Artificial societies and psychological agents

Stuart Watt

KMI-TR-33

September 1996

A revised version of this paper is to appear in the British Telecom Technology Journal special issue on Intelligent Agents, Autumn 1996

Artificial societies and psychological agents

Stuart Watt
Knowledge Media Institute and Department of Psychology
Open University
Walton Hall
Milton Keynes. MK7 6AA.

Email: S.N.K.Watt@open.ac.uk

Abstract

Agents have for a while been a key concept in artificial intelligence, but often all that the word refers to is a computational process or task with a capability for autonomous action, either alone or in an artificial society of similar agents. But the artificial nature of these societies restricts the flexibility of agents to a point where social interaction between people and agents is blocked by significant social and psychological factors not usually considered in artificial intelligence research. This paper argues that to overcome these problems, it will be necessary to return to the study of human psychology and interaction, and to introduce the concept of 'psychological agents.'

1. Introduction

There are several different kinds of artificial intelligence research. First, there is research on applications—on building a new generation of systems by borrowing problem solving techniques through analogies with physics, biology, and psychology, and creating a discipline that is mostly intended as an advancement on computer science. Second, there is research on artificial intelligence as a methodological tool in psychology, building psychological models and using them to study how the human mind works; trying to find insights on the nature of human intelligence. There are other kinds of research in this field, but it is the interplay between these two principal themes that is the focus of this paper.

Recently there has been a new link between these two research themes through a joint growth of interest in common sense—or 'folk'—psychology, particularly as expressed in the form of systems which can reason about explicit goals, intentions, and beliefs; either their own or other people's. In psychology, this has lead to a dramatic growth of interest in 'theory of mind' [e.g. 1, 5, 20]. In artificial intelligence, much of this research has been driven by distributed artificial intelligence, which also needed to build systems which can reason about each other—systems which are usually called ‘agents’ [2].

‘Agent’ is a difficult word for a difficult concept; covering a rag bag of concepts that span a whole gamut of different kinds of behaviour, including, for example, autonomy, learning, and social interaction. But there is a common ground. An agent will set out to do something, and do it; therefore it has competences for intending to act, for action in an environment, and for monitoring and achieving its goals. Of course, for adequate performance of these, other competences, such as learning, negotiation, and planning, may be helpful or even necessary.

This is not the whole story. Agency is a lot more than action in an environment, or, rather, the environment is not just a simple passive system. Often the environment will contain other agents—which is why social interaction and collaboration are so often stressed as a feature of agency. More interestingly, perhaps, the environment may even contain people, leading to the human kind of agency—the kind we talk about in terms like ‘estate agent.’ Agents are embedded in an environment, but this environment is social as well as physical; social not only in that an agent is working with other agents, but social in that an agent must also work with people. The environment, therefore, and the social rules that apply, are those of human social behaviour.

Work in artificial intelligence has never really addressed the problems of binding together its agents in human societies to the same degree as has the field of human-computer interaction. In artificial intelligence agents are designed to form unrealistic social systems, or, rather, they take valid models of realistic social systems and interpret the models too literally and too strictly. The human components of conflict, morality, and responsibility, for instance, are all simplified out of existence and, therefore, agents have real problems in human societies, except in small niche contexts where people can accept these limitations. The result is that agents are not usually flexible enough to be able to work effectively in human societies.

The true challenge for artificial intelligence is to remove this fault-line separating its agents from human societies. We do not need to do it all at once; we do not necessarily need to do it by making the agents truly intelligent; human societies can adapt to some extent, too. But, at the end of the day, we must make this shift for agents to become more than yet another temporary technological innovation.

In this paper I will argue that this separation can be overcome by drawing on agent ideas from human-computer interaction [e.g. 13, 16] and using them to create a more psychological and sociological background for agents. In the next section, I will discuss two dimensions against the concept of an agent can be measured, and in section 3 how agents face up to the pressures of action in human social environments. Section 4 introduces the key theme of the paper—psychological agents—and argues that a new dependence on human psychology is necessary for effective agents. Section 5 follows this up with a case study of how even simple agents can follow this path.

2. On agents

Before we can look at the fault-line between agents and human societies in more detail, I need to be a bit clearer about what I mean by an agent. People in artificial intelligence use the word ‘agent’ in several different senses, and it is important to be clear about which of these is meant. I will discuss two different dimensions along which the concept can be measured; first, the contrast between metaphorical and ideal interpretations of the concept, and, second, between internal and external points of view regarding an agent’s behaviour. I will look at the interpretation of the concept first.

