CHAPTER 12
EXTROPIAN ELITISM AND HUMANIST POSTHUMANISM

This chapter, more than any of the others in the book, has evolved significantly from its initial form. It began, like many of the others, as an article for the German newspaper *Frankfurter Allgemaine Zeitung*. Of all the articles Ben wrote for FAZ during the 1999-2001 time period, this was the one that excited them the most. It's not hard to see why: rather than explaining difficult technical content, it focused on social and moral issues. Admittedly, like a lot of successful journalism, it also had a bit of an over-sensationalistic tone – including a contentious Holocaust metaphor -- which one of the editors wanted to remove, but he was overruled by the publisher, Frank Schirrmacher, who felt pushing the boundaries to be one of the key purposes purpose of futurist writing in the first place.

The original article was a kind of critique of the Extropians, a specific group of futurist thinkers and activists. As a consequence of writing and publishing the article, Ben came into contact with a greater variety of Extropians than he'd known previously, and came to somewhat modify his views of the group (Stephan already had bumped into number of them from his work with Sasha Chislenko on hypereconomics and other issues). This chapter has ended up as a kind of chronicle of its own evolution, and of the evolution of our thinking about this group of futurists and the strengths and weaknesses of their conceptual approach to the issues of our mutual interest.

Ben's original FAZ article on Extropians started out as follows:

Nietzsche, my favorite philosopher, gave his book "Twilight of the Idols" the subtitle "How to philosophize with a hammer." It was the moral codes and habitual thought patterns of his culture that he was smashing. In a similar vein, the creed of the Extropians, a group of transhumanist futurists centered in California, might be labeled "How to technologize with a hammer." This group of computer geeks and general high-tech freaks wants to push ahead with every kind of technology as fast as possible – the Internet, body modification, human-computer synthesis, nanotechnology, genetic modification, cryogenics, you name it. Along the way they want to get rid of governments, moral strictures, and eventually humanity itself, remaking the world as a hypereconomic virtual reality system in which money and technology control everything. Their utopian vision is sketchy but fabulous: a kind of Neuromancer-ish, Social-Darwinist Silicon-Valley-to-the-n'th-degree of the collective soul.[1]

I sympathize with their techno-futurism and their lust for freedom. But their brand of ethics scares me a little.

Intuitively conceived as the opposite of entropy, Extropy is a philosophical rather than a scientific term. The Extropians website (www.extropy.org), the online Bible of the movement, defines Extropy as "A metaphor referring to attitudes and values shared by those who want to overcome human limits through technology. These values … include a desire to direct oneself in pursuing perpetual progress and self-transformation with an attitude of practical optimism implemented using rational thinking and intelligent technology in an open society."

---

1 As a side note: When Ben wrote his articles for the FAZ, they were submitted in English and translated into German. Neither of us knows German, and we never checked to see what the German translation of the phrase "Neuromancer-ish, Social-Darwinist Silicon-Valley-to-the-n'th-degreee of the collective soul" was.

"Transhumanism," as a general term, refers to philosophy that doesn't view human life as the ultimate endpoint of the evolution of intelligence. Extropianism is a particular form of transhumanism, concerned with the quest for "the continuation and acceleration of the evolution of intelligent life beyond its currently human form and limits by means of science and technology, guided by life-promoting principles and values, while avoiding religion and dogma." Working toward the obsolescence of the human race through AI and robots is one part of this; another aspect is the transfer of human personalities into "more durable, modifiable, and faster, and more powerful bodies and thinking hardware," using technologies such as genetic engineering, neural-computer integration and nanotechnology.

Along with this technological vision comes a political vision. Extropians, according to extropy.org, are distinguished by a host of sociopolitical principles, such as: "Supporting social orders that foster freedom of speech, freedom of action, and experimentation. Opposing authoritarian social control and favoring the rule of law and decentralization of power. Preferring bargaining over battling, and exchange over compulsion. Openness to improvement rather than a static utopia. … Seeking independent thinking, individual freedom, personal responsibility, self-direction, self-esteem, and respect for others." It is explicitly stated in Extropian doctrine that there cannot be socialist Extropians, although the various shades of democratic socialism are not explored in detail. In point of fact, the vast majority of Extropians are radical libertarians, advocating the total or near-total abolition of the government. This is really what is unique about the Extropian movement: the fusion of radical technological optimism with libertarian political philosophy. With only slight loss of meaning, one might call it libertarian transhumanism.

The characterization of Extropian philosophy that Ben gave in those paragraphs was based on conversations with a number of individuals identifying themselves as Extropians, both in person and via e-mail. Some of these individuals and conversations will be discussed later on in this chapter.

Conversations with other Extropians during the last couple years, however -- including Natasha Vita-Moore, one of the original Extropians -- have made us realize that this impression in the original article of the Extropian group was somewhat flawed and incomplete.

Everything observed in that article was true, but only of a certain subset of the Extropian community. But the Extropian community has a lot more diversity than the article which spawned this chapter gave it credit for. The statement that "the vast majority of Extropians are radical libertarians" was an overstatement. Many Extropians are radical libertarians, and very few are socialists, but all in all there is a much greater variety of political views in the community than appears at first glance.

This chapter represents a variant of that FAZ article, but with a less sensationalistic and hopefully more accurate flavor. The basic themes are the same, but we're pleased to be able to present them now as pertaining to a subset of the Extropian community rather than the Extropian community as a whole, and to give a more balanced overall perspective.

Even the new and mellower version of our take on Extropy and related ideas is unlikely to please everyone in the Extropian community – but we're calling it as we see it. As already noted above, we wish to emphasize the distinction between the "official Extropian line", as laid out on the Extropy website, and the actual belief systems that tend to be held by the majority of individuals associating themselves with Extropianism. Our concern here is mainly with these actual belief systems. We're not writing about Extropianism as a formal set of beliefs or as an official organization, but rather about the *cluster of individuals and ideas* that has aggregated around the Extropy concept over the last couple decades.

For example, libertarian politics is not part of the official Extropian philosophy; but it is a mighty common theme on the Extropy e-mail list and at Extropy conferences. The complaint that Extropianism isn't intrinsically connected to libertarianism is reminiscent of an argument Ben once had with a Sufi, who claimed that Sufism isn't a religion. He was right, formally speaking Sufism is not a religion – it's a "wisdom tradition" of Arabic origin, associated with Islam. And yet 99.9%, perhaps 100%, of Sufis are Muslims and are religious by the definitions of the rest of the world.

