AI & Soc (2011) 26:329–337
DOI 10.1007/s00146-010-0313-3

ORIGINAL ARTICLE

Advocating an ethical memory model for artificial companions from a human-centred perspective

Patricia A. Vargas · Ylva Fernaeus · Mei Yii Lim · Sibylle Enz · Wan Chin Ho · Mattias Jacobsson · Ruth Aylet

Received: 17 December 2009 / Accepted: 14 December 2010 / Published online: 18 January 2011

Abstract This paper considers the ethical implications of applying three major ethical theories to the memory structure of an artificial companion that might have different embodiments such as a physical robot or a graphical character on a hand-held device. We start by proposing an ethical memory model and then make use of an action-centric framework to evaluate its ethical implications. The case that we discuss is that of digital artefacts that autonomously record and store user data, where this data are used as a resource for future interaction with users.

Keywords Ethics · Privacy · Artificial companions · Robots · Memory modelling

P. A. Vargas (✉) · M. Y. Lim · R. Ayllet
School of Maths and Computer Science,
Heriot-Watt University, Edinburgh, UK
e-mail: p.a.vargas@hw.ac.uk

M. Y. Lim
e-mail: myl@macs.hw.ac.uk

R. Ayllet
e-mail: ruth@macs.hw.ac.uk

Y. Fernaeus · M. Jacobsson
SICS, Swedish Institute of Computer Science, Kista, Sweden
e-mail: ylva@sics.se

M. Jacobsson
e-mail: majac@sics.se

S. Enz
University of Bamberg, Bamberg, Germany
e-mail: sibylle.enz@uni-bamberg.de

W. C. Ho
Department of Informatics, Hertsfordshire University, Herts, UK
e-mail: w.c.ho@herts.ac.uk

1 Introduction

This paper reports on work conducted within the EU FP7 project LIREC—LIving with Robots and intEractive Companions (LIREC 2008). Unlike other related projects (Companions 2007; CHRIS 2008), LIREC sets out to advance the field of artificial companions by taking a longer term stance.

The objective of this paper is to review the ethics involved in a computational memory model in long-term artificial companions, including remembering and forgetting mechanisms (Halpern 2008). In this work, we will make use of the action-centric model of interaction (Fernaeus et al. 2008), designed specifically to put user action and experiences to the fore. Here, we use it as a lens to review the ethics involved in designing autonomous memory controls for computer systems, focusing on the ethical implications of different physical embodiments in relation to which user data are collected, stored and presented. More specifically, we will discuss ethical implications of the implementation of a digital memory in a social context, e.g. "forgetting" and "accessing" user data.

The idea of an "artificial companion" goes back to ancient society with its stories of sculptors falling in love with their creations and of statues coming to life. In more recent time, well-known film characters have dramatized the idea of companionship between humans and interactive artefacts. However, the idea of an artificial companion has not been widely accepted and sometimes not even considered or imagined. Certain scientists and philosophers clearly stand against the whole concept in all circumstances. Sparrow (2002, 2006) for instance argues that given that we cannot currently create robots with real personality, we are liable to create illusions about robot

capabilities that could involve a number of potential ethical dangers (Ambo 2007).

In an earlier paper (Fernaeus et al. 2009), we have argued that it is our responsibility as researchers to actively work towards the creation of systems that can meet existing or realistically feasible needs, within human–robot interaction, avoiding what we label as "a robot cargo cult". We see this as a matter of research ethics, and with this cargo-cult concept, we hope to initiate a discussion on how to develop methods that explicitly address such issues in future research in our field.

Kaplan's (Kaplan 2004) study on an "experienced value" variable in a long-term use of robotic products revealed that in order to increase this value, a robot should be endowed with an historical capacity and act as a repository for memories. It is here that this paper takes its starting point, addressing the specific ethical issues involved in such an endeavour.

Roboethics (Veruggio 2005; Veruggio and Operto 2006) is a field of robotics theory whose main objective is to provide scientific, cultural and technical tools that can be shared by different social groups and beliefs. Thus, it is believed that it should not only comply with the "Charts of Human Rights" but also take into account a variety of ethical theories by analysing possible negative effects of robotics such as abrogation of responsibilities, lack of access, deliberate abuse, terrorism and privacy amongst others, in a wide range of application fields, including the economy, society, law, interaction with the elderly, health and childcare. In this work, we concentrate on ethical issues related to recordings and access of digital material on different hardware platforms.

Here, we take the perspective of *human experience*. This requires a balanced discussion that does not focus on life and death (Asimov 1950; Clarke 1993; Anderson 2007; Arkin 2009). Relevant questions then concern how the technology that we build affects existing social practices, how this image in popular media affects us and our designs and the values that people in general associate with the kinds of technology that we build. Several of the systems we are considering will be endowed with an artificial memory and hence, the associated ethical issues should be investigated (Allen et al. 2006; Wallach and Allen 2009; Denning et al. 2009; Murphy and Woods 2009).