One possibility is that the concept of an agent is metaphorical. In this sense, the concept of an agent is mainly a tool for thinking with—a paradigm if you like. This certainly seems the intention behind some of the uses of the word [e.g. 9, 24]. But, for this, I do not think the case is especially clear that the concept of an agent is substantially different from a hybrid of two existing computational concepts, the task and the object. If people are using terms like ‘agent-oriented programming’ in order to introduce this new paradigm, it should be pointed out explicitly that this is a metaphorical use of the word ‘agent.’

The second possibility is that the concept ‘agent’ is something more of an ideal to which our computer agents are still only an approximation. In this sense, the ideal concept of an agent is the kind of agency that we humans are familiar with—and the technical concept is just the best we have been able to do so far.

As is probably obvious, I want to advocate the second, idealistic, interpretation. If the first, metaphorical, interpretation is to be accepted, then I would suggest that agents are a temporary technological innovation, one to abandoned in the face of later, more sophisticated ideas at the next paradigm shift in the programming community.

Despite this, there are some real advantages to using agents metaphorically, as characters if you like; they make it possible to develop solutions to problems which are better structured than might otherwise be the case. [31] shows that there may be many different ways to describe a system as a set of agents—some of which are traditionally rooted in distributed artificial intelligence, where others bring out human perspectives that are often hidden in knowledge-based systems. However, even this human emphasis already leans a little to the second, idealistic, interpretation. We find here that one of the key advantages of the agent approach is that the structure and format of interaction becomes far closer to the structures and formats of human interaction.

So there are advantages to the second, more idealistic, interpretation of the concept; it leaves room for a far more human kind of agency. Taken to the limit this would mean that agents have to be full-blown artificial intelligences, but this does not necessarily inhibit us from making good use of the technology in the meantime. We can still use the concept metaphorically, but explicitly as an interim step to keep our feet on the right path. If we do not make this explicit, the arguments about what constitutes an agent might eventually become a somewhat pointless terminological dispute.

There is a second way to cut up the concept of an agent, taking either an internal or an external stance to explaining its behaviour. In artificial intelligence agents are described principally in terms of the internal states, the desires, beliefs, and goals, over which the agent has control. This is appealingly close (perhaps too close) to the common sense notions of folk psychology, but raises many problems about what these ‘desires,’ ‘beliefs,’ and ‘goals’ really are, since they are clearly not the same as human desires, beliefs, and goals, at least, they are not in the current state of the field. The question is, is the difference one of degree or of kind? Are they metaphorical or ideal? Most workers in this field sidestep the issue completely by pretending that these words are being used as metaphors, mostly so they do not get attacked by philosophers on what they see as a non-critical issue.

Human-computer interaction, by contrast, uses the word ‘agent’ for any active entity that will take on a user’s goals and act on them. Typically it means “extending everything we do to be part of a *grand collaboration* with one’s self, one’s tools, other humans, and increasingly, with *agents*” [13, original emphasis]. The human-computer interaction literature avoids describing what is going on *inside* an agent, falling back to an intuitive definition of agent as something which initiates and performs actions. On one level, they are merely mirrors of a user’s goals, and are no more agents than the Eliza program [32] is a psychotherapist. On another level, they are clearly agents; they are capable of acting on their own, or, rather, their users treat them *as if* they are capable of acting on their own—and this is what really matters. It is the behaviour, the whole behaviour, and nothing but the behaviour that counts.

Of course, the internal and external views cannot be completely separated. Neither is an adequate description, and in practice most people prefer to adopt a view somewhere between the two. Besides, they both focus on an individual agent, so there is something missing from *both* views: society.

3. Getting into the social context

No agent is an island. It is the social context that helps to define the boundaries and the behaviour of any agent. Agency is a social and a psychological phenomenon rather than just a biological or physical one. It is the social structure that an agent participates in that shapes its action.

But human society is not constant. Society changes rapidly. Consider the telephone. It is more than just another form of communication, because there is a strong element of presence. “When you talk on the telephone your face and body still emit expression, even though you know full well that the person at the other end can see none of it” [19]. There is a perceptible difference in people’s attitudes to each other using the different media: some people are typically angrier on the telephone than in face-to-face conversation, and others more polite. Why should a new form of communication have such strange and subtle effects on people’s attitudes?

Agents can communicate in two different ways: either between themselves or with others outside their immediate group. This distinction is sometimes very clear and sometimes it is cloudy, but it can serve to highlight the distinction between the traditional artificial intelligent approach to agency and the psychological/social approach to agency that I am advocating. In the traditional artificial intelligence approach to agency, agents communicate with each other through a specialised language which is usually designed as a set of different kinds of speech acts [23].

