do not agree on what things there are in the world, on the ontology, to use a word in its original sense that is now normally used to mean something quite different. How should a Companion discuss relatives when it was sure Jack was Ethel's auntie, and its owner even said that the day before, but is now convinced they are quite different people? This is a deep matter to which we shall return.

Problems of identity will arise both in the context of representing individuals in a Companion's world, which is very much based on that of its user, one the Companion seeks to learn, and also in the wider world of information, which for convenience we will often identify with what can be found on the web or internet. It is quite normal now to hear people speak of 'virtual worlds' in connection with the web and internet, although it is usually unclear exactly what they have in mind. The only obvious place where virtual worlds belong naturally is in computer games, whose interaction with the web we will also need to consider, certainly in the case of Companions for the young, who spend more time in games worlds than the rest of the population.

The general interest here is in the interaction of these considerations with identity: having a verifiable identity is part of what it means to be a human being, or at least a modern human being. If a Companion is to have human-like characteristics, one will want to explore how its identity can have human-like features, as opposed to machine-like

features where identity is usually a trivial matter: a car is identified uniquely simply by the numbers stamped on its chassis and its engine and there is no interesting issue about that outside the world of car fraud.

If a Companion is to be an interface to the web, say for a user who is technologically incompetent yet who must conform to the standards of identity that society requires and imposes, then the Companion will have to understand identity to some degree and possibly be able to manipulate slightly different forms of it. In the US and UK, identity is currently established by a range of items with numbers, from passports through credit cards to health, driving licence and tax numbers (some with associated passwords or PINs), with the Social Security number having a definite primacy in the US. In most EU countries there is a single ID number, of which the clearest example is the lifelong single Personnummer in Sweden. States prefer a citizen to be identified by a single number, and in the UK there is currently strong pressure for something closer to the Swedish model, although UK law has at the moment no clear definition of identity (with legally registered unique names and addresses, as in most of the EU): there is no legal problem in the UK with having a number of identities simultaneously and bank accounts for each (as a famous case brought, and lost, by the Post Office showed some years ago) so long as there is no intention to defraud.

All this is important since identity checks are the basis of all web transactions, and if a Companion is to deal with an old person's affairs it will need something approaching a power of attorney or at least an understanding of how identity is established in web transactions, as well as a method for establishing that its owner approves of what it is doing in individual transactions, in case of later disputes (e.g. by angry relatives after an old person's money has been spent). A scenario has recently been described in which security of identity can be achieved without the imposition of unique identity, which security-minded authorities would certainly resist but which may be very important to someone who feels they have a right to buy something on the internet revealing only, say, their age but not who they actually are (see www.chyp.com/pubwebfiles/whitepapers/chyp_response.pdf).

All these issues are the subjects of active research programmes, but what I am introducing here is the possibility that a Companion, as a new kind of artefact among us, may focus our minds on a set of intellectual and practical issues in a new way, even though some of them are very traditional issues indeed.

Yorick Wilks (yorick@dcs.sheffield.ac.uk, www.dcs.shef.ac.uk/~yorick) is professor of computer science and Director of the Institute for Language, Speech and Hearing at the University of Sheffield. He was previously Director of the Computer Research Laboratory at New Mexico State University, and before that professor of linguistics and subsequently computer science at the University of Essex. His research interests include computational pragmatics, computational lexicons and information extraction. His most recent books are *Machine Conversations* (1999, Kluwer) and *Electric Words* (1996, MIT Press, with Louise Guthrie and Brian Slator).