In the following sections, we will introduce three classical stances within ethics. We will then describe our ethical memory model and available methods for "forgetting" and carrying out an action-centric analysis. Finally, we will discuss requirements and guidelines for an ethically "certified" memory model by focusing on what the system should and should not forget and its consequences when we also incorporate ethical theories into the companion's memory model.

2 Background and related work

Standard computationally focused analysis of human–robot interaction seems to miss important aspects of interaction in the physical rather than digital space. For instance, how a robot responds to touch in terms of direct and continuous movement is difficult to capture only through a straightforward diagram of explicit 'input', 'output' and 'data manipulations' as it would be the case in a computer-focused analysis. Defining the 'data' in physical interfaces is generally difficult as it is not always clear (to people) what a device is actually recording. An interesting question with regard to this is how the system treats these recordings, in other words, how they are interpreted, remembered and forgotten.

A well-known dilemma is that public discourse on ethical renegotiation focuses primarily on vague visions of the future or even robots from fiction, rather than on realistically implementable or commercially available systems and artefacts. More advanced robotic systems remain as research prototypes and seldom leave the laboratories so that any proper ethical discourse 'in the wild' can take place. Commercially available artificial companions come primarily in the form of children's toys. These have limited sensing and often a fixed repertoire of behaviour and language, driven by an inflexible and functionally limited finite state machine approach as with computer game characters. Thus, the level of autonomy these devices can display and the user data that they store are very limited. Nevertheless, such toys have been widely sold and thus provide some data aspects relevant to consider (Jacobsson 2009).

A related area of research can be found in the graphical domain where digital pets starting with the Tamagotchi[1] gathered a general acceptance from users. These may be digital versions of real animals such as dogs or cats, or animals that could never be real pets, like a dinosaur. The user can adopt a puppy and take care of it while it grows and make it learn new tricks. Yet these "digital pets" are little more than scripted characters that respond to certain actions by the user and lack the long-term memory of user interaction that we consider here.

2.1 Theories of ethics

Classical ethical theories may be classified into three types: deontological, consequentialist and virtue based. In deontological theory, an action is evaluated a priori as being moral or immoral irrespective of its consequences. Usually, a set of moral rules are defined describing a deontological moral system. A number of systems based on deontological

[1] ++ http://www.tamagotchieurope.com/.

ethics have been created (Gert 1988), an approach sometimes referred to as value-sensitive design (Friedman 1997). A focus is such systems may be e.g. the development of and adjustment to policies and societal laws regarding privacy and data security as well as sensitivity to the values expressed by a specific user group.

Consequentialist ethics states that the consequences should rule one's actions, i.e. that a "good" or "appropriate" action is one that results in "good" or "appropriate" consequences. In this sense, an ethical behaviour of a computer system should involve the ability to estimate or predict the result of an action and being able to evaluate the results of an action according to its intentions.

Virtue-based theory, on the other hand, considers *being* as opposed to only *doing* as the basis for estimating what is appropriate in terms of behaviour. Ethical behaviour is then considered a question of who is conducting an action, e.g. different sets of social rules may apply to different people, and different rules may apply to different technologies, and between different people around different technologies (e.g. different actions may be considered appropriate for a certain robot than for a virtual agent in a computer game).

Gips (1995) poses the question "What types of ethical theories can be used as the basis for programs for ethical robots?". The consequentialist theory could be implemented in a robot but prediction would be an issue. Deontological theory might also seem straightforward to implement but conflicting obligations would have to be dealt with. Virtue-based theory seems to resonate partially with the evolutionary robotics approach (Nolfi and Floreano 2000), but the unpredictability of evolved behaviours is an issue. We believe that in order to design an ethical memory model, one should consider incorporating aspects of all ethical theories (Wallach and Allen 2009). Here, we will focus on aspects of memory modelling and the related forgetting mechanisms. To the best of our knowledge, the combination of ethics and memory in artificial agents has not yet been addressed; hence, the following review of existing computational memory model focuses on efficiency, accuracy and adaptability.

3 A system for artificial memory control

We believe that forgetting mechanisms are useful to improve efficiency, scalability and adaptability of cognitive systems operating in dynamic task environments, such as a robot's interaction environments. Forgetting could be viewed as a way of controlling the memory of the companion since it could be used to regulate (Gold 1992) the type and amount of data stored in the memory, giving rise to a more consistent artificial companion, in terms of data security and thus privacy. In this section, we discuss a number of tentative computational models of human memory to date to thereafter present our proposal, which address the aforementioned issues of memory control.