For an agent to communicate with people, it needs to talk a human ‘language’—or at least to be able to communicate on human rather than machine terms. That, after all, was the whole point of the Turing test [28]. We can, of course, learn to infer a machine’s inner states from its expressions—be they panels of flashing lights or what have you, but this is, for us humans, a fundamentally foreign language. And instead of forcing people to learn a foreign language, would it not be better to teach the agents our language—after all, it is they who are beginning to participate in our society.

There is a fundamental difference between the ways people communicate by language and the ways that agents communicate by language. Usually agents interact through a formal made-up language, a kind of techno esperanto, perhaps something like KQML [7]; these languages are, of course, pretty hopeless for people. Proper agents must communicate using languages that people can understand. We should design agents which can interact using human natural language, as far as we can get them to understand it. Sure, this is a lot harder for us as designers, but the payoff, potentially, is a way of binding together human and agent societies far more effectively, as humans and agents begin to communicate through the same language. That does not necessarily mean that agents need to be full natural language systems to be proper agents, although in the long run that may be required. More, it is intended to suggest that agents should communicate with each other using the same channels that people do, and that the communications themselves should be in human forms rather than machine ones. Even a stylised, rigid, but at least human-comprehensible version of plain text, perhaps using forms, is better than an arcane language like KQML.

This focus on language can be misleading; by ‘language’ I mean something much richer than text—covering the whole range of human communicative acts. Consider this example: when I switch on my Macintosh computer it smiles at me, to tell me that the computer is ‘happy’. I recognise this because I know that in my human social context, people smile when they are happy. I have borrowed from my experience of human society to help me understand what the computer is feeling. (This is rather anthropomorphic. The designers of the Macintosh *used* this anthropomorphism for precisely this purpose: it helps people work with the computer.) I do not need to know anything about computers to ‘read’ this cue—even a child could do it—I borrow this skill from my natural, human, common sense psychology [17].

Agents in artificial intelligence need to be more human-like both in their behaviour in the traditional psychological sense as well as in their social context. That is, they need to be able to ‘read’ all our expressions of our inner mental states to be able to collaborate and interact with us appropriately. The study of agency in artificial intelligence has so far taken an over-simplified view of the effects of human psychology and society; it has created artificial societies, artificial social contexts in which its own kind of agency has a valid status. In order to create *real* artificial agents, these assumptions need to be lifted, the gap between these artificial social contexts and the reality of human society needs to be closed.

Artificial intelligence has usually tried to take its models from human behaviour, and when trying to build models of social interaction it was to human social systems that it turned. In practice, the models that were developed all represented more or less plausible models of the ways that agents could interact, but when turned into formal descriptions which could be implemented all the elasticity implicit in the original model was lost. Unfortunately, it was this elasticity that enabled the society to work effectively and to adapt to new circumstances, and these formalised models lost all this elasticity because they were all unrealistically rigid compared to human societies and organisations.

For example, even when a human social system is nominally called hierarchical, as a large company sometimes is, there may be many direct links between the members of the structure aside from those that make up the hierarchy. Engineers working on different projects may meet in the corridor over coffee, and the exchange of ideas can benefit all. A production-line worker may meet the managing director when they are in the greengrocer's, and each may gain insight into the problems of the other. The hierarchical structure on its own is too rigid to work effectively when real people are involved, and human social systems of any scale are never as pure and uniform as they are usually represented.

Natural human social structures are more complex than those applied in artificial intelligence because the human kind of intelligence has evolved hand-in-hand with these social systems. If we are to build agents which can live in our human social systems, we need to transfer some of psychology to them, by one means or another, so they can participate in and see our societies from our point of view. To do this, we need to look at how human psychology affects human interaction—and then use the lessons from this to restructure the concept of an agent to fit into these same psychological principles. This is what I mean by 'psychological agency.'

4. Psychological agents

The relationship between me and you—and even between me and a computer—is a social and psychological one, and a set of social rules apply which help me to interpret the behaviour of those I interact with, whether they are people or computers. Of course, the social relationships between me and you and between me and a computer are superficially very different, reflecting different sets of social assumptions, but there is all sorts of interplay between them and in many ways they are closely tangled: "people's expectations about human-computer interaction are often inherited from what they expect from human/*human* interaction" [3, original emphasis]. This very kinship opens up an immense possibility for conflict when there is a dissonance between these expectations and reality—when the expectations from human-human collaboration conflict with the reality of human-computer interaction.

