Intelligent agents in an e-literate society: some ethical considerations

Carolyn Dowling
Australian Catholic University, 115 Victoria Parade, Fitzroy 3065, Victoria, Australia

c.dowling@patrick.acu.edu.au

Key words: Ethical/Ethics, Future, Internet, Social Issues, Values

Abstract: One of the fastest growing applications of AI research is the implementation of computer programs commonly referred to as 'agents'. This type of software is distinguished from more traditional programs by a high degree of autonomy in decision making and action, the ability to 'learn' from experience and to adapt their behaviour accordingly, and often a highly personified interface. Many are specifically designed to process complex information, make decisions and initiate actions in 'mission critical' areas of human endeavour including health, scientific research, government, business, defence, the law and increasingly in education. While in some cases we are aware of our interactions with these electronic entities, in many contexts their activity takes place 'behind the scenes', at a level not apparent to the user. Implicit in our conception of an agent both in the physical world and in cyberspace is the notion of delegation. Important aspects of this concept are our understandings of features of human interaction such as trust, responsibility, privacy and our capacity to judge competence and intention. Consideration of these issues in relation to the activities of software agents could lead to the formulation of a broadly based code of 'agent ethics'. This could help in regulating some aspects of agent behaviour and act as a foundation upon which common expectations on the part of users might be formulated.

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35663-1_34

T. J. van Weert et al. (eds.), *Informatics and the Digital Society*

1. INTRODUCTION

Until relatively recently, the notion of intelligent software agents belonged in the realms of speculation and esoteric research. Today it is one of the fastest growing applications of Artificial Intelligence and is already a well accepted aspect of our everyday computing experiences. Some agents, such as the variety of 'personalities' available to assist us with anything from the smallest to the most complex computer based task, are visible to us and are, to a certain extent, amenable to our direct control. Others, particularly in online environments, undertake their activities at a level not apparent to the casual user.

These programs have several distinguishing features in common. One of the most important is a high degree of autonomy or independence. In computing contexts this means that the program is able to make decisions and initiate actions without the need for constant monitoring and intervention on the part of a human being. Another key attribute is the ability to 'learn' from experience and to respond flexibly to changing situations. This may cause such programs to behave in ways not clearly predicted by the original programmers. For many purposes, a highly personified interface suggesting the possession of a believable 'character' or personality provides an important basis for the agent's interaction with users.

Early definitions of agent software, based more on speculation than on experience, tended to emphasise the broader 'human' aspects of these programs, for example:

"an agent is a character, enacted by the computer, who acts on behalf of the user in a virtual environment" considered useful in mediating "a relationship between the labyrinthine precision of computers and the fuzzy complexity of man." (Laurel 1990)

Typical of more functional definitions is:

"An agent can be viewed as an object which has a goal and autonomously solves problems through interaction, such as collaboration, competition, negotiation and so on." (Kinoshita and Sugawara 1995)

From another perspective, an agent exhibiting these characteristics can be described as having:

"a set of beliefs about its environment and about itself; a set of desires which are computational states which it wants to maintain, and a set of

intentions which are computational states which the agent is trying to achieve." (O'Hare 2000)

The widespread implementation of agent based systems raises a number of issues that have clear ethical, even moral dimensions. They include the balance between autonomy and control and questions related to trust, responsibility and privacy.

2. IMPLICATIONS OF THE 'AGENT' METAPHOR

Why have we been so quick to accept software agents as an essentially unproblematic element in our computing environment? A large number of our day to day activities currently take place within computing environments characterised by rapid change, large quantities of extraordinarily complex information and a lack of common organisational structures through which information may be accessed and managed. It is clear that, as Laurel predicted, there are now many situations in which some form of 'intelligent' mediation is required between computer systems and the needs of users. Agents are a means of masking the complexity that lies behind our use of computers, thereby facilitating the achievement of our various aims and goals.

The term 'agent' strikes a comforting note of familiarity with many computer users, most commonly conjuring up an image of a docile and amenable servant undertaking particular tasks at the behest and in the interests of the user. As a way of conceptualising our relationship with the technology it has a great deal in common with an earlier metaphor which proved extremely popular, that of the computer as a 'tool'. Both of these images encompass the reasssuring suggestion that the technology is under the control of the user. In each case, however, the metaphor is less straightforward than it appears and the confidence and security engendered in users may well be misplaced. In many instances the notion of an agent masks not just a functional complexity, but also a range of activities which are not of the users' choosing and not necessarily even in their interest.

Both in the physical world and in cyberspace, the concept of an agent is understood and instantiated in a number of different ways. Some of the most interesting types of human 'agents' to consider in relation to the ethical and moral complexities of the roles undertaken by software agents are those which actually represent the interests of more than one party and must effect a balance between competing claims. Examples in the physical world include estate agents, theatrical or literary agents and employment agents.

