”Evolved altruism, ethical complexity, anthropomorphic trust: three factors misleading estimates of the safety of artificial general intelligence”.

Presented at the 7th European Conference on Computing and Philosophy (ECAP 2009) on July 4, 2009.

Kaj Sotala
Cognitive Science Unit, Department of Psychology, University of Helsinki

1000 word abstract

A true artificial general intelligence (AGI) is not a mere tool, but an autonomously acting agent. In this light, careful attention needs to be given to its goals and behavior. In the worst case, AGI may even become an ”existential risk”, threatening the survival of all of humanity [1]. I argue that there are three powerful factors which may bias humans to consider artificial intelligences more safe than they truly are.

First, evolutionary considerations may lead people astray. There exists a rich literature on game-theoretic reasons for why certain altruistic or cooperative behavior is likely to evolve. For instance, McNamara et al. [2] describe a simulation where individuals are represented by two traits, their willingness to invest effort in cooperative tasks and the amount of cooperative effort they demand from others. The authors show that in populations of non-identical agents, the population will evolve towards more cooperative individuals. Less abstractly, cases of reciprocal altruism, their effects on the community and their likelihood to evolve, have been extensively discussed in evolutionary biology [3].

One of the ways that fairness may evolve is by part of the population coming to express empathy, defined here as the property of making offers the agent itself would be willing to accept [4]. On these grounds, it may seem like an easy assumption that the AGIs prevailing in a competitive environment are those that can be deemed ”friendly”, not direct risks to humanity.

However, there is no reason why altruistically behaving AGI would need to be *inherently* altruistic. Having altruism as a ”hardwired” personality trait causes conventionally evolved agents to have a much higher likelihood of actually being altruistic than if they had to figure out the benefits of altruism themselves, on the basis of rational reasoning. Before humans, this kind of higher reasoning wasn't even possible, so all altruism was hard-wired. However, AGI *are* capable of advanced logical reasoning, as well as self-modification. They need not develop *sympathy* on top of empathy, as a mechanism making them help those that they perceive to be in need. They may simply choose to cooperate when it fits them, and discard such behaviors whenever it doesn't benefit them.

Second, a blindness to the arbitrariness and complexity of evolved ethical factors creates a false tendency to believe a system has adopted our goals. Our behavior and cognition is filled with evolutionary ”hacks”, created for very specific purposes based on specific regularities in our environment. As listed by [5]: ”the geometry of parallax gives vision a depth cue; an infant nursed by your mother is your genetic sibling”. As seen above, ethical goals are also similarly arbitrary. Moral philosophers run into problems that seem impossible to solve without highly counterintuitive consequences [6]: our ethics weren't built for universal consistency.

Simply giving the subsystems in their AI's sufficiently suggestive names has occasionally misled researchers, leading them to attribute to the systems properties they didn't actually have [7]. When we see something having one property that's commonly associated with specific other properties,

we tend to automatically assume it also has the other properties. Being blind to the deep structure of our ethical systems, we are in danger of thinking AGI to have morals more like ours than they actually do. A closely related risk is the halo effect, where a person's appearance in one field biases our global evaluations of them in the same direction (positive or negative). Subjects are not always aware of the halo effect, and may even believe that the direction of influence is opposite to the true direction [8].

Even if an AGI would have adopted our goals in one environment, it may not reason similarly about them in others. This has direct paralells in human behavior and our evolved goals. In order to produce more offspring, we evolved to consider sex enjoyable, but these days we employ contraception in order to disconnect the act of sex from actual conception.

Finally, it is hard to combat anthropomorphic tendencies when trying to evaluate AGI reliability. Trust in humans is at least partially mediated by oxytocin - higher levels of oxytocin lead to more trusting behavior [9]. Trusting somebody and then not being betrayed by the trustee increases oxytocin levels [10], and the hormone has been linked to pair bonding. Testing an AGI for reliability and then having one's trust repaid seems likely to trigger the same mechanism. Thus people may believe that an AGI that has cooperated with them for a long time has ”earned their trust”, and feel protective whenever the AGI's friendliness is questioned. However, unlike humans, AGIs need not have psychological mechanisms that would make it hard for them to systematically deceive us. Even if we do begin to trust them more and more, they don't necessarily need to be any more hesitant to betray us later on than they were in the beginning. Their reliability cannot be ensured on the basis of behavior alone, without inspecting their internal workings. But researchers who have grown to trust the AGIs might be reluctant to do so.

Furthermore, social psychologists have documented the phenomenon of correspondence bias. People very readily infer other people's behavior to be caused by their internal disposition, even when situational forces would explain the behavior just as well [11]. For example, people listening to either a pro-choice or pro-life speech prepared by someone else will readily attribute that person with pro-choice or pro-life attitudes himself. This happens even when the listeners are specifically told that a position had been *randomly* chosen for the speaker, and the person had been instructed to prepare a speech defending that randomly chosen position! [12] This kind of effect will make an objective evaluation of an AGI's true trustworthiness even harder.

The three factors surveyed here are strongly interrelated, and though they may seem obvious in retrospect, one easily ignores them if not made explicitly aware of them. Regardless, if we are to design true AGI, taking them into account is paramount.

References

[1] Yudkowsky, E. (2006) Artificial Intelligence as a Positive and Negative Factor in Global Risk. *Global Catastrophic Risks*, Nick Bostrom & Milan Cirkovic (eds.) Oxford University Press, 2007.
[2] McNamara, J.M. & Barta, Z. & Fromhage, L. & Houston, A.I. (2008) The coevolution of choosiness and cooperation. *Nature* 451, 189-192.
[3] Trivers, RL. (1971) The evolution of reciprocal altruism. *The Quarterly Review of Biology,* vol. 46, no. 1, 35-57.
[4] Page, K.M. & Nowak, M.A. (2002) Empathy Leads to Fairness. *Bulletin of Mathematical Biology* 64, 1101-1116.
[5] Tooby, J. & Cosmides, L. (2009) The Great Pivot: Artificial Intelligences, Native Intelligences, and the Bridge Between. *The Edge Annual Question 2009.*
http://www.edge.org/q2009/q09_11.html#tooby
[6] Ryberg, J. & Tännsjö, T. & Arrhenius, G. (2008) "The Repugnant Conclusion",

The Stanford Encyclopedia of Philosophy (Fall 2008 Edition), Edward N. Zalta (ed.). http://plato.stanford.edu/archives/fall2008/entries/repugnant-conclusion/
[7] McDermott, D. (1976) Artificial Intelligence Meets Natural Stupidity. *SIGART Newsletter,* no. 57.
[8] Nisbett, R.E. & DeCamp Wilson, T. (1977) The Halo Effect: Evidence for Unconscious Alteration of Judgments. *Journal of Personality and Social Psychology.* Vol 35, no. 4, 250-256.
[9] Kosfeld, M. & Heinrichs, M. & Zak, P.J. & Fischbacher, U. & Fehr, E. (2005) Oxytocin increases trust in humans. *Nature* 435, 673-676.
[10] Zak, P.J. (2008) The Neurobiology of Trust. *Scientific American*, June 2008.
[11] Gilbert, D.T. & Malone, P.S. (1995) The Correspondence Bias. *Psychological Bulletin*, vol. 117, no. 1, 21-38.
[12] Gilbert, D.T. & Pelham, B.W. & Krull, D.S. (1988) On Cognitive Busyness: When Person Perceivers Meet Persons Perceived. *Journal of Personality and Social Psychology.* Vol. 54, no. 5, 733-740.