Although people will readily attribute some kind of agency to many computer systems, this is really anthropomorphism. People inevitably anthropomorphise their computer—not because they are told to—but because it is part of the way people relate to each other, and they use this to 'read' the computer. Computers are, after all, social objects rather than just physical ones—and people apply social and psychological principles when interacting with them [18]. "At the grossest level, people simply attribute agency to the computer itself ('I did this, and then the computer did that'). They also attribute agency to application programs ('My word processor trashed my file')" [15].

Human societies and individuals have a human flexibility which can be added, if needed, to make sense of a situation. This requires at the minimum a kind of “naive physics” [11]; an ability to make commonsense predictions about the behaviour of objects, but also—and more importantly—a kind of ‘naive’ (what [12] calls “natural”) psychology. This natural psychology is not the same as academic psychology, more it is the ability of humans to understand and predict the behaviour and feelings of other humans. It is the psychology of motivation as well as that of cognition, dealing with feelings, emotions, and moods, recognising them and interpreting their effects on peoples’ behaviour. This links back to the origins of agent theory, as it was this common sense psychology that was the source of the explicit goals, intentions, and beliefs that lead to the advent of agents in artificial intelligence. The difference is that, from necessity, artificial intelligence has over-formalised these concepts in adopting them, and shut out the human psychology that originally underpinned them.

Humans have this natural psychology, this ability to understand their own and other people’s essentially human mental states, including their goals, intentions, and beliefs. It is part of the glue that holds human society together. For artificial intelligence agents to gain first-class status within our human societies, they must be able to reason with and communicate about these same—essentially human—mental states. ‘Alien’ machine kinds of mental state can, and possibly even do, exist, but that is simply not relevant; it is human mental states that are the fabric of human society and for artificial intelligence agents to have a status within our society these agents must have the same kind of human common sense psychology.

So human psychology has a fundamental effect on what it is to be an agent. Now we can start to reconstruct the concept of an artificial intelligence agent in these terms. To show this most clearly, I will propose a new model for agents which links three different levels—and kinds—of agency.

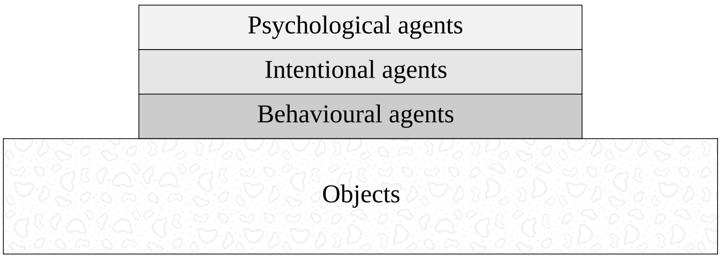

Figure 1. A three layer model of agency

The most primitive level is what we might call ‘behavioural’ agency. This is the aspect of agency which is most directly concerned with action. At this level, something is an agent because it acts with apparent autonomy. This is the level at which an alarm clock is an agent. There is nothing in this kind of agent which would normally correspond to a ‘belief’ for example, about another agent. A behavioural agent can act, even act autonomously, but it has no knowledge of or ability to reason about another agent’s goals, intentions, or beliefs.

The next level is what we might call ‘intentional’ agency. At this level, agents have intentional states up to an arbitrary number of levels. This means that they can not only behave as before, they can also have explicit beliefs about other agents, and intentions and goals which involve acting indirectly through other agents, as well as or instead of acting directly for themselves. At this level agents are capable of reasoning about themselves and other agents in their society with the whole gamut of beliefs, desires, and intentions, to an arbitrary degree. Intentional agents are capable of shared commitment, negotiation, and deception, for example, and all the other behaviours that are associated with the beliefs, desires, and intentions model [4, 21].

So far so good. This is the traditional artificial intelligence model. I now want to suggest that a third level is necessary, ‘psychological’ agency. In part this is a stopgap, because research does seem to suggest that a “representational theory of mind” [8] which corresponds to second level, intentional, agency, is not enough, *even in principle*, to build human agents [22]. The distinction I have in mind here is that at this third level, agents can use common sense heuristics to guess at, as well as to reason about, human intentional states. They are also capable of ascribing rationality to agents—that is, distinguishing autonomously between things which are agents and things which are not [6, 25]. Finally, they should be able to perceive and recognise all the ‘backchannels’ in human interaction, the half smiles, nods, and frowns that frame and can completely change the meaning of any linguistic interaction they accompany. All this requires changes in an agent’s perceptual apparatus as well as its psychology.