3.1 Computational models of human memory

Modelling a human-like memory has been researched for some time in AI, and there are a variety of memory models like the *Scripts* (Schank and Abelson 1995).

In recent years, modelling temporal sequences of episodic events, in both robotic and virtual agents research, has been a growing area. By collecting relevant events that are perceived and actions that are conducted, a robot exploring its environment is able to reduce its state-estimate computation in localising itself and building a cognitive map in a partially observable office environment (Endo 2007). Also, storage of such episodic memory sequences with attributing emotions may help a virtual robot to predict rewards from human users, thus facilitating human–robot interactions in a simple Peekaboo communication task (Ogino et al. 2007).

Mirza et al. (2006, 2007) uses the concept of interaction histories, defined as the "temporally extended, dynamically constructed and reconstructed, individual sensory-motor history of an agent situated and acting in its environment including the social environment".

Research modelling a complete "human-like memory", as in the episodic memory of Soar (Nuxoll and Laird 2004) and a generic episodic memory module (Tecuci and Porter 2007), establishes a common structure that consists of context, contents and outcomes/evaluation for agents to remember past experiences.

Brom et al. (2007) attempted to create a "full" episodic memory storing almost everything that could be recorded around the agent for the purpose of storytelling. Forgetting processes were also partially implemented in their work where less emotional-tag-rich records were deleted. Brom and Lukavsky (2009) show an extension of this investigation for graphical characters but claimed that it was possible to apply it to physical robots as well.

Previous research (Ho et al. 2009) modelled the psychological concept of autobiographical memory and integrated it into computational synthetic agent architecture. With this memory included, agents are not only capable of recognising and ranking significant events which originate in the agents' own experiences, but also can remember, recall and learn from these experiences.

Functional decay theory (Altmann and Gray 2000) has been found to be useful in making quantitative predictions for human performance in dynamic task environments. The core idea of this theory is that the most recent information must be the most active in memory to allow reliable and fast retrieval.

Forgetting has also been adopted in many learning algorithms. In structural learning with forgetting (Ishikawa 1996), it is applied to two of the three phases: learning with forgetting and selective forgetting, and hidden units clarification. Koychev (2000) utilises a gradual forgetting method in learning drifting concepts by applying time-based forgetting function. The idea is comparable to functional decay theory that the most recent information is the most active in memory. The result of experiments showed an improved predictive accuracy and adaptability of the systems that adopt learning algorithms with gradual forgetting.

Despite all the effort hitherto made to create a more realistic and reliable computational memory model, none of the aforementioned models have adequately accounted for its intrinsic ethical implications when considering a long-term interaction with the user or fully exploited these characteristics (Ho et al. 2009; Vargas et al. 2009).

3.2 A proposed system for an ethical memory control

The fact that our proposed system makes use of an artificial memory model for long-term companions underlines the necessity of using a forgetting mechanism as pointed out by Vargas et al. (2009). Moreover, ethical issues related to data security, i.e. privacy (Allen et al. 2006; Wallach and Allen 2009; Denning et al. 2009) amongst others (Anderson 2008; Murphy and Woods 2009) should also be addressed while designing such a model.

Note that we are not proposing to closely mimic the particular natural processes of human memory but rely on a technical solution that can actually adjust to the dynamics of social interaction between artificial companions and users. By following this research direction, different artificial companions developed in the LIREC project will complement the human user's limitations in certain physical aspects (e.g. carrying objects) and cognition (e.g. remembering events), taking on the role of assisting human users in their everyday activities.

In a previous work, Ho et al. (2009) proposed an initial memory model for long-term companion technology. This memory model enables the artificial companion to remember events that are relevant or significant to itself or to the user. The two main components in the model are working and long-term "memories". Working memory (WM) supports agents in focusing on the stimuli that are relevant to their current active goals within the environment. Long-term memory (LTM) contains episodic events that are chronologically sequenced and derived from an agent's interaction history both with the environment and with the user. Meanwhile, LTM also produces concepts as knowledge about the world in order to help in formulating and processing new goals.

Lim et al. (2009) proposed an extension to the same memory model that included forgetting mechanisms not only through time-based decay for short-term (STM) and LTM but also considering *repression* or *motivated forgetting*.

Vargas et al. (2010) developed the model further by suggesting a memory model that supports forgetting of events through the processes of generalisation and memory restructuring.

In this work, we try to go a step further towards an ethical memory control by suggesting the incorporation of ethical theories to the previous models (Ho et al. 2009; Lim et al. 2009). Apart from privacy, we think that other ethical issues should be addressed. Therefore, it does not suffice, for instance, to incorporate Asimov's three laws in our future artificial companions as already highlighted by other researchers (Anderson 2008; Murphy and Woods 2009).

