Robust Normative Systems: What happens when a normative system fails?

Peter Wallis

The NLP Group, Department of Computer Science, The University of Sheffield
pwallis@acm.org

Abstract. Computers that can hold a conversation, such as chatbots on the web, embodied conversational agents (ECA) or automated call handling systems are, by the agent model of software, autonomous agents situated in a social world. As social animals, we humans rely on social norms that we are barely conscious of. In this paper it is argued that 1) these normative systems have a layered structure, and 2) current conversational agents only work at the top layer. People abuse such systems, not because they fail, but because their response to failure is inappropriate.

1 Introduction

Creating a machine that can hold a conversation is a difficult problem, the solution to which would have many applications. In the eighties it was felt that the problem was one of simply having the resources to collect enough data, but today it seems there is something fundamental missing from our understanding of how language works. The agent metaphor of provides an alternative to the idea of computers as strictly information processors. The classic approach to language is to see it as a conduit for meaning [10]. Parsing is seen as one step in mapping text to its meaning [9], and dialog is seen as a means of updating the information state of the hearer [7]. The agent paradigm suggests a different model in which a conversational agent *acts* in a social context. If we are treating conversational agents as social actors – and not just conveyors of information – the question arises then to what extent must they rely on other social skills. What is it that makes a human trust the information presented by a automated call centre or by a virtual tutor? What makes a character in a computer game engage us emotionally, and to what extent can a virtual sales assistant get a visitor to divulge personal information about his or her spending habits and interests? In a previous paper I've argued that intentionality is key [11], but it is not intentionality 'all the way down.' This paper is about the nature of language generation once we stop thinking about it.

The observation made here is that in human / human conversation, people fail gracefully, and what is more, they do it without thinking. The hypothesis is that our ability to do this is part of our social intelligence – the process is part of our mechanism for dealing with the intra group pressures of being a social animal. In this paper the mechanism we use to cope with other people is

characterised as a *normative system.* That is, individuals have sets of behaviours that they normally do, and these individual behaviours fit together like a jigsaw puzzle to form the fabric of society. The key issue is that these norms or protocols can be broken. People can cheat, and the idea presented here is of a *robust* system of norms in which abuse is a key mechanism.

2 Norms for Social Actors

Although Margaret Thatcher didn't think so, societies are more than a collection of individuals. In societies people cooperate to do things such as build cathedrals and go to war. The nature of cooperation can be *described* with a set of rules. Some of these rules are explicit and prescriptive, while others are hardly available to the conscious mind. One can imagine for instance that each honey bee in a hive works to a set of shallow rules that make its behaviour mesh with that of other members of the hive. Bees can navigate past each other in a crowded passage, pass information about the location of food sources, and defend the hive all as part of a cooperative behaviour, presumably, without understanding their role in the process.

Sometimes this cooperation is not in the interests of the individual. Honey bees, famously, will sacrifice themselves to defend the hive. From the perspective of the selfish gene [2] one can see how such altruism would come about. A queen bee creates worker bees that have rules of behaviour that cause the individual to sacrifice itself for the good of the hive. This provides an environment for the queen's genes to prosper, which creates more bee societies with selfless worker bees.

The argument is that such rules work in human society as well. Sure people can reason about their behaviour, but such reasoning is constrained. I am polite[1] to strangers and enjoy going to the pub; I get nationalistic in the face of terrorism, I gossip [3], and buy dolls with big eyes [6]. Why? Because I am a human and humans are programmed to do those things. Without those things, I would not trust the bank, I would have to hoard food through the winter and worry about protecting such resources from my neighbour. The distinction, between rational mean-ends reasoning about action, and reactive behaviours, is made in economics. This is Jon Elster [4] introducing social norms:

> One of the most persistent cleavages in the social sciences is the opposition between two lines of thought conveniently associated with Adam Smith and Emilie Durkheim, between *homo economicus* and *homo sociologicus.* Of these, the former is supposed to be guided by instrumental rationality, while the behaviour of the latter is dictated by social norms. The former is "pulled" by the prospect of future rewards, whereas the latter is "pushed" from behind by quasi-inertial forces (Gambetta, 1987). The former adapts to changing circumstances, always on the lookout for improvements. The latter is insensitive to circumstances, sticking to the prescribed behaviour even if new and apparently better options become

Fig. 1. Images from the dogAttack movie discussed in Kubinyi et al.

> available. The former is easily caricatured as a self-contained, asocial atom, and the latter as the mindless plaything of social forces.

He goes on to discuss attempts by economists to reduce norm-oriented action to some type of optimising behaviour. The interest here is not in discussing the nature of economic good and evil however.

The problem is of course that, unlike honey bees, individual humans (actually their genes) have their own interests to look after. What is more, we are often smart enough to be able to reason about the outcomes of our actions. With a little thought an agent might become an 'asocial atom' and cheat.

3 Robust Normative Systems

Unlike the protocols of computer science, social norms have a certain robustness about them. Rather than building a formula one racing car where every piece is optimised up to, but not beyond, the point of failure, normative systems in human societies are more like military aircraft where structures are often designed such that no individual component is critical. When an A130 hits the supports for a cable car, it is the people in the cable car who are killed, not the air crew. This robustness of design is key to effective normative systems where there is a chance individuals might cheat. The fabric of society must have some means of handling cheats and the proposal is that the mechanism is simply another norm. In order to make a system of norms robust, there must be *second order norms* that guide individuals back in (see [5]) and keep society operating. Buying rounds in a pub is a social norm that gives advantage to individuals that can skip their turn to buy. Such an individual is however soon bought into the fold.