–And there's no question: Some Extropians carry their anti-socialist libertarianism to a remarkable, ultra-radical extreme. For instance, visionary roboticist Hans Moravec, a hero to many Extropians, had a somewhat disturbing exchange with writer Mark Dery in 1993 (see Dery, 1997). Dery asked Moravec about the socioeconomic implications of the robotic technology he envisioned. Moravec replied that "the socioeconomic implications are… largely irrelevant. It doesn't matter what people do, because they're going to be left behind like the second stage of a rocket. Unhappy lives, horrible deaths, and failed projects have been part of the history of life on Earth ever since there was life; what really matters in the long run is what's left over." Does it matter to us now, he asks, that dinosaurs are extinct[2]? Similarly, the fate of humans will be uninteresting to the superintelligent robots of the future. Humans will be viewed as a failed experiment – and we can already see that some humans, and some human cultures, are worse failures than others.

Dery couldn't quite swallow this. "I wouldn't create a homology between failed reptilian strains and those on the lowermost rungs of the socioeconomic ladder."

Moravec's reply: "But I would."

Put this way, Extropianism starts to seem like a dangerous, profoundly ethically deficient philosophy – one which equates relative success in a particular system at a particular point in time with absolute worth. Thus the outrage of the

2 An obvious flaw in his argment by analogy is that, for a substantial number of people, the answer is "yes, it does matter to us". The is related to the fact that the vast preponderance of humans are capable of empathy, sympathy, and compassion. But we'll get there.

FAZ article that served as the seed of this chapter. But one must remember (which Ben didn't when he wrote the original article, overheated with moral outrage) that Moravec is just one voice among many – and a decent percentage of Extropians would be just as annoyed by Moravec's ethics as we are.

Although Moravec is often intentionally confrontational and alienating, overall, the Extropian perspective isn't *that* far out on the fringe these days. Luminaries associated with it in one way or another include Marvin Minsky the AI guru, Eric Drexler the nanotechnologist, Kevin Kelly of Wired Magazine, and futurist writer Ray Kurzweil. At this point, Extropianism does not rank as one of our more prominent cultural movements. But it is active, vibrant, and growing. The Extropy magazine had several thousand subscribers before it moved on the Web in 1997; and the Extropy e-mail discussion list is a hugely active one (though greatly uneven in quality). A vast amount of online literature exists, related to various aspects of Extropian thinking, linked to the http://www.extropy.org site. This is definitely one of the more important online communities. Whatever its strengths and weaknesses, it's worth paying some attention to. Discussion of Extropian ideas brings up all sorts of interesting topics, which are highly pertinent to the future of technology, life and intelligence.

MAX MORE: THE ORIGINAL EXTROPIAN

The man who got this all started was Max More, a philosophy Ph.D. with a knack for rational argumentation and an impressive, convincing personal style. In 1995, Jim McClellan interviewed More for the UK newspaper Observer and noted, "The funny thing about Max is that while his ideas are wild, he argues them so calmly and rationally you find yourself being drawn in."

More started his career studying philosophy, politics and economics at St. Anne's College at Oxford University, in the mid-1980's. At that point his main focus was on economics, from the libertarian perspective. While doing his degree, Max became strongly interested in life extension, and he was the first person in Europe to sign up for cryonic suspension with the US firm Alcor. In 1995, when he received his philosophy degree from the University of Southern California for

research on mind, ethics and personal identity, he was already deep into organizing the Extropian movement, bringing his political and technological interests together. Technology, he felt, was ready to push mind into new spaces altogether, such as virtual realities where the notion of "I" as currently conceived had no meaning. Governments were holding us back, preventing or slowing research in crucial areas.

The first edition of Extropy magazine came out in August/September 1988 with just 50 copies, co-edited by Max More and his friend T.O. Morrow. It was a wild mix of sci-fi-meets-reality thinking (much like this book!) -- life extension, machine intelligence, and space exploration, to intelligence augmentation, uploading, enhanced reality, alternative social systems, futurist philosophy, and so on. The magazine seeded the social network that led to the e-mail list (1991), the first Extropy conference (1994) and the website (1996), which soon (1997) obsoleted and incorporated the paper magazine.

In terms of philosophical precedents it's not too inaccurate to call More's credo a mix of Ayn Rand-ian anti-statist individualism with Nietzschean transmoralism, held together by a focus on future technologies. In Extropy #10 he explicitly equates the "optimal Extropian" with "Nietzsche's Uebermensch." But he cautions, in another essay ("Technological Transformation: Expanding Personal Extropy"), that "the Ubermensch is not the blond beast and plunderer." Rather, the Extropian Ubermensch "will exude benevolence, emanating its excess of health and self-confidence." That's reassuring, yet hard to reconcile with Moravec's Olympian detachment regarding the destruction of the human race. This contradiction, we believe, is both Extropianism's core weakness and a primary source of its energy.

In spite of More's forceful argumentative style, Extropianism is certainly not an orthodoxy. Within the general "party line" of Extropianism there's room for a lot of variety. This is one of the movement's strengths, and surely a necessary aspect of any organization involving so many overly-clever, individualistic, oddball revolutionaries. Moravec and More don't agree with each

other entirely, and don't necessarily agree with all their own past opinions. Consensus isn't critical; progress is the thing.

SASHA CHISLENKO

We've had intellectual exchanges with plenty of Extropians, including Marvin Minsky, Max More, Max's wife Natasha Vita-More (a highly creative thinker), Eliezer Yudkowsky (whom we'll discuss below), and too many others to name. But the Extropian we've known best on a personal level was Sasha Chislenko – a visionary cybertheorist and outstanding applied computer scientist. Sasha's work, thought and life exemplify the brilliance and power, and weakness and danger, of the Extropian perspective in an extremely vivid way.

As with many Russian emigrants to the US, Sasha's libertarianism was borne of years of oppression under the Soviet Socialist regime. Having seen first hand how much trouble an authoritarian government can cause, he was convinced that government was intrinsically predominantly evil. After he left the Soviet Union, Sasha was a man without a country, lacking a Russian passport due to his illegal escape from Russia and lacking an American passport because of his status as a political refugee. Once Sasha and Ben were invited to give a lecture together at Los Alamos National Labs in New Mexico, but were informed that since he lacked a passport he couldn't get the security clearance required to enter the lab grounds. They ended up turning down the lecture invitation, both disgusted at the government's closed-mindedness.