We believe that in order to create an ethical robot, one should consider incorporating aspects of all the ethical theories discussed above (Wallach and Allen 2009). These could be combined into a master "roboethical" theory, which would encompass the positive features of each one, while attempting to overcome the shortfalls. For instance, appropriateness in behaviour of a digital artefact could be programmed as a set of rules (deontological theory), which may be acted out differently on different hardware platforms (virtue-based theory) and also by applying predictions that could be learned by practice (consequentialist theory).

In order to illustrate our approach, an artificial companion can learn (consequentialist-based ethics) data privacy regarding contents as well as contexts from making mistakes that are later rectified by the user. Each time the system makes a mistake e.g. by presenting an unwanted recording (which might be negotiable) and the user corrects the system, the memory architecture should create a new rule (deontological) to handle the same type of data under the remembered context, namely the current environment and other people's presence. Once the user reinforces the new rule, it allows the system to be attentive to a particular type of information while interacting with the environment and perceiving data through sensors. This new rule can help processing the information and enable a "situational forgetting" mechanism, which allows the system to "forget" a piece of sensitive information under specific circumstances to satisfy the user's expectations and attend its requests via making accurate predictions (consequentialist).

To recapitulate, by considering ethical issues, a priori knowledge (a deontological system), a learning process taking into account the properties of the particular platform (virtue based) and a prediction scheme (consequentialist) should be part of the "ethical" system as described earlier together with the aforementioned forgetting mechanisms.

The proposed model tries to incorporate all three major ethical theories in a single system, albeit not considering how this model could actually promote a behaving artificial memory or an ethical artificial memory. This is part of our next step in the design of such memory control. Thus, in the next section, we will apply an action-centric framework (Fernaeus et al. 2008) to aid our memory model development by providing us with a concrete reference to facilitate our further enhancement of the prototype.

4 The ethical aspects of the proposed model

Taking an ethical or a user- and experience-centred perspective seriously means that digital artefacts must be described from a perspective of *human* action and experience, rather than only from a *system* perspective, which is currently more common in robotic research and AI. Moreover, relying on data-centric models (e.g. input-process-output) for user interaction is not sufficient in describing the experience of interacting from a perspective of use. In an attempt to illustrate how a user-centred perspective may be taken further into account when describing interaction with robotic and other interactive artefacts Fernaeus et al. (2008) have developed a new framework to be used as a resource for describing these aspects of interaction. In contrast to data-centric models, the framework describes socially and contextually oriented actions performed around the robotic artefact, as well as actions related to the computational system running on the machine.

The framework emphasises four experiential dimensions of interactive artefacts, emphasising different ways that it may work as a *resource* for human action. The four dimensions are (1) physical manipulation, (2) perception and sensory experience, (3) contextually oriented action and (4) digitally mediated action. These four dimensions are theoretically based on a phenomenological view of human sense making as well as on recent theoretical development based on empirical studies in the areas of human–computer interaction and interaction design and is intended to provide a concrete guideline for designers in addressing the experiences of using an artefact.

Table 1 illustrates what is meant by the four dimensions outlined in Fig. 1, based on the three ethics dimensions described in Sect. 2. Several examples from commonly observed human–technology interactions are provided, as a way of showing how the model could be used to assist in the design of appropriately behaving memory systems.

4.1 Physical manipulation

This dimension concerns how physical objects may be moved and interacted within space, how they may be physically combined, be brought to different environments, how they allow for action and interaction to be performed concurrently, with both hands, jointly or individually. From a perspective of human experience, it is sometimes difficult or even irrelevant to distinguish the physical manipulations that are treated computationally by the system, from those that are not.

An example is how users of a mobile robot may get it to move in another direction by obstructing its path, by physically pushing it or even carrying it to a new location. Naturally, physical manipulation with interactive technology also covers some digital aspects, e.g. the importance of physical nearness when using Bluetooth or RFID, directing the robot to respond to IR signals and pressing of hardware buttons to control the system running on the device. Importantly, physical manipulation also includes physical management, such as procedures for changing or charging batteries, switching the robot on and off, how it can be cleaned and groomed and how the robot may be physically moved and stored. Completely different sets of physical manipulations are involved if the system instead runs on, say, a mobile phone handset.

In terms of digital memory storage, a relevant question is then how these manipulations would be sensed and recorded by the system, and if so, whether and how these recordings could be accessed by the user (or someone else). This is highly relevant as from a perspective of use, it may be unknown how the actions are treated computationally or mechanically inside the device. This is particularly relevant as in most models for artificial memory control physical manipulation is omitted.

