Identifying Your Accompanist*

Will Lowe
will.lowe@uni-mannheim.de
Mannheim Centre for European Social Research (MZES)
University of Mannheim

> Computer scientists are designing digital companions for people – that is, virtual conversationalists, or digital confidants.
>
> They would [...] build a life narrative of the owner. [...] You could call it autobiography building for everyone,
>
> [...] inner content could be elicited by long-term conversations.

These quotations are taken from the first paragraph of the invitation to this forum on Artificial Companions in Society. They express a view about the *nature* of such companions, the *outcome* of extended interaction with such companions, and the *method* of their interaction. In short, Companions are agent-like others; they help us understand ourselves; and they do by way of conversation.

But there is an obvious tension between the first and second parts – the mention of autobiography gives it away: Are these companions truly *others*, or are they *us*?

Rather than attempt to say whether they are, or should be thought of as other agents or as filtered versions of ourselves, I would like to take one step back. What Companions are or should be will depend on what we want them to do for us. And this depends on what they *can* do for us. In the end, I shall argue that whether Companions should best be thought of, programmed, and regulated as others, or as extensions of ourselves, is a tactical question. One path will simply be more effective than the other. Either way, it's going to be all about us. To make a start, switch to the second person: What can a Companion do for you?

We begin with a practical and not-to-distant use case: a Companion can remind you to take your medicine on time, and to ensure that you finish the course. For medical issues it this will be easier and work better if the Companion has access to your medical records. This is both empowering, because you are immediately more integrated in the project of maintaining your health, and controlling, because it ensures that advice you get from doctors cannot be so easily ignored. And it has significant social benefits, allowing large amounts of accurate epidemiological data to be harvested via your Companion's other (one might hope suitably anonymised) interactions with those interested in public health, from both research and policy perspectives. Indeed, Companion-like projects already exist to keep the

*This paper was written for the *Artificial Companions in Society* meeting held at the Oxford Internet Institute, University of Oxford, October 26th, 2007 while the author was Research Fellow at the Methods and Data Institute at the University of Nottingham. An edited version appears as: Lowe, W. (2010) 'Identifying your accompanist' in *Close Engagements with Artificial Companions*, Y. Wilks (Ed). Benjamins. pp.95–99

demented out of institutional care for longer by making their environments more 'intelligent'[1]. Other cognitive deficits may also be best addressed by a Companion approach: the National Health Service already makes use of computer-based cognitive behavioural therapy courses for phobias, panic attacks, and depression. Indeed, such approaches may work better with a Companion involved; it will have more information about your local environment to tailor your treatment and monitor your progress. More generally, the idea of an extended period of agent interactions in order to provide a framework for understanding your choices, obligations, and relationships could well be described as autobiography-building. It is also called therapy.

One of the more optimistic possible roles for a Companion is to make you more rational. Romanticism aside, it is better and less frustrating if your preferences order consistently over currently available goods, and over time[2]. Consider these two types of well-ordering.

It is well known that psychological preferences are sensitive to the framing of choices. Usually, the framing is done by another, such as the store whose window you are looking into, or the state whose tax code you are attempting to negotiate. The advantage of a Companion is that it can reframe your choice set, perhaps according to criteria that you my have told it about beforehand.

The possibility of using a Companion as a level of indirection for choices is also helpful when the choices are difficult because they require expertise you do not have, or because they are easy to make now, but hard to make later. This latter is a ubiquitous planning problem, referred to generically as inter-temporal choice. Whether it is the cream cake forbidden by your diet, the last week of antibiotic treatment, or contributions to your holiday fund, when the sirens of plan-breaking temptation sing, it may be your Companion that ties you to the mast.

The problem of inter-temporal choice can be described by individual discounting curves: assume that for any good, the utility of having it today is greater than tomorrow, and still less the day after. Classically rational agents discount their utilities at a constant rate, e.g. every extra day spent waiting reduces utility by 10%. This generates a curve of expected utility that decreases exponentially towards the present moment from its maximum when the good is available. The exponential form guarantees that utilities retain their rank ordering, so preferences over goods will never reverse. Most mammals, however, discount at a variable rate depending on the proximity of a good, typically tracing 'hyperbolic' curves that can cross, leading to situations where the expected utility of a larger but later good is initially greater than and then at a certain point less than the utility of a smaller sooner one, leading the latter to be chosen. Fortunately for planning, the curve traced by a good explicitly constructed by bundling together many lesser goods over time is more nearly exponential than the curve of any of its components[3].

While there is agreement on the form of the cognitive problem, diverse mechanisms have been suggested to explain it, including folk-psychological theories of 'will power', religious theories of 'temptation', and most recently in neurophysiological accounts[4]. For the purposes of thinking about Companions, George Ainslie's account is particularly suggestive[5].

[1] e.g. work at Newcastle University's Institute for Aging and Health

[2] or as economists would put it: if you *have* preferences, rather than simply unrationalisable choice behaviours.

[3] See e.g. K. N. Kirby and B. Guastello (2001) Making choices in anticipation of future similar choices can increase self-control, *Journal of Experimental Psychology: Applied*, 7, pp.154–164.

[4] See e.g. S. McClure, K. Ericson, D. Laibson, G. Loewenstein, and J. Cohen (2007) Time discounting for primary rewards. *Journal of Neuroscience*, 27, pp.5796-5804

[5] See e.g. G. Ainslie (2005) Précis of Breakdown of Will, *Behavioral and Brain Sciences* 28, pp.635–673; G. Ainslie (1991) Derivation of "rational" economic behaviour from hyperbolic discount curves, *American Economic*

Begin by noting the structural similarity between the problem of inter-temporal choice and the prisoners dilemma in game theory. In a one shot game the equilibrium policy is to defect, gaining more utility than the other player but less than if you both co-operated. Other policies are equilibria if the game is repeated. Concrete proposals usually suggest reputation (as a co-operator) as a mechanism to maintain long term optimal behaviour despite the temptations of short term gains realised by defection.

