Michael Vassar

Proposed paper for Transvision 2004

Memes and Rational Decisions

Although Transhumanism is not a religion, advocating as it does the critical analysis of any position; it does have certain characteristics which may lead to its identification as such by concerned skeptics. I am sure that everyone here has had to deal with this difficulty, and as it is a cause of perplexity for me I would appreciate it if anyone who has some suggested guidelines for interacting honestly with non-transhumanists share them at the end of my presentation. It seems likely to me that each of our minds contains either meme complexes or complex functional adaptations which have evolved to identify "religious" thoughts and to neutralize their impact on our behavior. Most brains respond to these memes by simply rejecting them. Others however, instead neutralize such memes simply by not acting according to the conclusions that should be drawn from such memes. In almost any human environment prior to the 20th century this religious hypocrisy would be a vital cognitive trait for every selectively fit human. People who took in religious ideas and took them too seriously would end up sacrificing their lives overly casually at best, and at worst would become celibate priests. Unfortunately, these memes are no more discriminating than the family members and friends who tend to become concerned for our sanity in response to their activity. Since we are generally infested with the same set of memes, we genuinely are liable to insanity, though not of the suspected sort.

A man who is shot by surprise is not particularly culpable for his failure to dodge or otherwise protect himself, though perhaps he should have signed up with Alcor. A hunter gatherer who confronts an aggressive European with a rifle for the first time can also receive sympathy when he is slain by the magic wand that he never expected to actually work. By contrast, a modern Archimedes who ignores a Roman soldier's request that he cease from his geometric scribbling is truly a mad man. Most of people of the world, unaware of molecular nanotechnology and of the potential power of recursively self-improving AI are in a position roughly analogous to that of the first man. The

business and political figures that dismiss eternal life and global destruction alike as plausible scenarios are in the position of the second man. By contrast, it is we Transhumanists who are for the most part playing the part of Archimedes. With death, mediated by technologies we understand full well staring us in the face; we continue our pleasant intellectual games. At best a few percent of us have adopted the demeanor of an earlier Archimedes and transferred our attention from our choice activities to other, still interesting endeavors which happen to be vital to our own survival. The rest are presumably acting as puppets of the memes which react to the prospect of immortality by isolating the associated meme-complex and suppressing its effects on actual activity.

OK, so most of us don't seem to be behaving in an optimal manner, what manner would be optimal? This ISN'T a religion, remember? I can't tell you that. At best I can suggest an outline of the sort of behavior that seems to me least likely to lead to this region of space becoming the center of a sphere of tiny smiley faces expanding at the speed of light.

The first thing that I can suggest is that you take rationality seriously. Recognize how far you have to go. Trust me; the fact that you can't rationally trust me without evidence is itself a demonstration that at least one of us isn't even a reasonable approximation of rational, as demonstrated by Robin Hanson and Tyler Emerson of George Mason University in their paper on rational truth-seekers. The fact is that humans don't appear capable of approaching perfect rationality to anything like the degree to which most of you probably believe you have approached it. Nobel Laureate Daniel Kahneman and Amos Tversky provided a particularly valuable set of insights into this fact with their classic book "Judgment Under Uncertainty: Heuristics and Biases" and in subsequent works. As a trivial example of the uncertainty that humans typically exhibit, try these tests. (Offer some tests from Judgment Under Uncertainty)

I hope that I have made my point. Now let me point out some of the typical errors of Transhumanists who have decided to act decisively to protect the world they care about from existential risks. After deciding to rationally defer most of the fun things that they would like to do for a few decades until the world is relatively safe, it is completely typical to either begin some quixotic quest to transform human behavior on a grand scale

over the course of the next couple decades or to go raving blithering Cthulu worshipping mad and try to build an artificial intelligence. I will now try to discourage such activities.

One of the first rules of rationality is not to irrationally demand that others be rational. Demanding that someone make a difficult mental transformation has never once lead to them making said transformation. People have a strong evolved desire to make other people accept their assertions and opinions. Before you let the thought cross your mind that a person is not trying to be rational, I would suggest that you consider the following. If you and your audience were both trying to be rational, you would be mutually convinced of EVERY position that the members of your audience had on EVERY subject and vice versa. If this does not seem like a plausible outcome then one of you is not trying to be rational, and it is silly to expect a rational outcome from your discussion. By all means, if a particular person is in a position to be helpful try to blunder past the fact of your probably mutual unwillingness to be rational; in a particular instance it is entirely possible that ordinary discussion will lead to the correct conclusion, though it will take hundreds of times longer than it would if the participants were able to abandon the desire to win an argument as a motivation separate from the desire to reach the correct conclusion. On the other hand, when dealing with a group of people, or with an abstract class of people, Don't Even Try to influence them with what you believe to be a well reasoned argument. This has been scientifically shown not to work, and if you are going to try to simply will your wishes into being you may as well debate the nearest million carbon atoms into forming an assembler and be done with it, or perhaps convince your own brain to become transhumanly intelligent. Hey, its your brain, if you can't convince it to do something contrary to its nature that it wants to do is it likely that you can convince the brains of many other people to do something contrary to their natures that they don't want to do just by generating a particular set of vocalizations?