There are a number of implicit assumptions here. First and foremost, by asserting that psychological agency is different from intentional agency, I am also implicitly asserting that there is something more to human psychology than intentional states. This is a claim that is not certain either way, but the claim that intentional states are sufficient (as opposed to merely being necessary) for psychology is not really supported by the evidence [20, 22]. There does seem to be a gap between intentional states and human psychology.

The fundamental reason for my emphasis on human psychology lies in the claim that true machine intelligence must be isomorphic with true human intelligence. Turing once said of the Turing test [28], “might not a machine do something which might be called thinking, but which is unlike what people do?” I think he was wrong about this. Machines could already be fully conscious beings on that basis—if we accept that people can not recognise them as intelligent, because they possess some ‘alien intelligence.’ Computers could already be classed as superintelligent in an alien sense for their exceedingly fast numerical processing, but that is not what I class as intelligent. I want to argue that for intelligence to mean anything, it must be the kind of thing that other people—other *people*—recognise and are prepared to call ‘intelligence.’

As things stand, then, a useful goal for artificial intelligence research would be to study and to develop models of agency which are truly psychological, not just intentional in level. Of course, this is a long term project—one which will undoubtedly take at least my lifetime—but that does not mean we should not try. It also does not mean that we can not reap a useful benefit from systems which are only vague shadows of this in the short term, and I will discuss this, with an example, in the next section.

5. Steps to the grand collaboration: Luigi

Agents change the way we work with programs. In the future, neither an individual nor a social, not an internal nor an external, view, will be sufficient to describe how agents will work. Agents will work both with people and with other agents, in a form we might call ‘heterogeneous groupware.’ Agents have ceased to become objects, and have become a medium in their own right—a distinction that [13] describes as the shift from manipulation (of objects) to management (of agents). This offers both problems and opportunities for agent design.

In human-computer interaction an interface agent is often a character living in the computer acting on behalf of someone in a virtual environment, with a degree of autonomy. Sometimes this is taken to extremes, giving the agent a human, name, face, and voice—making the agent seem like a virtual person. But interface agents are not without their problems. There is a psychological price to be paid for the anthropomorphism that is built into many interface agents. Just because an agent has a human name, looks human, or speaks like a human, it does not mean that people will interact with them as if they are human; the behaviour of the agent has to live up to this expectation, and if it fails there is a kind ‘anthropomorphic dissonance’ which undermines the collaboration. (This is another way of seeing the conflict I mentioned earlier between expectations inherited from human-human interaction and the reality of human-computer intention.) People get frustrated if they have to negotiate work indirectly through an over-anthropomorphised agent [“some dip in a bow tie,” 15] rather than acting directly when they already know what to do.

In order to study how agents interact with people, we selected a domain where people interacted with each other and with computer programs: diary and meeting management. Meeting management software has been available for many years, and has been the focus of efforts by project management specialists, user interface specialists, workgroup software specialists, and artificial intelligence specialists. But despite this interest, diary and meeting management software has never been as successful as the more conventional forms of collaborative communication, such as the telephone or electronic mail.

This kind of system has been permanently a victim to the ‘weakest link in the chain’ phenomenon: unless everyone in the workgroup plays by the rules and keeps an on-line accessible electronic diary up to date—a diary which is compatible with everybody else’s software—the underlying premise of the workgroup software falls apart. This is typical of the failure of much computer-supported cooperative work, in that there are different people in different roles, and the benefits of cooperation fall unequally on the roles and leads to a breakdown in cooperation when cooperative actions cost people in some roles more than they benefit. We can call this ‘role conflict’ [29].

These observations, together with in-house experience of one of the major commercial packages and observation of the ongoing patterns of electronic mail among human meeting makers led us to investigate a ‘least common denominator’ approach. ‘Luigi’ is an agent which embodies this approach.

Luigi communicates using a widespread medium, and interacts directly with people, so that they do not need to keep up a diary, or run any special software at all. A proposer can send a message to Luigi requesting a meeting, saying who is invited, how long the meeting will last, and suggesting a number of possible dates. Luigi then sends mail to the delegates and manages the meeting for the proposer, negotiating the possible dates and keeping the proposer informed on the progress of the meeting, and when a date has eventually been agreed, it asks the proposer to confirm the meeting for that date. Luigi, then, acts for the proposer but interacts with all the potential participants.