Sasha was impatient for body modification technology to advance beyond the experimental stage – he was truly, personally excited to become a cyborg, to jack his brain into the Net, to replace his feeble body and brain with superior engineered components. Not that there was anything particularly wrong with his body and brain – in fact he was in fine shape, with the exception of some intermittent emotional problems – it just wasn't as good as the best synthetic model he could envision. As a creature of a world of abstraction he could easily envision a million ways to improve upon his natural state and was anxious to do so. He was a strong advocate of various "smart drugs," some legal, some not,

which he felt gave him a superhuman clarity of thought. He was outraged that any government would consider it had the right to regulate the chemicals he chose to put into his body to enhance his intelligence.

His own technical work focused on "active collaborative filtering," technology that allows people to rate and review things they see on the Net and then recommends things to people based on their past ratings and the ratings of other similar people (Chislenko, 1996)[3] which, in collaboration with Stephan and others, he evolved into the idea of Hypereconomics (Chislenko, et. al., 1998)[4]. Popular websites like amazon.com and bn.com have primitive collaborative filtering systems embedded in them – when you log on to buy a book, they give you a list of books you might be interested in. Sometimes these systems work, sometimes they don't. Recently Ben logged onto Amazon to buy a "Bananas in Pyjamas" movie for his young daughter, and their recommendation system suggested that he might also be interested in the movie "Texas Chainsaw Massacre II." How it came up with that recommendation is anyone's guess: Perhaps the only previous person to by "Bananas in Pyjamas" had also bought the Texas Chainsaw Massacre film. The recommendation systems that Sasha designed were far more sophisticated than this one, probably the most advanced in the world. He led a team implementing some of his designs at Firefly, a company later acquired by Microsoft.

Compared to body modification, cranial jacks and superhuman artificial intelligence, active collaborative filtering might seem a somewhat unexciting path to the hypertechnological future, but to Sasha, it was a tremendously thrilling thing – a way for humans to come together and enhance one another's mental effectiveness, passing along what they'd learned to one another in the form of ratings, reviews and recommendations. Recommendation and filtering technology was a kind of collective smart drug for the net-surfing human race. Through recommendations and voting, Sasha believed that global intelligence could be increased, freedoms expanded, and social and environmental stability ensured

[3] For this and other of Sasha's writings, see his website, at http://www.lucifer.com/~sasha/home.html

with a hypereconomic system which allowed the aggregated values of individuals in the system to trump the centrally-imposed valuations of traditional economics, even market econcomics. His goal with first recommendation systems, and ultimately hypereconomics, was to eliminate all forms of central planning in economics entirely and create a system in which the values of communities were directly reflected in how resources were allocated throughout the entirety of the economic system.

The recommendation system aspect of Sasha's vision in this area has become somewhat mainstream by this point. An example is the Website epinions.com, which pays users to give their reviews of consumer products and other things. The higher that others rate your reviews, the more you get paid. Sasha had nothing to do with his site but it epitomized is ideal. He strongly felt that, as the economy transformed into a cyber-powered hypereconomy, intellectual contributions like his own would finally get the economic respect they'd always deserved. People would be paid for writing scientific papers to the extent that other scientists appreciated the papers. The greater good would be achieved, not by the edicts of an authoritarian government, but by the self-organizing effects of people rating each others productions, and paying each other for their ratings and opinions. He coined the word "hypereconomics" to refer to the complex dynamics of an economy in which artificial agents dispense perpetually renegotiated, often small payments for everything, and in which complex financial instruments emerge even from simple everyday transactions – AI agents paying other agents for advice about where to get advice; your shopping agent buying you not just lettuce but futures and options on lettuce, and maybe even futures and options on advice from other agents.

But there was a painful contradiction lurking here, not far beneath the surface. This personal contradiction, we believe, cuts close to the heart of Extropian philosophy – at least, in the form that it takes in the mind of many Extropians. The Libertarian strain in Sasha's thinking was highly pronounced: On different occasions he told each of us, tongue only halfway in cheek, that he thought air should be metered out for a price, and that those who didn't have the

money to pay for their air should be left to suffocate! We later learned this was a variation on a standard libertarian argument, sometimes repeated by Max More, to the effect that the reason the air was polluted was that nobody owned it[4] – ergo, air, like everything else, should be private property (although like any other extreme privatist, Extropians of this ilk are unable to explain how they'd propose to undo the unfair economic playing field the wealthy would have during privatization of shared resources based on their long history of accepting corporate welfare and government support of coercive, monopolist practices which contradict the Libertarian ideal). At the 2001 Extropy conference a speaker give an even more extreme variation: In the future, every molecule will be bar-coded so its owner can be identified. People or their descendants will have to pay for the oxygen they breathe, molecule by molecule. While this seems both absurd and offensive, our economic system is currently headed in basically that direction, as evidenced by government permission for patents and monopoly claims on genes, mathematical formulae, and a variety of other things considered "part of nature" and previously barred by patent tradtion.

Sasha equated wealth with fundamental value, and his vision of the cyberfuture was one of a complex hypereconomic network, a large mass of money buzzing around in small bits, inducing people and AI agents to interact in complex ways according to their various personal greed motives (although he didn't feel it was based on greed, but rather mutual benefit with hypercompetition buffered out by a détente created by a situation akin to the games in Robert Axelrod's *The Evolution of Cooperation* and the greatly lessened ability of any individual to game the system). But Sasha was by no means personally wealthy and this fact was highly disturbing to him. He often felt that he was being shafted, that the world owed him more financial compensation for his brilliant ideas, and that the companies he'd worked for had taken his ideas and made millions of dollars from them, of which he'd seen only a small percentage in the form of salary and stock options.