In Ainslie's framework you are engaged in just such a repeated game with your future selves. If you can frame sets of goods that will occur at different times as parts of a larger good then discounting for the larger good will become more exponential, and thus less liable to be trumped by a sooner smaller rewards. Ainslie describes this cognitive process as bundling, typically a mixture of asserting an equivalence across actions – all drinks are equally bad for me – and a personal rule – because I don't drink. To the extent that breaking such a rule by succumbing to a nearby pint affects the probability that it will happen again, rule breaking incurs a reputation cost (your reputation to yourself). There is therefore always a motivation to interpret each rule breaking as a justifiable exception – drinking at a birthday – which in turn may perversely motivate systematic misunderstanding of oneself and the world. The dynamic of setting personal rules, sometimes breaking them, and then dealing with the consequences for your self image can therefore be described as inter-temporal bargaining.

Ainslie describes the mechanisms of decision in an unaided agent, but the presence of a Companion changes things. Crudely, a Companion may act as enforcer, distracting you from proximate but self-defeating pleasures, or telling you off afterwards. More interestingly, a Companion may simply inform you about what you've been doing, making it harder to maintain a transient pleasure-justifying narrative that you would be ashamed of later. Companions are also plausible repositories for personal rules that discourages opportunistic rewriting – a commitment mechanism similar to illiquid retirement savings plans[6]. A Companion may also independently assess certain predictive probabilities. This may be a problem: If it is necessary for maintaining your current course of action to believe that every 'exception' signals the end of your personal rule and triggers a steep reputation cost, then it may not be helpful to have a Companion inform you that there is in fact a 0.8 probability that your latest 'lapse' will not be repeated[7]. In short, a Companion might support you in making choices where "Those sinkings of the heart at the thoughts of a task undone, those galling struggles between the passion for play and the fear of punishment, would there be unknown."

The previous passage is taken from Jeremy Bentham's discussion of the Panopticon and its advantages as a school design[8]. The possibilities sketched above spell out the irreducibly Panoptical nature of Companionship. Our two possible Companion types appear as extremes: Either the Companion is another – e.g. the watching eye and punitive hand of the state, even if only its health service – or the Companion is a cooler headed version of yourself, a programmable super-ego. Interesting Companions no doubt lie between these extremes, but the the notion of inter-temporal bargaining brings out the common Panoptical core of each. This core promises (or threatens) to make you transparent to others, and to yourself.

It is conventional to deplore Bentham's Panopticon, but the common core of surveillance and intervention may be as desirable for constantly renegotiating the extended bargain of a coherent self as it is offensive when imposed by a state. The watchful Companion can

Review 81, pp.334–340

[6]See e.g. C. Harris and D. Laibson (2001) Dynamic choices of hyperbolic consumers *Econometrica*, 69(4), pp.935–957

[7]Naturally, a sophisticated Companion would do its calculation conditioning on both the value itself and its general criteria for telling you, leading fairly directly to a version of Newcombe's Problem.

[8]in J. Bentham (1995) *The Panopticon Writings.* M. Bozovic (ed.) Verso, London.

exist at any point between, so its construction obliges us to be clear about the nature of the difference.

While a Companion might make inter-temporal planning easier when all goods have been bundled and choices framed by increasing the information available to us about ourselves in choice situations, we have not addressed the question of how choices become framed to start with. Considering this prior process – "which behaviours count as 'my diet', and what constitutes a 'lapse'?" – and the permanent possibility of on-the-fly redefinition demonstrates the limits of Companion-driven transparency. For example, the probability of not lapsing again is only 0.8 under some bundling of goods and plans, some set of equivalence classes over actions. But it is different or undefined in other (possibly non-commensurable) decompositions of the social world. Every probability needs a sample space, and your Companion will partly constitute it.

In the first example: You tend to accept the information and interventions of your doctor. This is partly because you share her assumptions, and because if you cease to trust her judgement you can find another doctor, resort to prayer, or switch to herbal remedies – each embodying a distinct conceptual scheme and rule set that a Companion could help you work within[9]. Panoptical objections arise most quickly when your Companion is chosen for you, particularly by a state. This is reasonable: Not only is there no general reason to assume that state's bundling of goods and plans is the best one for you, but states too have inter-temporal choice problems[10]. But problems also arise when you choose for yourself. A Companion can impose the thinking and values of an entire community on you, in a way that is unaffected by subsequent changes in that community's thinking. For example, a Companion may still be 'helping' you ward off spells when a ceasefire agreement has already been signed between the other members of your religion and modern medicine. This is problematic despite the fact that you chose your Companion freely, perhaps to bolster your faith.

Returning to the original question. It is an empirical matter whether the inter-temporal bargaining we seem to need to stabilise our choices and construct coherent selves are best facilitated by Companions that are better informed, cooler-headed reflections of ourselves, or by distinct agents interested in us in the same benign but instrumental way as our therapist or financial advisor. Either way, they will need to be distinct and independent enough for us to treat their advice, information, and representation of our rules as real constraints on our action, but sufficiently closely aligned and sensitive to our goals that we do not feel them as an imposition. Companions will inevitably structure our choices, but we will choose them to do precisely that.

[9] Expect to see people shopping around for Companions with ideologically appealing frameworks.

[10] See e.g. P. Streich and J. S. Levy (2007) Time horizons, discounting, and inter-temporal choice, *Journal of Conflict Resolution*, 51 pp.199–226.