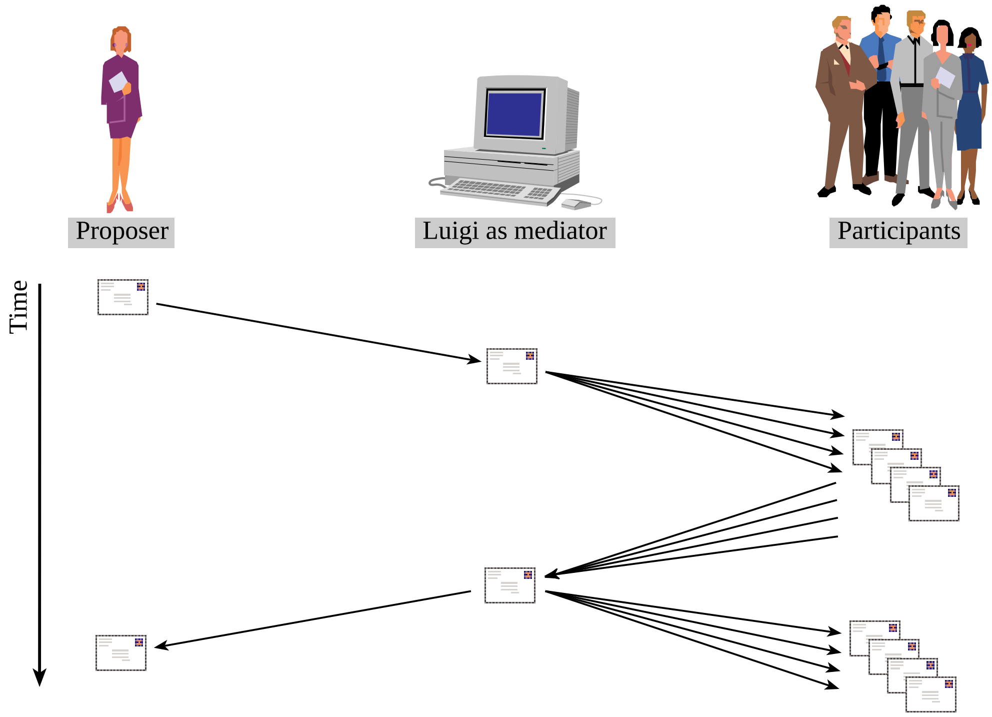

Figure 2. Messages to and from Luigi

Figure 2 shows how Luigi acts as an intermediary between the meeting's proposer and the invited participants. Luigi takes away some of the administrative burden, but has to be careful to ensure that the proposer does not feel they have lost control of the meeting.

This strategy means that Luigi, building on a medium with relatively few implicit role biases, such as electronic mail [10], can help where other meeting systems cause a breakdown between the roles by assisting some roles (meeting organisers, in this case) without any additional cost to others in keeping electronic diaries up to date. That is, Luigi avoids role conflict. More agents can be added, each addressing a specific role, and gradually a complex society of human and artificial agents will emerge. Separate agents can be added for each participant to delegate automatic acceptance of certain meeting requests—these agents can then negotiate directly with Luigi without having to bother the meeting's proposer.

Luigi is a complete prototype system. It can successfully plan meetings using a forms interface, either with simple textual forms in mail messages or using forms on the world wide web. But the prototype is more a sketch to let us try out ideas for this kind of agent, and to explore the psychological effects of different media and different message content on the collaborative processes.

How does Luigi fit into our model of psychological agency? Luigi clearly is not anything remotely resembling a full-blown artificial intelligence—in fact, it more shows a greater attention to the psychology of the human proposer and participants than it does to fancy reasoning with beliefs, desires, and intentions. This is, though, an area where we expect Luigi to develop substantially in the future.

But even so Luigi is a lot more than a naive scheduler. Luigi has to accept real responsibility for a sensitive kind of discussion—that involved in planning a meeting. The texts that it uses for its messages have to be very carefully toned to avoid invoking feelings (such as those of alienation or frustration) in those it is bringing together. It is a system which must interact with people and with agents, and where it has to use well phrased natural language for communication and for negotiation. With this in place, we can work to make Luigi gradually better at dealing with the human psychological processes which make real meeting planning and negotiation so hard.

The important part of the approach for us is that we do not regard the successes and failures of Luigi as ‘bugs’ but as hints about how people, psychologically and socially, respond to agents. These hints can then be used to in turn to refine the designs of future agents to work better with people.

Luigi shares some features with other interface agents for meeting scheduling, such as those of Kozierok and Maes [14] and Sycara and Zeng [27], but there are important differences. Maes’ approach to interface agents [16] stresses collaboration with an individual user in a single human-computer interface, where Luigi collaborates with a group rather than an individual. Secondly, Maes’ agents operate within the user’s work environment, where Luigi takes on a task to the extent that it passes partly outside that single work environment. Sycara and Zeng’s agent is primarily a agent-oriented implementation of a solution to the problem of planning a visit with collaborating information agents; there is no strong connection with the user. On both scores, the fundamental difference between Luigi and similar agents is that Luigi is situated, in the sense of Suchman [26], but situated in the human social world of collaborating to set up meetings.