Consider Figure 1 showing three stills from the dogAttack movie by Kubinyi et al [8] who have been using a Sony Aibo to study animal interactions. The movie shows a dog eating some meat, and on the other side of the food is an Aibo – the Sony robot dog – in a crouching position. The Aibo stands up. The dog stops eating, bares its teeth and growls. The dog goes back to eating. The Aibo takes a step forward toward the food and the dog stops eating; growls; and grabs the Aibo by the neck and throws it across the room.

What is happening here? One interpretation of what we are seeing is that the Aibo is pressing the dog's buttons in such a way that it triggers a hard-wired

social norm. The warning growl is a *second order* norm that leaves space for young dogs to learn the first order norm that adult male dogs eat first. If dog societies had no interest in socialising the young (in the same way as spiders and crocodiles) there would be no warning and the adult would simply kill any transgressor in order to keep its food. Puppies, and humans, are hardwired to recognise such second order norms but Aibos, and chatbots, are not.

4 Conversational Agents

In 2001 we set up a wizard of oz experiment with a student as the wizard, pretending to be a machine. The very first conversation went like this:

PROGRAM 0:00 Hello and welcome to the Melbourne University campus guide. How can I help you?
CALLER1 0:00 Moron.

What should the wizard say? The caller is obviously testing the program and it is perhaps interesting that our understanding of computers leads us, these days, to start the conversation with testing. Here is a conversation between a machine and a human from 1996 in which the human doesn't start by testing, but ends up doing so:

PROGRAM 0:47 G'day mate.
JUDGE02 0:55 G;day. How's it going?
PROGRAM 1:11 My mother once said "How it is going is a mystery." She was a bit nuts.
JUDGE02 1:22 Uh-huh.
PROGRAM 1:31 Are you still learning to speak, my friend?
JUDGE02 1:52 NO, no. But your anecdote took me a bit by surprise.
JUDGE02 1:52 How was your morning?
PROGRAM 1:59 Oh well, that's a pity!
JUDGE02 2:15 I'm sorry; what's a pity?
PROGRAM 2:20 Apologies accepted.
JUDGE02 2:35 I feel compelled to insult you, and we've only been talking a minute.
PROGRAM 2:49 We are bloody well talking, aren't we?
JUDGE02 3:14 Oh baby. We sure are. So here are some keywords for you: shit mother baseball

Whatever it is that has annoyed him or her, the judge offers a chance for the system to recover, which the system fails to take up. Like the dog in the video, the judge warns the machine; the machine's response to this second order norm is not right, and the judge 'attacks.' Note that the judge's response starts out quite angry and abusive – the agent has pressed the buttons of the judge – but then the keywords become more acceptable. I would suggest that the milder (and more rational) behaviour is primarily a product of knowing that his or her response would be viewed by others. On the positive side, note that the machine – in both the chatbot and Aibo cases – is being treated as a social actor by the interactant. The problem is not to make a machine that is accepted, but to make it behave itself once it is accepted as an actor in the appropriate social context.

5 Conclusion

Ants and bees live in communities where the fabric of society can be expressed as a normative system. Each agent is given a set of norms that make it fit within the mechanisms that enable the nest/hive to survive and reproduce. The proposal is that people still use such rules, but can also think about their actions. Whereas insect communities might use a normative system that pulls action from individuals, self conscious agents can reflect on their role and start to act based on self interest. The normative system for these agent communities must be *robust*, and abuse is part of this process. Abuse is the fore runner to actual harmful action and as such leaves space for individuals to change their anti social behaviour. Whereas humans and puppies are hardwired to know what these second order behaviours mean, Aibo's and chatbots need to be told. This is the challenge, I believe, that stands between us and the creation of effective human-machine conversation.

References

1. Penelope Brown and Stephen C. Levinson. *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987.
2. Richard Dawkins. *The Selfish Gene*. Oxford University Press, 1976.
3. Suzanne Eggins and Diana Slade. *Analysing Casual Conversation*. Cassell, Wellington House, 125 Strand, London, 1997.
4. Jon Elster. Social norms and economic theory. *Journal of Economic Perspectives*, 3(4):99–117, 1998.
5. Harold Garfinkel. Conditions of successful degradation ceremonies. *American Journal of Sociology*, 61:420–424, 1956.
6. Daniel Harris. *Cute, Quaint, Hungry and Romantic: the aesthetics of consumerism*. Basic Books, 10 East 53rd St, New York, 2000.
7. Jörn Kreutel and Colin Matheson. Modelling dialogue using multiple inferences over information states. In *Proceedings of ICOS-2, 2nd Workshop on Inference in Computational Semantics*, Dagstuhl, 2000.
8. Enikö Kubinyi, Ádám Miklósi, Frédéric Kaplan, Márta Gácsi, ózsef Topál, and Vilmos Csányi. Social behaviour of dogs encountering AIBO, an animal-like robot in a neutral and in a feeding situation. *Behavioural Proceses*, 65:231–239, 2003.
9. I. Mel'cuk. Semantic primitives from the viewpoint of the meaning-text linguistic theory. *Quaderni di Semantica*, 10(1):65–102, 1989.
10. Michael J. Reddy. The conduit metaphor: A case of frame conflict in our language about language. In Andrew Ortony, editor, *Metaphor and Thought*. Cambridge University Press, 1993.
11. Peter Wallis. Believable conversational agents: Introducing the intention map. In Catherine Pelachaud, Elisabeth Andre, Stefan Kopp, and Zsófia Ruttkay, editors, *Creating Bonds with Humanoids (Proceedings of the Workshop at AAMAS'05)*, Utrecht University, the Netherlands, july 2005.