Another type of agent with which we are all familiar but are probably less comfortable, is the 'secret agent', generally a gatherer of supposedly confidential information operating under an assumed identity. This model is particularly relevant to some current practices on the Internet. Far from being initiated through a contractual arrangement entered into voluntarily by the user, interaction with many such agents is an invisible and frequently unsought component of an apparently innocuous task such as a search for information.

Common to all of the understandings of the roles and functions of human agents outlined above is the assumption that their usefulness derives from the possession of specialised skills, which qualify them to mediate between an individual and a particular environment for the more effective achievement of various ends. Furthermore, there exists a commonly held understanding that most human agents can be relied upon to act in accordance with an understood set of ethical principles that ensure the client or user is not disadvantaged. Such assumptions play a significant part in encouraging the acceptance of programs characterised as 'agents'.

3. THE ROLE OF PERSONIFICATION

In circumstances where the user is intended to be aware of the activities of the software and to interact knowingly with it, the agent metaphor lends itself particularly well to reinforcement through a personified interface. Anthropomorphic elements have been implicit in most computer interfaces from the earliest days. Intelligence and language use are widely accepted as the key criteria distinguishing us both from other living things and from inanimate objects. The very use of the term 'intelligence', albeit 'artificial', in relation to computer programs, in combination with the fact that for the most part we interact with them through language, adds support to the perception that a software agent is 'one of us', and should be subject to the types of expectations, including that of trust, that govern our interactions with our fellow human beings.

Not only are we accustomed to interacting with computers as though they share with us a degree of 'humanity', but in a number of areas of activity we value 'social interaction' particularly highly. An example is the field of education, where our current understandings of learning depend very much on an acceptance of the role of the social construction of knowledge. The high value we place on the 'social' also lends acceptability to the interlocking activities of multiple agents when they are metaphorically characterised as a 'society' (Franklin and Graesser 1996; Costa and Perkusich 1997).

However, while we might acknowledge as inevitable the inadvertent attribution of human qualities to computers and their programs, the deliberate cultivation of personified or anthropomorphic interfaces is more problematic. Both research and experience suggest that a mismatch between realism in appearance and the apparent knowledge level of the agent as revealed through its use of language and other capabilities can have a deleterious effect on credibility and on acceptance. As Masterton, writes:

"A common problem with AI programs that interact with humans is that they must present themselves in a way that reflects their ability. Where there is a conflict between the ability of the system and the users' perception of that ability a breakdown occurs and users may either fail to exploit its full potential or become frustrated with its shortcomings." (Masterton 1998)

Agents that 'look' smart and 'act' or 'talk' dumb are poorly received by many users, who express a higher tolerance for the limitations of a 'character' more sketchily represented, for instance through cartoon-like graphics, including those characterised as normally inanimate objects or as animals (Chan 1998). This may provide some counterbalance to our tendency to unquestioningly invest a high level of trust in such entities. Here too, however, there are complexities, although of a psychological rather than an ethical nature. While in many instances we may feel more comfortable giving instructions to a 'character' that is not fully human, we may not feel the same way about receiving advice or even correction from such an entity. Users of Microsoft Office will be all too familiar with the indignity of being subjected to the whims of an animated paperclip.

The issue of the degree to which software agents can and should be personified is extremely important when considering the ethical aspects of agent systems. It is this aspect of agency that propels questions of right and wrong to a level beyond that pertaining to our use of more traditional types of programs. To the extent that personification is successfully implemented, the expectations of users in relation to the behaviour of the program move into the arena governing the understandings upon which we base our interpersonal interactions, rather than merely those upon which we base our use of technology. The problems are accentuated in cases where the agent is not depicted visually but emulates human dialogue to the extent that users may not be aware that they are interacting with a computer program. Such a situation is inherently and deliberately misleading and is a poor basis for the development of the type of trust in agency that arguably should be an important part of a fruitful ongoing relationship between human beings and software agents.

4. AUTONOMY, TRUST AND RESPONSIBILITY

The delegation of any task to a software agent raises questions in relation to its autonomy of action and decision, the degree of trust which can be vested in its outcomes and the location of responsibility, both moral and legal, for those outcomes.

Many of these agents are specifically designed to weigh up complex information, make decisions and initiate actions in 'mission critical' areas of human endeavour including government, health, scientific research, commerce, the law and defence, with a significant degree of autonomy being intrinsic to their usefulness. Implicit in the recognition of the need for a capacity to exercise initiative is an acknowledgment that some outcomes of agent activity may not be easily predictable by the user. In some cases they may even be contrary to what the user might perceive as his/her interests and wishes.

A further issue emanating from consideration of agent autonomy is that of responsibility for outcomes and actions resulting from decisions which are out of the user's control. Indeed they may relate to capacities of the program of which the user had no knowledge at all. It is inevitable that such instances will occur, precipitating the need for renewed examination both of community and legal understandings of liability.