The fundamental principles of this approach: heterogeneous groupware mixing people and agents, careful avoidance of any extra work burden on any participant, and a wide variety of forms of interaction, seem both theoretically sound and practically useful in both the short term and the long term. Luigi shows the immediate practical utility of agency in a modified form, in a safe middle ground between the Scylla of anthropomorphic agency and the Charybdis of pure manipulation. It allows us to explore the psychology of interaction between humans and agents in a relatively controlled environment. As such, I believe that this is where research on intelligent agents should be focused in the near future.

6. Conclusions

There are some themes which vaguely resemble hype in the current interest on agent technologies. At the core, however, there are some very deep psychological issues, some of which are old and some new. All potentially offer a way to a new generation of systems which are designed to collaborate with each other and with people, and which are better able to cope with the social systems in which they have to operate.

First, I have suggested that agents should collaborate with people on human terms, even human psychological and social terms, rather than on formal abstract mathematical or logical ones. This is clearly a lot harder for us as researchers now, but I believe that gradually elevating agents to our human level is going to be necessary, eventually.

By advocating a three layer model of agency, I intend to throw new emphasis on the psychology of agent interaction—namely on the issues of how people recognise something as an agent, and how people ascribe new mental states to these agents. This is in contrast to the previous, intentional states, model, which mainly focuses on the logical procedures which can be used to make inference about existing mental states.

As a research programme, this is, fundamentally, as hard as true artificial intelligence—indeed, it can be argued that this is the key to true artificial intelligence [30]. It should not be regarded as the kind of problem that we can crack in a matter of a few years.

In the interim, then, I suggest that we just take the ideas and not be too worried about the purity of the psychological principles that underpin them. Artificial intelligence has got on very well borrowing ideas from psychology without feeling that it has to adhere to them too closely. We can already build systems which fit into human societies, which take on responsibilities, and which interact autonomously with each other and with people. We can already use psychological principles in the design of these systems. And, furthermore, by looking at the successes and failures, the strong and the weak points of these systems, we can indirectly discover the pure psychological principles that are necessary for true agency.

This approach to agency is fundamentally quite radical. Instead of building progressively more complex agents, I propose building progressively more *human* agents. At first we will not succeed; the agents will both be and appear to be quite mechanical, but that must not put us off our long term goal. We must not get distracted into designing more and more artificial societies for our agents to act and interact in; instead we must remember that we are human, and begin to design more and more human agents. And even now, even in the short term, we must start to see our agents less as tools, and more as assistants—assistants which respond to our needs as people and talk to us in our language, without expecting us to talk to them in their language. Underneath all that hype, there still lurks the glimmer of true psychological agency—we must not let it get buried under technology or caged in artificial societies.

Acknowledgements

Luigi was designed collaboratively by Marc Eisenstadt, Zdenek Zdrahal, and the author. The author is especially indebted to Marc Eisenstadt for moral and textual support, to Tamara Sumner and Hyacinth Nwana for making this paper much better than it was, and to all the participants in British Telecom’s 1995 agents workshop for an exciting exchange of ideas and perspectives.

References

[1] S. Baron-Cohen, H. Tager-Flusberg, and D. J. Cohen, “Understanding Other

Minds: Perspectives From Autism," . Oxford: Oxford University Press, 1993.

[2] A. H. Bond and L. Gasser, "An Analysis of Problems and Research in DAI," in *Readings in Distributed Artificial Intelligence*, A. H. Bond and L. Gasser, eds. San Mateo: Morgan Kaufmann, 1988, pp. 3-35.

[3] S. E. Brennan, "Conversation as Direct Manipulation," in *The Art of Human-Computer Interface Design*, B. Laurel and S. J. Mountford, eds.: Addison-Wesley, 1990.

[4] P. R. Cohen and H. J. Levesque, "Intention Is Choice with Commitment," *Artificial Intelligence*, vol. 42, pp. 213-261, 1990.

[5] M. Davies, "The Mental Simulation Debate," in *Objectivity, Simulation and the Unity of Consciousness*, C. Peacocke, Ed., 1994, pp. 99-128.

[6] D. C. Dennett, *The Intentional Stance*. Cambridge, Massachusetts: MIT Press, 1987.