What users remember or pay attention to will naturally be different if the interaction with ‘data’ takes different physical forms. Thus, the data that may be recorded are dependent on the physical platform and the way that users experience its physical manipulations. This relates fundamentally to the dimension of ‘virtue’ in ethics theory, i.e. that ethics depends on the form of ‘being’ as well as the actions that it performs.

4.2 Perception and sensory experience

This dimension concerns personal, bodily and emotional engagement with technology, e.g. how the artefact feels to hold, touch, to look at and to listen to. Affective experiences such as fear, curiosity and attachment are also included in this dimension. This not only includes device-specific qualities of the hardware, e.g. weight, texture and hotness, but also digital expressions, such as the experience of sound and visuals on a screen. Importantly, perception and experience are here understood as actions *performed by people*, rather than passively imposed from the artefact.

Table 1 Action-centric framework to evaluate the three ethics dimensions used in the proposed memory model

	Deontological	Consequential	Virtue-based (user roles)	Virtue-based (device)
Physical manipulation	What are the appropriate manipulations?	What may be the consequences of the manipulations?	Are there different appropriate manipulations depending on who you are?	Are there different appropriate manipulations depending on the device?
Perception and sensory experience	What should be possible for users to access/perceive?	What may be the consequences of accessing and perceiving? e.g. violation of privacy.	Who should be allowed to perceive/access the system?	Should you be able to perceive different things depending on the device?
Contextually oriented action	What are the appropriate behaviours *around* the system? e.g. social rules.	What are the consequences of the actions people may do around the system?	What may be the different appropriate *roles* in the context of use?	Would you expect different actions around it depending on the device?
Digitally mediated action	What should be allowed in user-control, access, manipulation and sharing of data?	What could be the consequences of various user actions with the data?	What are the different user roles in the control of the data? Who should be allowed to do what?	Should different things be recorded/mediate depending on the device?

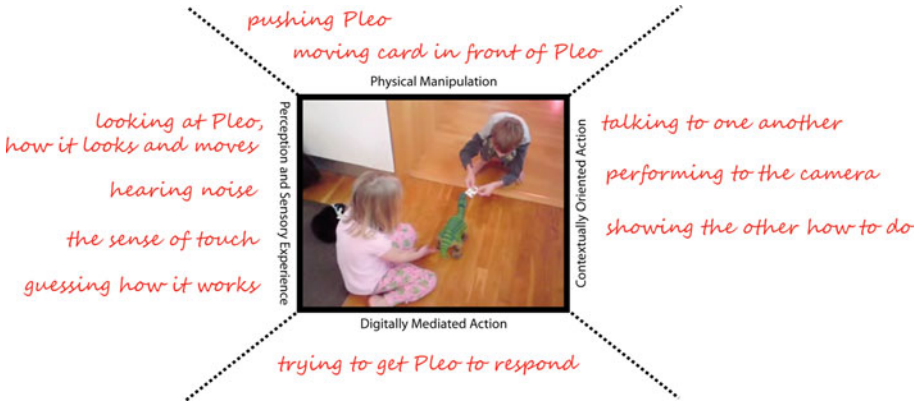

Fig. 1 The framework applied to the case of two children interacting with robotic toy Pleo for the first time, emphasising what the *children* are doing with the robot, rather than how the *robot* reacts to different forms of input

From a perspective of user experience, large motor actions are fundamentally different from animations displayed on a small screen, not only through the ways in which they can be visually accessed, but also in terms of noise, dimensionality in space and the possibility to get an individual viewpoint from different angles.

Related to perception and sensory experience are concepts of skill and body memory, important features that are often neglected in the design of new interactive systems (Klemmer et al. 2006). A concrete example is how an experienced user may control a system by moving its physical parts without having to actually look at the device, actions that are fundamentally shifted when moving the interactive experience, for instance, from a handheld mobile device to that of a physical robot.

An ethical consideration in terms of memory control may for instance include if the system (whether it is an onscreen or physical gestalt) has something that 'looks like' eyes, users may get the *experience* of being watched, even though it in fact does not record anything. Users may, for instance, worry that researchers (or companies) may be able to spy on their actions *through* the technology (Denning et al. 2009). That is a rational concern, given the capabilities of commonly available technology (TV, radio, telephones). An implication of this is that issues such as privacy and data security may be relevant to discuss not only in terms of what the system is *actually* recording, but also in what people *experience it as capable of* recording. To explicitly design the system so that users can perceive when and what information is being recorded is therefore an appealing research challenge.

4.3 Contextually oriented action

For any technology to be used by people, it is relevant to consider how it could be taken up and used as a resource in existing socio-technical contexts. As designers, we need to ask ourselves what situations the technology is meant to support, who would benefit from it, what existing practices they would be engaged in when using

it, and what kinds of technology they currently use in these practices.