My recommendation that you not make an AI is slightly more urgent. Attempting to transform the behavior of a substantial group of people via a reasoned argument is a silly and superstitious act, but it is still basically a harmless one. On the other hand, attempts by ordinary physicist Nobel Laureate quality geniuses to program AI systems are not only astronomically unlikely to succeed, but in the shockingly unlikely event that they do succeed they are almost equally likely to leave nothing of value in this part of the

universe. If you think you can do this safely despite my warning, here are a few things to consider.

a) A large fraction of the greatest computer scientists and other information scientists in history have done work on AI, but so far none of them have either begun to converge on even the outlines of a theory or succeeded in matching the behavioral complexity of an insect, despite the fantastic military applications of even dragonfly equivalent autonomous weapons
b) Top philosophers, pivotal minds in the history of human thought, have consistently failed to converge on ethical policy.
c) Isaac Asimov, history's most prolific writer and Mensa's honorary president, attempted to formulate a more modest set of ethical precepts for robots and instead produced the blatantly suicidal three laws (if you don't see why the three laws wouldn't work I refer you to the singularity institute for artificial intelligence's campaign against the three laws)
d) Science Fiction authors as a class, a relatively bright crowd by human standards, have subsequently thrown more time into considering the question of machine ethics than they have any other philosophical issue other than time travel, yet have failed to develop anything more convincing than the three laws.
e) AI ethics cannot be arrived at either through dialectic (critical speculation) or through the scientific method. The first method fails to distinguish between an idea that will actually work and the first idea that you and your friends couldn't rapidly see holes in, influenced as you were by your specific desire for a cool sounding idea to be correct and your more general desire to actually realize your AI concept, saving the world and freeing you to devote your life to whatever you wish. The second method is crippled by the impossibility of testing a transhumanly intelligent AI (due to the fact that it could by definition trick you into thinking it had passed the test) and by the irrelevance of testing an ethical system on an AI without transhuman intelligence. Ask yourself, how constrained would your actions be if you were forced to obey the code of Hamurrabi but you had no other ethical impulses at all. Now keep in mind that Hamurrabi was actually FAR more like you than an AI will be. He shared almost all of your

genes, you're very high by human standards intellect, and the empathy that comes from an almost identical brain architecture, but his attempt at a set of rules for humans was a first try, just as your attempt at a set of rules for AIs would be.

f) Actually, if you are thinking in terms of a set of rules AT ALL this implies that you are failing to appreciate both a programmers control over an AI's cognition and an AI's alien nature. If you are thinking in terms of something more sophisticated, and bear in mind that apparently only one person ever has thought in terms of something more sophisticated so far, bear in mind that the first such "more sophisticated" theory was discovered upon careful analysis to itself be inadequate, as was the second.

If you can't make people change, and you can't make an AI, what can you do to avoid being killed? As I said, I don't know. It's a good bet that money would help, as well as an unequivocal decision to make singularity strategy the focus of your life rather than a hobby. A good knowledge of cognitive psychology and of how people fail to be rational may enable you to better figure out what to do with your money, and may enable you to better co-operate your efforts with other serious and rational Transhumanists without making serious mistakes. If you are willing to try, please let's keep in touch. Seriously, even if you discount your future at a very high rate, I think that you will find that living rationally and trying to save the world is much more fun and satisfying than the majority of stuff that even very smart people spend their time doing. It really really beats pretending to do the same, yet even such pretending is or once was a very popular activity among top notch Transhumanists.