[7] T. Finin, D. McKay, R. Fritzson, and R. McEntire, "KQML: an information and knowledge exchange protocol," in *Knowledge building and knowledge sharing*, K. Fuchi and T. Yokoi, eds.: Ohmsha and IOS Press, 1994.

[8] J. A. Fodor, "Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade-Mecum," *Mind*, vol. 94, pp. 55-97, 1985.

[9] M. R. Genesereth and S. P. Ketchpel, "Software Agents," *Communications of the ACM*, vol. 37, pp. 48-53, 1994.

[10] J. Grudin, "Groupware and Cooperative Work: Problems and Prospects," in *The Art of Human-Computer Interface Design*, B. Laurel and S. J. Mountford, eds.: Addison-Wesley., 1990.

[11] P. J. Hayes, "The Naive Physics Manifesto," in *Expert Systems in the Microelectronic Age*, D. Michie, Ed. Edinburgh: Edinburgh University Press, 1979, pp. 242-270.

[12] N. K. Humphrey, "The Social Function of Intellect," in *Growing Points in Ethology*, P. P. G. Bateson and R. A. Hinde, eds. Cambridge: Cambridge University Press, 1976.

[13] A. Kay, "User Interface: A Personal View," in *The Art of Human-Computer Interface Design*, B. Laurel and S. J. Mountford, eds.: Addison-Wesley, 1990, pp. 191-207.

[14] R. Kozierok and P. Maes, "A Learning Interface Agent for Scheduling Meetings," in the proceedings of the ACM SIGCHI International Workshop on Intelligent User Interfaces, Orlando, Florida, 1993.

[15] B. Laurel, *Computers as Theatre*. Reading, Massachusetts: Addison-Wesley, 1991.

[16] P. Maes, "Agents that Reduce Work and Information Overload," *Comunications of the ACM,* vol. 37, pp. 31-40, 1994.

[17] J. McCarthy, "The Little Thoughts of Thinking Machines," *Psychology Today*, vol. 17, 1983.

[18] C. Nass, J. Steuer, and E. R. Tauber, "Computers are Social Actors," in the proceedings of CHI'94, 1994.

[19] N. Negroponte, "Hospital Corners," in *The Art of Human-Computer Interface Design*, B. Laurel and S. J. Mountford, eds.: Addison-Wesley, 1990, pp. 191-207.

[20] J. Perner, *Understanding the Representational Mind*. Cambridge, Massachusetts: MIT Press, 1991.

[21] A. S. Rao and M. P. Georgeff, "BDI agents: from theory to practice," in the proceedings of the First International Conference on Multi-Agent Systems, ICMAS'95, San Fransisco, USA., 1995.

[22] J. Samet, "Autism and Theory of Mind: Some Philosophical Perspectives," in *Understanding Other Minds: Perspectives from Autism*, S. Baron-Cohen, H. Tager-Flusberg, and D. J. Cohen, eds. Oxford: Oxford University Press, 1993, pp. 427-449.

[23] J. R. Searle, *Speech Acts*. Cambridge: Cambridge University Press, 1969.

[24] Y. Shoham, "Agent Oriented Programming," *Artificial Intelligence*, vol. 60, pp. 51-92, 1992.

[25] T. R. Shultz, "From Agency to Intention: A Rule-Based Computational Approach," in *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading*, A. Whiten, Ed. Oxford: Basil Blackwell, 1991, pp. 79-95.

[26] L. A. Suchman, *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge: Cambridge University Press, 1987.

[27] K. Sycara and D. Zeng, "Visitor-Hoster: Towards an Intelligent Electronic Secretary," in the proceedings of the CIKM-94 (International Conference on Information and Knowledge Management) workshop on Intelligent Information Agents, 1994.

[28] A. M. Turing, "Computing Machinery and Intelligence," *Mind*, vol. LIX, pp. 433-460, 1950.

[29] S. N. K. Watt, "Role conflict in groupware," in the proceedings of First

International Conference on Intelligent Cooperative Information Systems, Rotterdam, Netherlands., 1993.

[30] S. N. K. Watt, “A Brief Naive Psychology Manifesto,” *Informatica*, vol. 19, pp. 495-500, 1995.

[31] S. N. K. Watt, Z. Zdrahal, and M. Brayshaw, “Multiple Agent Systems for Configuration Design,” in *Frontiers in artificial intelligence and applications*, J. Hallam, Ed.: IOS Press, 1995, pp. 217-228.

[32] J. Weizenbaum, “ELIZA: A Computer Program for the Study of Natural Language Communication between man and machine,” in *Communications of the ACM*, vol. 9, 1966, pp. 36-45.