Syrdal et al. (2007) conducted an exploration experiment using a human-sized robot, which was operated under remote control while interacting with 12 participants in an experimental trial using "The Wizard of Oz" method (Kelley 1983). One important aspect raised by the analysis of the results was the influence of nationality and thus cultural differences between the participants. As expected, people were mostly concerned with "what" was being stored on the memory of the robot and "how" this data would be processed and to "whom" this information would be further disclosed. It was concluded that systems that are meant to be used by general public should not only explicitly justify any data captured from its users but should also address privacy and data protection issues that are relevant in a particular context of use.

One can use different technologies such as robots and mobile phone handsets as resources for getting attention by others, as indicators of the current state of an activity or as triggers for new conversations. As a concrete example, owners of the commercial toy dinosaur Pleo reported that they placed the robot in their office reception as a form of 'ice breaker' with their customers (Jacobsson 2009). Moreover, we may need to understand how an artificial companion could be used for socially appropriate co-located interaction, for remote communication, and how it might work with other tools in an existing social practice. This class of action includes all of the actions that people perform that are not directly directed towards the system or artefact, but that yet seem important to how users interact with or around it.

Gesture as well as physical manipulations of the device may be directed to the social context, e.g. handing over a handheld robot to a friend, pointing to draw attention to a certain feature of the technology or teaching another person how to use a system. Of particular importance then are aspects such as intended and unintended *audiences* of the interaction. A mobile phone may be discretely manipulated within the privacy of one's personal pocket, while a large robot may demand a more public performance. Similarly, the sound generated on a handheld device may be constrained to a particular user via headphones, which is not always an available option with robotic devices. This means that the 'data' presented through these different channels are accessed differently, including the meanings associated with that data, and how these 'memories' can or will be used. These actions are here defined as contextually oriented, as the primary targets of these may not be towards the interactive artefacts *per se*, but the context around them.

Ethical aspects of contextually oriented action also include who will be responsible for maintenance of the technology, what will be the required skills for updating the software, or how the product conforms to established guidelines for health and safety, cultural norms and sustainability. Once more, all these aspects will fundamentally differ with different hardware platforms.

4.4 Digitally mediated action

This class of action concerns how the technology supports users in controlling and performing with a computational system. This includes the design of all forms of applications made accessible through the device, and how different forms of media can be captured, generated, communicated, controlled and manipulated. As resources for digitally mediated actions, both robots and handheld devices may provide new possibilities and precision in manipulation and navigation in virtual spaces, richer ways of accessing recorded and interactive media, or for remote communication between people.

This dimension also includes actions where the robot performs autonomously, for instance an industrial robot that with great precision assists a worker by repeatedly performing a complex manoeuvre, a robotic toy that can be trained to perform new actions or robots designed to entertain and amaze by performing on stage (e.g. a Rubic's Cube solving robot). Note however that actions performed autonomously by the robot but unnoticed by people may in a sense be irrelevant in terms of user experience.

Importantly, from an action- and experience-centred perspective, software is understood in terms of tools and resources, asking designers to address the *interactive* features of the technology, in other words, how actions that users perform are taken up and mediated by the device (rather than the other way around). This includes considerations of what users do to control, program, update and communicate digitally through the robot, and how the technology succeeds in responding to those expectations.

In terms of computer memory, central to this is therefore how the recorded data are presented and communicated to its users as well as to others. As this will necessarily have to be performed slightly differently on different interactive hardware platforms, the memory model may need to be taken into consideration also how the model is used in terms of controlling the recording, communication and access of the data.

5 Discussion

Based on the action-centric framework, our ethical memory model design needs to consider four dimensions of interaction: physical manipulation, perception and sensory

experience, contextually oriented action and digitally mediated action.

The first one is the physical manipulation and concerns the way data in the memory may be physically recorded, accessed and manipulated. How this should and could be done necessarily differs between users as well as between the physical properties of the different devices that the recordings are stored on. Another aspect to take into account is that the user is usually unaware of how the data are physically stored, possibly resulting in a poor interaction and jeopardising the sense of "companionship".

The second is the perception and sensory experience, where the user perception of how and when data will be recorded and how it is presented is important. In other words, the external design of the artificial companion could have an impact on the user's perception and thus his/her expectations of when and what are being recorded in the companion's memory.

As for the contextually oriented action, consideration needs to be given on establishing rules or guidelines to ensure appropriate socio-technical deployment. This could mean that the device should not be allowed to collect and record information indiscriminately. That includes the role definition in terms of one's responsibility during the maintenance and manipulation of data in the context of use. We should also reflect on the social context in which the artificial companion is inserted for this could dictate its behaviour in terms of which information could be disclosed and when.