4 An idea Extropians have cribbed from extreme Libertarians in the pre-Cyber era.

Sasha worked for us for a while in 1999 and 2000; our company Webmind Inc. hired him away from Marvin Minsky's lab at MIT. While we enjoyed Sasha very much as an intellectual collaborator and a friend, he wasn't in any way an easy employee. We were excited to bring him into the Webmind team, but were relieved as well as sad when he quit in mid-2000, having been offered a position as CTO of a tech incubator in Boston. He contributed many interesting technical and conceptual ideas to our Webmind work – mostly to the Webmind Classification system product (which focused on automatically placing documents into categories), but to some extent to the AI R&D codebase as well. But his staunch individualism combined with his aggressively Socratic demeanor and preference for theory over praxis caused him to excel neither as a practical implementor nor as a manager, and so it was sometimes hard to fit him into the work process of a start-up company based on collaborative teamwork. He spent very little of his life in a university research context, but this would probably have been the most natural place for him – he had diverse skills and ambitions, but above all, he was a visionary deep thinker and conceptual guru par excellence.

Toward the end of 1999, he frequently told each of us how he had conquered many of the intellectual puzzles he had been struggling with for decades, and was now focusing on mastering his own mind and emotions. The gravity with which he declared this scared us a little. In one way or another, we each told him, probably too flippantly, that we found emotions were sometimes more fun if you left them unmastered. Sasha's response to such comments was to become extremely serious and philosophical, warning us that we all needed to master our own minds lest we leave ourselves open not only to control by others (his lingering Soviet paranoia) but also the destructive tendencies of our own weaknesses. While this sounds not unlike many self-help books, Sasha couched it in very critical, technical, and philosophically Randian contexts which made it clear he considered this task to be beyond mere self improvement. But he wasn't always so serious: he was also an avid disco dancer, occasionally observed leaving the Webmind office in the evening with a girl half his age on his arm,

heading for a dance club or a rave, where he would move to the beat in his peculiarly robotic yet wonderfully vivacious way.

When Sasha committed suicide in mid-2000, we wondered at first whether it had been an act of philosophical despair. Had there been a problem at his new company – were they unwilling to implement his latest designs for online collaborative filtering? Had he received one more devastating piece of evidence that the world just wasn't going to compensate him appropriately for his ideas, that the hypereconomic cyberfuture was far too slow in arriving? Had we, as not only collaborators but friends, somehow caused him to feel philosophically abandoned by not agreeing with him on various substantial points of how to design AGI and hypereconomics? As it turned out, his terrible action was more directly motivated by a complicated and painful romantic relationship – good old-fashioned, low-tech human passionate distress.

His 19 year old girlfriend had jilted him. The situation was complex, as such things often are, but the crux of it seems to be that he had wanted a more exclusive sort of relationship than she had. She later created some controversy by posting his final love letters to her on her public website, along with her wide-ranging, youthful musings on life, the universe and everything. Ben corresponded with her briefly and found her sweet, intelligent, creative, understandably upset, and more than a little confused. Fresh out of high school, she'd been overwhelmed by the mind and affections of this 40-year-old, sometimes-depressed genius. She'd known he was both depressed and jealous, but was as shocked as anyone else to learn that he'd hung himself in his apartment. She felt sure she had had a brief spiritual contact with him from beyond the grave.

In some important ways, Sasha was similar to Nietzsche, who as we've seen was one of the Extropians' philosophical godfathers. Both Sasha and Nietzsche were intellectual superstars who explicitly enounced one moral philosophy but lived another. Nietzsche preached toughness and hardness[5], but in his life he was a sweet person, respectful of the feelings of his mother and sister

5 Though not necessarily of the form that too many people have come to assume, usually through never having read any of his work.

(whose beliefs he despised). On the day he went mad, he was observed hugging a horse in the street, sympathetic that its master had whipped it. He preached the merits of the robust, healthy man of action and criticized intellectual ascetics, yet he himself was sickly, nearly celibate, and sat in his room thinking and writing day in and day out. Similarly, Sasha extolled the money theory of value, yet lived his own life seeking truth and beauty rather than cash, trying to transform the world for the better and distributing his ideas for free online – one one level assuming that a meritocracy was not only the intended order of society, but that it actually existed unimpeded in America and his rewards could come because of rather than in spite of such generosity of ideas, and on another level knowing better and feeling that by his example he could lead the world by example into a implementing its meritocratic ideals. He argued that air should be metered out only to those who could pay for it, yet was unfailingly kind and generous in real life, always willing to help young intellectuals along their way without asking for anything in return.

For what it's worth, it's impossible to avoid observing in this context that Sasha, the would-be-cyborg transhumanist, manifested a remarkable number of cyborg-like personality traits. His body movements were sometimes oddly robotic – in fact he looked most natural when dancing to techno music, with it computer-generated beats. It would be an unfair exaggeration to say that his voice had something of the manner of a speech synthesizer – but it did have a peculiar stiffness to it, that one might describe as wooden or metallic. Of course, this point should not be made too strongly: Sasha was an outgoing, friendly human being, easily hurt and in some circumstances quick to anger; he was by no means devoid of affect, and was generally a warm and giving friend. But when, about six months before his death, a group of us were coming up with silly e-mail nicknames for our co-workers (Sasha was among them at the time), the one we picked for Sasha was robotron@webmind.com. It was clear to everyone who knew him that he had difficulties dealing with the ambiguities and subtleties of human attitudes and relationships. He acknowledged this himself, and sometimes said it was something he was working on, while at other times he embraced it as

superior, admonishing us all to bring order and rigor to our social and emotional interactions. He was a poor politician, which is partly why he so often got himself into positions where his ideas weren't adequately appreciated by his co-workers or employers. Extropianism, a clear-cut, simple philosophy, seemed to provide him a welcome respite from the human complexities and contradictions that caused him so much grief in ordinary life. In the end, however, Extropianism failed him severely in that the Moravecian strains of dogmatic hypercriticism of all things human left little room for the errors and omissions that each individual makes as they try to improve themselves. Like any other system of dogma, the imperfect (i.e. everyone) are find in their daily life that they have bought into little more than an unattainable ideal, as such inflexible systems lead necessarily to collapse, rather than adaptation, in the face of any critical failure of their tenets to prove irrefutable.

Of course, not all Extropians have Sasha's personality characteristics, and not all impose the contradictions of a dogmatic personal philosophy coupled with an actually humanist lifestyle upon themselves. It would be a mistake to overgeneralize, to create a psychology of Extropians from this one example. Max More, for example, is extremely politically adept in his own way; and Max and his wife Natasha are both body-builders with much grace and naturalness in their physical motions, devoted to living well-rounded lives as well as to deep futurism and the life of the mind. Many Extropians have above average mastery of human relations, happy personal lives, and so forth. But everyone has their own philosophical quandaries they're looking for answers to and it's impossible not to hypothesize that the role Extropianism played for Sasha – providing crisp certainties to serve as welcome relief from the puzzling, stressful confusion of everyday life – tells us something about the role Extropianism plays for some other individuals as well.