Aiming at true rationality will be very difficult in the short run, a period of time which humans who expect to live for less than a century are prone to consider the long run. It entails absolutely no social support from non-Transhumanists, and precious little from Transhumanists, most of whom will probably resent the implicit claim that they should be more rational. If you haven't already, it will also require you to put your every-day life in order and acquire the ability to interact positively with people or a less speculative character. You will get no VC or Angel funding, terribly limited grant money, and in general no acknowledgement of any expertise you acquire. On the other

hand, if you already have some worth-while social relationships, you will be shocked by just how much these relationships improve when you dedicate yourself to shaping them rationally. The potential of mutual kindness, when even one partner really decides not to do anything to undermine it, shines absolutely beyond the dreams of self-help authors.

If you have not personally acquired a well-paying job, in the short term I recommend taking the actuarial tests. Actuarial positions, while somewhat boring, do provide practice in rationally analyzing data of a complexity that denies intuitive analysis or analytical automatonism. They also pay well, require no credentials other than tests in what should be mandatory material for anyone aiming at rationality, and have top job security in jobs that are easy to find and only require 40 hours per week of work. If you are competent with money, a few years in such a job should give you enough wealth to retire to some area with a low cost of living and analyze important questions. A few years more should provide the capital to fund your own research. If you are smart enough to build an AI's morality it should be a breeze to burn through the 8 exams in a year, earn a six figure income, and get returns on investment far better than Buffet does. On the other hand, doing that doesn't begin to suggest that you are smart enough to build an AI's morality. I'm not convinced that anything does.

Fortunately ordinary geniuses with practiced rationality can contribute a great deal to the task of saving the world. Even more fortunately, so long as they are rational they can co-operate very effectively even if they don't share an ethical system. Eternity is an intrinsically shared prize. On this task more than any other the actual behavioral difference between an egoist, altruist, or even a Kantian should fade to nothing in terms of its impact on actual behavior. The hard part is actually being rational, which requires that you postpone the fun but currently irrelevant arguments until the pressing problem is solved, even perhaps with the full knowledge that you are actually probably giving them up entirely, as they may be about as interesting as watching moss grow post singularity. Delaying gratification in this manner is not a unique difficulty faced by transhumanists. Anyone pursuing a long-term goal, such as a medical student or PhD candidate, does the same. The special difficulty that you will have to overcome is the difficulty of staying on track in the absence of social support or of appreciation of the problem, and the difficulty of overcoming your minds anti-religion defenses, which will be screaming at you to cut

out the fantasy and go live a normal life, with the normal empty set of beliefs about the future and its potential.

Another important difficulty to overcome is the desire for glory. It isn't important that the ideas that save the world be your ideas. What matters is that they be the right ideas. In ordinary life, the satisfaction that a person gains from winning an argument may usually be adequate compensation for walking away without having learned what they should have learned from the other side, but this is not the case when you elegantly prove to your opponent and yourself that the pie you are eating is not poisoned. Another glory related concern is that of allowing science fiction to shape your expectations of the actual future. Yes it may be fun and exciting to speculate on government conspiracies to suppress nanotech, but even if you are right conspiracy theories don't have enough predictive power to test or to guide your actions. If you are wrong, you may well end up clinically paranoid. Conspiracy thrillers are pleasant silly fun. Go ahead and read them if you lack the ability to take the future seriously, but don't end up in an imaginary one, that is NOT fun.

Likewise, don't trust science fiction when it implies that you have decades or centuries left before the singularity. You might, but you don't know that, it all depends on who actually goes out and makes it happen. Above all, don't trust its depictions of the sequence in which technologies will develop or of the actual consequences of technologies that enhance intelligence. These are just some author's guesses. Worse still, they aren't even the author's best guesses, they are the result of a lop-sided compromise between the author's best guess and the set of technologies that best fit the story the author wants to tell. So you want to see mars colonized before singularity. That's common in science fiction, right? So it must be reasonably likely. Sorry, but that is not how a rational person estimates what is likely. Heuristics and Biases will introduce you to the "representativeness heuristic", roughly speaking the degree to which a scenario fits a preconceived mental archetype. People who haven't actively optimized their rationality typically use representativeness as their estimate of probability because we are designed to do so automatically so we find it easy to do so. In the real world this doesn't work well. Pay attention to logical relationships instead.

Since I am attempting to approximate a rational person, I don't expect e-mails from any of you to show up in my in-box in a month or two requesting my cooperation on some sensible and realistic project for minimizing existential risk. I don't expect that, but I place a low certainty value on most of my expectations, especially regarding the actions of outlier humans. I may be wrong. Please prove me wrong. The opportunity to find that I am mistaken in my estimates of the probability of finding serious transhumanists is what motivated me to come all the way across the continent. I'm betting we all die in a flash due to the abuse of these technologies. Please help me to be wrong.