Finally, the digitally mediated action dimension calls for attention on the control and communication of memory data in terms of interaction features. A relevant aspect here is to define how the artificial companion will react to the user actions of *control*. This would impact on how to better design the companion's memory in order to comply with these *orders*. We believe that adaptation and the ability to evolve are going to be intrinsic properties of the memory model in respect of the long-term interaction and particularly crucial in this case.

Following these guidelines, our companion's memory can, for instance, be personally tailored to suit particular user *needs* while initializing the system. The same memory architecture, with different levels of forgetting mechanisms to handle sensitive contents, can support various user groups with regard to personal privacy. As an example, a system working in an office can be *personalised* to remind workers their schedule, meeting appointment and regular break times; however, this system may be *taught* to avoid remembering workers' personal information such as someone's home address or salary, because these are believed to be sensitive issues to individuals in the office environment. Remind that this will be due to repression of personal information (i.e. the information still accessible in other circumstances). In contrast, a system used in the home environment can store more personal information at users' *request*. This system can also help user with daily tasks at home and also remind the user when to take medicine, appointments with a doctor and/or of personal dates.

6 Conclusion

We argue that if technologies that autonomously collect and store digital data are to become acceptable to users, then the computational requirements, as well as the social consequences must be understood and addressed. From a technological perspective, interaction style and mechanisms, visual appearance, memory, responsiveness to human display of affect, security and privacy, are only some of the areas that must be investigated in order to develop long-life personalized companions.

By considering ethical issues, we have proposed a novel ethical memory model for artificial companions. Basically, we advocate that an established set of moral rules (a deontological system), dependency on specific user and system and a learning process (virtue based) and a prediction scheme (consequentialist) should be part of the "ethical" system together with the aforementioned forgetting mechanisms. In this way, the user could control "what" is being stored, "how" it is being encoded and to "whom" it would be available.

We have also analysed the possible ethical impacts of the proposed memory model by applying an action-centric framework. This framework proved to be extremely useful to evaluate our model by providing insightful and interesting views on the interaction between users and their artificial companions. As a result, we could highlight many relevant aspects that should be taken into account while developing future prototypes of an ethical memory.

This work is a step further towards the development of an enhanced memory model taking into consideration ethical issues involved. Hence we continue to explore what information an artificial companion should remember, forget and also generalise in order to generate appropriate ethical behaviours and thus smooth the interaction with the human user in a long-term perspective. To the best of our knowledge, this research question has not yet been proposed or discussed in our field of study.