For some of its adherents, Extropianism serves the role of providing a simple, optimistic world-view and a community of like-minded believers. Like most religions, and other religion-like belief systems such as Marxism or Neoconservativism, via its focus on a better future world Extropianism can

encourage avoidance of the difficult ambiguities of human reality. Through focus solely on the better future or a return to the idealized past, especially coupled with an assumption of inevitability, a dogmatic belief system helps people deal in one way or another with existential questions and impose certainty onto a system – life – which is inherently ambiguous. Extropianism is explicitly anti-religious, but it's not a new observation that rabid anti-religiosity can, for some people, serve almost as a religion itself. As Dostoevsky said, the atheist is one step away from the devout. Atheism and theism provide the mind with the same kind of rigid certainty. For some people, this kind of definitive cutting-through of the Gordian knots of messy human reality can be indispensable, providing the comfort level prerequisite to a healthy and productive state of mind. Unfortunately, an otherwise comforting belief system coupled with the human drive to share ideas and to be certain about things leads to the prostelytizing mindset, and the belief that anyone who can not be convinced to believe the same *existential particulars* as oneself is inferior and omitting them from society is not just acceptable, but mandatory.

And of course, for some other Extropians, the Extropian perspective does not serve this sort of role at all, but rather is a conceptual and practical philosophy fitting in naturally with an intellectually, emotionally and physically healthy life. In such cases, Extropianism is not a dogma but a model of an optimistic future – a profound hope to be worked towards, but not an absolute mandate which must be achieved no matter how terrible the cost .

INHUMAN TRANSHUMANISM?

As Max More realized from the start, the moral-philosophical aspects of Extropianism are key. Like Nietzsche, Extropian thinkers recognize that morals are biologically and culturally relative, rather than absolute. Who hasn't been struck by this at one time or another? We consider it OK to eat animals but not humans; Hindus consider it immoral to eat cows; Maori and other tribes until quite recently considered it acceptable to eat people. Even within the supposedly absolute morality of contemporary religion we find all kinds of relativist

arguments, such as a litany of situations in which it is acceptable to kill[6]. Or, consider sexual morals. Why are female sexual infidelity and promiscuity considered "worse" than similar behaviors on the part of males? This is common to all known human cultures; it comes straight from the evolutionary needs of our selfish DNA.

Given this blatant arbitrariness, it seems quite attractive to ignore human values altogether and focus one's attention on knowledge, understanding and power – qualities which seem to have an absolute meaning that morality lacks. In this vein, Nietzsche focused on personal power achieved through mental exploration and self-discipline; whereas the Extropians focus, by and large, on power achieved through technological advancement. They also share a focus on intellectual brilliance – and many, though not all, Extropians seem to take a worrisomely dismissive attitude toward those whom they feel don't have what it takes to make the next step on the cosmic evolutionary path (as exemplified in the Moravec quote above). Considering their emphasis on self-improvement, this attitude almost negates all their arguments and essentially deflates their claims to any moral high ground – if technology really can overcome all problems and raise us to the next level of advancement, and with technology guiding evolution it is now subject to the explicit decisions of intelligent beings rather than randomness + selection, what reason other than a defective moral and ethical system would there be for leaving anyone behind?

Moravec was playing Devil's Advocate in the quoted interview, but what we'd like to see is more Extropians taking an opposite point of view, and focusing on the value of transhumanist technologies for advancing the well-being of *every* sentient being. Whomever doesn't "have what it takes" now, oughtn't they, if the Extropian ideal is correct, be able to be improved through amazing advances in technology until they have got it? The further Extropian culture moves in this direction, the more we'll like it.

---

6 The argument that the admonition from God is against murder – that is, it is against killing without a good reason, not all killing – is a morally relativistic argument itself. If one merely needs to have a "good reason" to escape the judgment of murder, one can invent a wide variety of plausible excuses for killing essentially whomever they like. Weakly defined "explicit exceptions" such as self-defense leave quite a bit of room for creative relativism.

4 or 5 years ago, Ben posted a question on the Extropian e-mail list, either of us even started thinking about writing about Extropianism. Intellectually fishing, Ben posited in the post to the list that compassion, simple compassion, was an ethical universal, although it might manifest itself in different ways in different cultures and different species. He suggested that compassion, in which one mind extends beyond itself to feel the feelings of others and act for the good of others without requiring anything in return, might be essential to the evolution of the complex self-organizing systems we call cultures and societies – and expressed an essential disbelief that all human interaction is, or should be, economic in nature.

If you're expecting a story about a deep intellectual and ethical discussion on humanist and transhumanist philosophy in the context of an Extropian dialogue on compassion – well, no such luck. There was a bit of flaming, some impassioned but shallow Ayn Rand-ish refutations, and then they went back to whatever else they'd been talking about, unfazed by the heretical position that perhaps transhumanism and humanism could be compatible, that technological optimism wasn't logically and irrefutably married to the extremes of Libertarian politics. At that time, you could only belong to their e-mail list for free for 30 days; after that you had to pay an annual subscription fee. After Ben's 30 days expired, he chose not to pay the fee, bemused that this was the only e-mail list he knew of that charged members money, but impressed by their philosophical consistency in this matter. (Stephan never chose to pay the fee telling the Extropians he knew, such as Sasha, who invited him to participate that they were being inconsistent – participants ought to be paid for posting as well as the organization for hosting the list, thus the list should be free on the basis of barter. Now the list is free, though who knows if a number of people presenting similar arguments is why?)

We have more respect for the diversity of the Extropian community than we did when we first encountered it, or when Ben wrote the FAZ article that was the first version of this chapter. It's definitely a loose conglomeration of individualists – heck, even though we're not formally a members of the Extropian

group, Ben's spent some time on their e-mail list and has been to one of their conferences, and Stephan has spent a lot of time in-person with Extropians and at local Extropian get-togethers, so from the outside world's perspective we're virtually Extropians ourselves. Though some of the key philosophical *habits* of the group trouble us, any statement made about "Extropians" as a philosophical whole is bound to be a bad overgeneralization, more so than would be the case for many other social subgroups.