Closely associated with our ambivalence towards agent autonomy is the issue of trust. How might we decide whether or not to 'trust' an agent? In the physical world a reputation for credibility in most fields is contingent on a verifiable history demonstrating qualities such as accuracy, reliability and efficiency. Where such assurances are available in relation to the software agents to whom we delegate responsibility, perhaps we can be justified in taking the risk of 'trusting' them. As with some forms of delegation to other human beings, however, we may have to accept that there are no absolute guarantees.

Even if we are prepared to trust the agents with whom we deal directly, a further question remains as to how agents might reasonably 'decide' how to trust one another. Implications for users include important considerations in regard to privacy. For instance, the main threats to privacy have been defined by one researcher as:

"Threats caused by agents acting on behalf of the user (through loss of control of the user on (sic) his agent and the disclosure of the user's personal information)" and "Threats caused by foreign agents that act on behalf of others (information extraction via traffic flow monitoring, data mining and even covert attempts to obtain personal information directly from the user's agent)." (Van der Lubbe 2000)

5. A CODE OF ETHICS FOR SOFTWARE AGENTS?

An important element in our decisions to invest trust in expert human agents is our belief that their actions will be governed by a code or codes of ethical behaviour, both at a personal and a professional level. Adherence to a code of ethics is, in fact, a key element in the definition of a profession.

What might a code of ethics for software agents look like? Given that they could well be described as the disembodied 'robots' of cyberspace, an appropriate place to start might be Isaac Asimov's Three Laws of Robotics, familiar to several generations of science fiction readers. A generalised version of the original laws is the directive that a robot may not injure humanity or, through inaction, allow humanity to come to harm.

AI researcher Marvin Minsky acknowledges issues similar to those recognised by Asimov in relation specifically to software agents when he writes:

"There's the old paradox of having a very smart slave. If you keep the slave from learning too much, you are limiting its usefulness. But, if you help it to become smarter than you are, then you may not be able to trust it not to make better plans for itself than it does for you." (Minsky1994)

Some current research in this area expresses the problem in terms of the need for the types of 'norms' that constrain behaviour within human society (Verhagen 2000).

6. CONCLUSION

There is no doubt that the use of those AI applications we refer to as 'agents' is increasing in a range of areas critical to our well being both as individuals and as members of society. Without necessarily implying any intent on the part of software developers to promote agent behaviours which we would regard as unethical, it would seem prudent for ethicists and programmers to collaboratively give serious consideration to the formulation of an appropriate ethical code that, if accepted, might give users of agent technologies a level of confidence in the performance of these entities comparable with that which is currently vested in human beings.

REFERENCES

Chan, T. -W. (1998), *The past, present, and future of educational agents.* [http://www.apc.src.ncu.edu.tw/apc/ppt_chan.html]

Costa, E.deB. and Perkusich, A. (1997) Designing a Multi-Agent Interactive Learning Environment. In *Proceedings of ICCE 1997*, International Conference on Computers in Education, Z. Halim, T. Ottmann and Z. Razak (eds.), Charlottesville, VA.

Franklin, S. & Graesser, A. (1996) Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In *Proceedings of the Third International Workshop on Agent Theories, Architectures and Languages*, Springer-Verlag, Berlin.

Kinoshita, T. and Sugawara, K. (1995) *Agent oriented computing*. Soft Research Centre, Tokyo.

Laurel, B. (1990) Interface agents: metaphors with character. In *The Art of Human-Computer Interface Design*, B. Laurel (ed.), Addison-Wesley, Reading, Massachusetts.

Masterton, S. (1998) Computer support for learners using intelligent educational agents: the way forward. In *Global Education on the Net Vol. 1*, T. W. Chan, A. Collins and J. Lin (eds.), Proceedings of ICCE '98, the Sixth International Conference on Computers in Education. China Higher Education Press and Springer-Verlag.

Minsky, M. (1994) A conversation with Marvin Minsky about Agents, *Communications of the ACM*, July 1994, vol. 37, no. 7.

O'Hare, G. (2000) Agency, Mobility and Virtuality: A Necessary Synergy. In *Proceedings of ISA 2000*, International ICSC Congress on Intelligent Systems and Applications. ICSC Academic Press, Canada, pp. 15 -21.

van der Lubbe, J. (2000) Incorporating Privacy Enhancing Technology in Intelligent Software Agents. In *Proceedings of ISA 2000*, International ICSC Congress on Intelligent Systems and Applications. ICSC Academic Press, Canada, pp. 222-228.

Verhagen, H. (2000) Trust and Norms for Artificial Agents. . In *Proceedings of ISA 2000*, International ICSC Congress on Intelligent Systems and Applications. ICSC Academic Press, Canada, pp. 277-280.