References

Allen C, Wallach W, Smit I (2006) Why machine ethics? IEEE Intel Sys 21(4):12–17

Altmann EM, Gray WD (2000) Managing attention by preparing to forget. Int J Cogn Ergon 1(4):152–155
Ambo P (2007) Mechanical love, a film directed by Phie Ambo
Anderson JR (2007) Kogniotive psychologie. Eine Einführung (6th edn) Heidelberg
Anderson SL (2008) Asimov's "three laws of robotics" and machine metaethics. AI Soc 22(4):477–493
Arkin RC (2009) Governing lethal behavior in autonomous robots. CRC Press
Asimov I (1950) I robot. Doubleday and Co, New York
Brom C, Lukavsky J (2009) Towards virtual characters with a full episodic memory II: the episodic memory strikes back", AAMAS, pp 1–8
Brom C, Peskova K, Lukavskyz J (2007) What does your actor remember? Towards characters with a full episodic memory. In: Cavazza M, Donikian S (eds): ICVS. vol 4871 of lecture notes in computer science, Springer, 89–101
CHRIS (2008) The CHRIS project, http://www.chrisfp7.eu/index.html
Clarke R (1993) Asimov's laws of robotics: implications for information technology. Cambridge University Press, Cambridge
Companions (2007) Companions project, http://www.companions-project.org/
Denning T, Matuszek C, Koscher K, Smith JR, Kohno T (2009) A spotlight on security and privacy risks with future household robots: attacks and lessons. In Proceedings of the UbiComp 09, 30 sept–3 Oct 2009
Endo Y (2007) Anticipatory robot control for a partially observable environment using episodic memories. Georgia Tech Mobile Robot Lab Tech Rep, GA
Fernaeus Y, Tholander J, Jonsson M (2008) Beyond representations: towards an action-centric perspective on tangible interaction. Int J Arts Technol 1(3/4):249–267
Fernaeus Y, Jacobsson M, Ljungblad S, Holmquist LE (2009) Are we living in a robot cargo cult? In: 4th ACM/IEEE international conference on human robot interaction, 9–13 March 2009, La Jolla, California, USA
Friedman B (ed) (1997) Human values and the design of computer technology, Cambridge P
Gert B (1988) Morality: a new justification of the moral rules. Oxford University Press, Oxford
Gips J (1995) Towards the ethical robot. In: Ford K, Glymour C, Hayes P (eds) Android epstemology. MIT Press, cambridge
Gold PE (1992) A proposed neurobiological basis for regulating memory storage for significant events. In: Winograd E, Neisser U (eds) Affect and accuracy in recall: studies of "flashbulb" memories Vol 4. Cambridge University Press, New York, pp. 141–161
Halpern S (2008) Can't remember what i forgot: the good news from the front lines of memory research. Harmony Books, New York
Ho WC, Lim M, Vargas PA, Enz S, Dautenhahn K, Aylett R (2009) An initial memory model for virtual and robot companions supporting migration and long-term interaction. In the 18th IEEE ROMAN'2009, Japan, pp 277–284
Ishikawa M (1996) Structural learning with forgetting. Neural Netw 9(3):509–521
Jacobsson M (2009) Play, belief and stories about robots: a case study of a pleo blogging community In proceedings of Ro-Man: IEEE
Kaplan F (2004) Everyday robotics: robots as everyday objects. In proceedings of joint sOc-EUSAI conference, pp 59–64
Kelley JF (1983) An empirical methodology for writing user-friendly natural language computer applications. In proceedings of ACM, pp 193–196 ACM
Klemmer SR, Hartmann B, Takayama L (2006) How bodies matter: five themes for interaction design. In proceedings of DIS '06. ACM Press. 140–149
Koychev I (2000) Gradual forgetting for adaptation to concept drift. In: Proceedings of ECAI 2000 workshop current issues in spatio-temporal reasoning, Berlin
Lim M, Aylett R, Ho WC, Enz S, Vargas PA (2009) A socially-aware memory for companion agents. IVA'2009, Amsterdan, pp 20–26
LIREC (2008) The LIREC project, http://www.lirec.org/
Mirza N, Nehaniv C, Dautenhahn D, te Boekhorst R (2006) Interaction histories: from experience to action and back again. In: Proceedings of the 5th IEEE international conference on development and learning (ICDL), 2006
Mirza N, Nehaniv C, Dautenhahn D, te Boekhorst R (2007) Grounded sensorimotor interaction histories in an information theoretic metric space for robot ontogeny. Adapt Behav 15(2):167–187
Murphy R, Woods DD (2009) Beyond Asimov: the three laws of responsible robotics. IEEE Intel Syst 24(4):14–20
Nolfi S, Floreano D (2000) Evolutionary robotics: the biology, intelligence, and technology of self-organizing machines. MIT Press
Nuxoll A, Laird J (2004) A cognitive model of episodic memory integrated with a general cognitive architecture. In: International conference on cognitive modeling
Ogino M, Ooide T, Watanabe A, Asada M (2007) Acquiring peekaboo communication: early communication model based on reward prediction. In: Proceedings of IEEE international conference in development and learning (ICDL)
Schank RC, Abelson R (1995) Knowledge and memory: the real story. In: Robert S, Wyer Jr (eds) Knowledge and memory: the real story. Lawrence Erlbaum Associates, Hillsdale, NJ, pp 1–85
Sparrow R (2002) The march of the robot dogs. Centre for applied philosophy and public ethics, The University of Melbourne
Sparrow R (2006) Killer Robots. J Appl Philos 24(1):62–77
Syrdal DS, Walters ML,Otero NR, Koay KL, Datenhahn K (2007) He knows when you are —privacy and the personal robot. AAAI-07
Tecuci D, Porter B (2007) A generic memory module for events. In proceedings of the 20th Florida Artificial intelligence research society conference
Vargas PA, Ho WC, Lim M, Enz S, Aylett R (2009) To forget or not to forget: towards a roboethical memory control, to appear at the AISB social understanding of AI workshop, Edinburgh, Scotland, pp 18–23
Vargas PA, Freitas AA, Lim M, Ho W, Enz S, Ayllet R. (2010) Forgetting and generalisation in a memory model for robot companions: a data mining approach. In the proceedings of the AISB2010 convention, De Montfort University, UK
Veruggio G (2005) The birth of roboethics," ICRA 2005, IEEE international conference on robotics and automation workshop on Robo-Ethics, Barcelona
Veruggio G, Operto F (2006) Roboethics: a bottom-up interdisciplinary discourse in the field of applied ethics in robotics. International review of information ethics, vol 6, pp 3–8
Wallach W, Allen C (2009) Moral machines: teaching robots right from wrong. Oxford University Press