We admire Extropianism's courage in going against conventional ways of thinking, in recognizing that the human race is not the end-all of cosmic evolution, and in foreseeing that many of the moral and legal restrictions of contemporary society are going to be mutated, lifted or transcended as technology and culture grow. We too are outraged and irritated when governments stop us from experimenting with our minds and bodies using new technologies – chemical or electronic or whatever (collective responsibility against destroying ourselves is of course a good thing, but the idea that big-G Government as an entity is the mechanism for collective care against dangerous experimentation has proved rather fallacious upon even a cursory look at 20th century history). We find Extropian writings vastly more fascinating than most things we read. Extropian individuals are looking far toward the future, exploring regions of concept-space that would otherwise remain unknown, and in doing so they may well end up pushing the development of technology and society for the better. But yet, we're a bit vexed by the strain of Extropian thought that envisions Extropian human beings as supertechnological proto-Uebermensches, presiding over the inevitable obsolescence of humanity through the promotion of selfishness and the absolute worship of power and money. As a philosophy and a group of people, Extropianism is simultaneously profound, attractive, amusing and disturbing.

Nietzsche, like Sasha Chislenko, was generally an exemplary human being in spite of the inhuman aspects of his philosophy. Yet many years after his death, Nietzsche's work played a role in atrocities, just as he'd bitterly yet resignedly foreseen. In the back of our minds is a vision of a far-future hyper-technological Holocaust, in which cyborg despots dispense air at fifty dollars per cubic meter,

citing turn-of-the-millenium Extropian writings to the effect that humans are going to go obsolete anyway, so it doesn't make much difference whether we kill them off now or not. Extropian philosophy should be read, because explicitly Extropian thinkers have pondered some aspects of our future more thoroughly than just about anyone else. But in our view, some of the key themes in the Extropian community – particularly, the alliance of transhuman technology with simplistic, uncompassionate Ultra-Libertarian philosophy – must be opposed with great vigor. We ourselves are extremely far from being anti-Freedom, but recoil at the reduction of all human and environmental interaction to uncompassionate, self-obsessed, and dangerously (even in a practical, amoral sense) hypersimplistic models such as the monetary theory of economics (in which a single parameter – price – determines the absolute value of all things and beings).

Many of the freedoms the Extropians seek – the legal freedom to make and take smart drugs, to modify the body and the genome with advanced technology – will probably come soon. But we not only hope for but work towards and implore others to work towards, a situation in which these freedoms do not come along with a cavalier disregard for those living in less fortunate economic conditions, who may not be able to afford the latest in phosphorescent terabit cranial jacks or quantum-computing-powered virtual reality doodaddles, or even an adequately nutritional diet for their children.

Although both of the authors have had the privilege to grow up and live in a wealthy First World nation, neither of us has yet been particularly wealthy nor sheltered; we have both experienced the expectable batch of unpleasant and ambiguous life-experiences. Nevertheless, we both believe that we humans, for all our greed and weakness, have a compassionate core, and hope and expect that this aspect of our humanity will carry over into the digital age, trumping our less altriustic tendencies, and carrying over even into the transhuman age, outliving the human body in its present form and that this trait follows us throughout our evolution—wherever that may take us. We love the human warmth and teeming mental diversity of important thinkers like Max More, Hans Moravec, Eliezer

Yudkowsky and Sasha Chislenko, and great thinkers like Nietzsche – and hope and expect that these qualities will outlast the more simplistic, ambiguity-fearing aspects of their philosophies. Well aware of the typically human contradictoriness that this entails, w're looking forward to the development of a cyberphilosophy accepting what is great in Extropianism and moving beyond it in the explicit direction of compassion – a *humanist transhumanism*.

ELIEZER YUDKOWSKY

The typical techno-futurist guru has a huge variety of domains of interest and knowledge, but one, maybe two special obsessions. Moravec is robotics-focused; More is particularly into life extension and libertarian politics; Ben is an AI guy at heart, in spite of recent forays into biotechnology and other domains; Stephan has his preferences for philosophy and AI. On the other hand, Eliezer Yudkowsky, one of the more interesting young folks in the Extropian circle, focuses his thinking almost exclusively on ensuring the Singularity is beneficial by creating what he calls "Friendly AI." We find Eliezer's work particularly interesting in that it is, in its premise at least, a kind of humanist transhumanism. Unlike Moravec, Eliezer specifically and intensely wants the Singularity to help all humankind. He takes his altruism very seriously. "If one human dies," he says, "it subtracts from me."

It is uncertain how deeply Eliezer considers himself "Extropian," but he's a member of the Extropy organization, and we include him in this chapter because we first encountered him via his frequent postings on the Extropians email list. He has been an active member of the Extropian community for many years, in spite of having some profound disagreements with Max More and other Extropian stalwarts.

Yudkoswky shares with usthe idea that the probable best course to the delirious and universally beneficent cyberfuture is to create a computer smarter than us, one that can figure out all these other puzzles for us. To accomplish this goal of real computer intelligence, Eliezer champions the notion of a "seed AI," in

which one first writes a simple AI program that has a moderate level of intelligence, and the ability to modify its own computer code, to make itself smarter and smarter. His design for a "seed AI" is still evolving, and so far appears not to have achieved nearly the level of concreteness that we have with our Novamente system, but he's a clever guy and we don't doubt he'll come up with something interesting. Discussions on his "Singularitarian" e-mail list led to the formation of the Singularity Institute devoted to the creation of seed AI, and to the (apparently now defunct) company Vastmind.com, developer of a distributed processing framework that allows a collection of computers on the Net to act like a single vast machine (a project Webmind was also undertaking when we ran out of money).

Like many of the leading Extropians, Eliezer started his life as a gifted child; and, like many gifted children, he grew up neglected by the school system and somewhat misunderstood by his parents. He's followed a unique psychological trajectory: After the seventh grade, he was stricken with a peculiar lack of physical energy, which to some degree plagues him to this day. His parents tried to help him cope with this in various ways, but without success: only when they allowed him to take control of his own life and his own mind was he able to work his way back to a productive and functional state of mind. This experience, he says, taught him that even well-meaning, loving people who want to help you can do you a lot of damage, due to their lack of understanding. He cites this as one of the sources of his Libertarian political philosophy (do you see a trend emerging here?). Just as his parents tried to guide his life but failed in spite of good intentions, so does the government try to guide the lives of its citizens, but fails – and fails particularly where the vanguard of technology is concerned. Of course, every vanguard of new ideas feels this way about government in particular and society in general, but that is the mechanism by which the vanguard compels itself to enact change rather than merely pontificate about it.

Eliezer runs an e-mail list called "SL4" (sl4@sl4.org). He and his helpers moderate the list with a sense of humor and an iron hand (there are good, practical

reasons for this approach, but like Nietzsche and Chislenko, Yudkowsky also can't escape the contradictions in trying to make practical an absolutist philosophy like Ultra-Libertarianism). The control they exert on the list is occasionally a little overbearing, but, all in all, it keeps the list's signal-to-noise quality orders of magnitude higher than on the Extropians list (extropians@extropy.org). "SL4" stands for "shock level 4," where he defines a shock level as a measurement of "the high-tech concepts you can contemplate without…experiencing future shock." According to his Web page, http://sysopmind.com/sing/shocklevels.html, the first 4 shock levels are defined roughly as follows:

* SL0: The legendary "average person" is comfortable with this level of modern technology. SL0 technology is not on the frontiers of modern technology, but is the technology used in everyday life.

* SL1: Virtual reality, living to be a hundred, the frontiers of modern technology as seen by Wired magazine. Scientists, novelty-seekers, early-adopters, programmers, and technophiles are completely comfortable with this technology, but the "average person" is skeptical, perhaps wary.

* SL2: Medical immortality, interplanetary exploration, major genetic engineering, and new ("alien") cultures. The average Science Fiction fan is, or at least believes they are, comfortable with such things coming to be. Most people consider such things to be improbable, if not absolutely impossible.

* SL3: Nanotechnology, human-equivalent AI, minor intelligence enhancement, uploading, total body revision, intergalactic exploration. Extropians and transhumanists consider such things to be inevitable and are comfortable with this. Other kinds of futurists are skeptical, even wary, and the "average person" doesn't consider such things, except perhaps as Science Fiction scenarios vaguely familiar to those in the post-industrial nations – if any came to pass, the "average person" would be no less shocked than if God himself appeared before them.

* SL4: The Singularity, Powers (a term taken from Vernor Vinge's fiction, meaning superintelligent godlike-beings), complete mental revision, ultraintelligence, the total evaporation of "life as we know it." All but the

Singularity faithful are skeptical, even wary, and the "average person" considers any such possibility to be solely the domain of the supernatural and in the hands of God.

According to Yudkowsky, "The use of this measure is that it's hard to introduce anyone to an idea more than one Shock Level above - and Shock Levels measure what you accept calmly, not what you know about. There are very few SL4s…. If somebody is still worried about virtual reality (low end of SL1), you can safely try explaining medical immortality (low-end SL2), but not nanotechnology (SL3) or uploading (high SL3). They might believe you, but they will be frightened - shocked."

Periodically someone comes along and claims to have achieved SL5, but the claims are never very convincing. Perhaps the best "SL5" story heard on the SL4 list came in a discussion on psychedelic drugs, where someone said that SL5 is the realization that all human categorizations, including shock levels, are the almost-meaningless cognitive masturbations of our limited ape-like minds. But of course, the beauty of SL4 is that it pretty much encompasses SL5 in this sense. Once you've realized that mind and reality as we know it may be superseded by something totally different, you've got to question whether any of our ideas make any sense at all – or whether, from the perspective of our future super-superhuman "selves," our current ideas will seem about as profound and interesting as the metaphysical ruminations of mildly retarded dung beetles.

Of course, the "shock level" categorizations are a formalism about an individual's hopes and beliefs. There is no empirical evidence that anyone will accept SL4 with calmness or glee, because we can define so little about it. Rather, someone who claims to be SL4 has a specific notion of what a possible Singularity could be like and would like to see it implemented – not unlike a Christian waiting for Rapture, except that rather than have faith and wait, a Singularitarian must actively work to *create* the Singularity and through ethical and engineering actions set in motion now, guide the creation of Singularity such that they can achieve a positive Transcendent state (and hopefully not be cast asunder by the reign on earth of an evil AI superbeing – hence the emphasis on

Friendly AI and its predecessors, like Asimov's Laws of Robotics, and its successors, like our Voluntary Joyous Growth philosophy, which you'll hear about in Chapter 15).

Yudkowsky has introduced the the term "sysopmind" (sysopmind.org was the former Internet domain replaced by sl4.org) to refer to the notion of the Sysop, a superintelligence that has achieved near-complete control over the structure of matter and energy in some region of spacetime, and thus plays the role of a (hopefully benevolent) system administrator for some portion of the world. A long thread on the SL4 group discussed the possibility of some lesser mind, at some future time, hacking into the Sysop and co-opting its powers for its own devious purposes. That's a rather literal extrapolation of the analogy between present-day computer technology and a possible future AGI, but not an uncommon one in sci-fi literature, and it reflects a very real concern about the seduction of a seemingly flawless system and the care people must take in assuming that someday a being will exist which frees all other beings from responsibility towards their own well-being and that of the system as a whole.

Of all the various and wacky discussions on SL4, the one that amused us most was the one about the contrasts between Eliezer's and Ben's personal lives. Eliezer was raised in a strict Jewish family, and he shows this influence very strongly in his life and his work, even though he is an avowed atheist. His devotion to the Singularity is definitively monastic. As he has publicly declared many times, he does not "fight, drink alcohol, take drugs or have sex." He recognizes the pleasures that can be obtained from these things – not through experience but through accepting others abstractions about such things into his own thinking – but he does not want to get involved with activities that will evoke strong animal emotions and thus distract him from 100% focus on the Singularity. He specifically laid out, in an e-mail to the SL4 list, the only conditions under which he could see violating these precepts. For instance, if a wealthy woman approached him and told him she would fund his work on the Singularity, but only if he would marry her – then, he says, he would marry her, not for her sake, but for the Singularity.

In e-mail dialogues with Eliezer on SL4, Ben presented doubts as to the necessity of this monkish approach, saying he was highly devoted to pushing toward the Singularity as well, but didn't see why it would be necessary to give up having a rewarding personal life in order to manifest this devotion in a highly effective way. We work long hours because we enjoy it, and believe what we're doing is extremely important, both for ourselves (we want to build an AI that will help us figure out how to live forever!) and for the human race, and the evolution of intelligence overall. But we still take time for my wife and children, and various other pursuits like composing music, playing the piano (not that expertly, but enthusiastically), occasional outdoor sports, writing this book, and even other jobs (running companies, writing software, working on movies – whatever pays the bills). Our own quest for the Singularity is partly altruistic, but partly a consequence of our boundless curiosity and desire for adventure and excitement – the same thing that pushes us to try new sports, to travel to different countries; and the same thing that, in our college years, impelled us to experiment with various mind-altering substances (those which we had judged to be no less safe than the dangers imposed by the vagaries of daily life – alcohol, driving cars, weather, and so on).

Yudkowsky responded that his own quest to bring about the Singularity through creating seed AI was purely altruistic in motive. He also called himself a true romantic, stating that he knew a real love relationship would take too much of his time, so he was just going to steer away from that domain of life altogether. Waxing poetic, he said "...love is a cathedral that you build together, a rose that you grow and water together *for its own sake* "

Upon reading this characterization of love, Ben couldn't help but respond to him with the Charles Bukowski line, "Love is a dog from hell."

Both tired of e-mailing for the day, Eliezer went back to helping humanity, Ben returned to his own endless work with the somewhat maudlin thought that love didn't really have much to do with any of these silly words anyway – but he sure hoped it would survive the Singularity in one form or another.

About a year later, Eliezer posted a very funny Web page on the theme of his own celibacy, entitled Love and Life Just Before the Singularity[9] . The page ended with the following survey:

****

*** Table 12.1 goes here ***

****

Table 12.1 Eliezer's web page about Love and Life Just Before the Singularity

In a later version of the page, the following note appeared next to a report of the results:

(Note: 22 votes were cast for the fifth option by a multiple voter… You can subtract 22 votes from the fifth option to arrive at the correct total. This kind of spoiler action is not welcome; please do not do so.)

As it happens, that multiple voter was Ben's girlfriend at the time – an attractive young physics student who was a big fan of eternal love, flowers, Puccini and all things romantic (as well as, incidentally, an obsessive fan of black metal music). I think she hoped to influence Eliezer to experience the glories of romance – but, predictably enough, she succeeded only in annoying him, and his dogma about the subject seemed to hold firm. In 2002, following the "Cathedral" thread, there followed a hilarious SL4 e-mail thread in which a group of others discussed the theme of the "monk versus the warrior." Somehow Ben had become a warrior – which seems rather amusing since in point of fact, much like Eliezer, he spends an excessive proportion of his waking hours on his butt in front of the computer, and is generally inclined towards cooperation rather than conflict. Then, in early 2005, Eliezer showed up at a Bay Area social gathering with a woman whom he proudly and humorously introduced as his "consort." When questioned by a friend (neither of us were there), she reported that she had been attracted to him via his online essay about why he would never have a girlfriend.

[9] http://yudkowsky.net/essays/lovelife.html

In the end, it seems we have to give Eliezer some kind of prize for finding a creative way to use the Internet to pick up girls! We are relieved, however, that there is at least some flexibility in his dogma, and that he's open to change his ideas in the light of new information. Not only *can* one contribute to the positive evolution of human society while having a fulfilling social life, but one *should* do so to the best of their abilities, because these developments proceed in the context of human relations and how can you change the world for the better if you don't relate to the people in it?

The main practical consequence of Eliezer's extreme altruism – apart from his abstemious lifestyle – is his focus on the notion of Friendly AI. He wants the Singularity to benefit all people, which in our view is a vast improvement of the Moravecian "to hell with the poor" attitude. Yudkowsky believes the Singularity will be brought about by a seed AI transforming itself to superintelligence and then making endless further inventions and innovations, until – if his formulation of a positive scenario occurs – it becomes a Sysop, a system administrator for the computational system that is the universe, regulating the dynamics of the universe according to human-friendly and sentience-friendly ethical principles. It follows from this that making the seed AI as benevolent as possible to humans is an important idea. Of course, it can't be known that a human-friendly seed AI will become a human-friendly Sysop. But, in Eliezer's view, the lack of absolute knowledge in this regard is a lame excuse for not trying.

We'll dig into, refute some of, and expand upon some of Eliezer's ideas – which are quite subtle – in more detail in Chapter 15. On a crude level, however, his key idea is that Friendliness (to humans) should be at the top of any seed AI's goal system. He doesn't mean "friendliness" in quite the typical sense, and has recently introduced a concept of "humaneness" as a way of explaining his notion of Friendliness, and we'll go into this in great detail in Chapter 15. For now, the common notion of humaneness will suffice, and to understand that in his perspective the Friendliness goal should be paramount to any AI that is created. Other goals, such as learning things or surviving, should be represented within the system as subgoals of Friendliness. The system should try to survive, but not

because survival is its ultimate goal – but rather, because surviving will allow it to help people more.

When we invited him to give a talk at Webmind Inc. in late 2000, he lectured us passionately on the need to give our AI system a friendly goal system. He was a little concerned that the Webmind AI Engine might undergo a "hard takeoff" – a rapid transition from intelligence to superintelligence via progressive self-modification – and that if it didn't have the right goal system inside it at that point, the future of humanity might be a bleak one. Since we involved with the Webmind project were painfully aware of the incomplete state of our codebase, we were not so concerned about this possibility (our concern about running out of "hard cash" before the system was able to produce any "hard results" proved, sadly, to be more well-founded in that particular instance).

Reactions to his talk amongst the Webmind Inc. staff ranged from deep interest, to distant amusement, to outright disgust at the silliness and impracticality of the topic. Generally speaking, our Webmind colleagues were absorbed with the practical problems of trying to create real digital intelligence, whereas Eliezer was then more concerned with the various philosophical and futuristic issues that would arise once a truly intelligent AI system is completed. In his talk – reflecting his conceptual bias as of 2000, which is different from his bias as of 2005 -- he focused on the issue of programming AI systems with Friendly goal systems, more than on educating AI systems or the philosophy of humaneness. His emphasis on "wiring in Friendliness" definitely struck everyone powerfully, one way or another. Among the milder responses, one of our Brazilian software engineers raised his hand and politely said: "But perhaps the most important thing is not the in-built goal system, but whether we teach it by example." The friendlier we are, in other words, the friendlier our AI systems are going to be.

The issue is clear and poignant. What the Brazilian engineer was suggesting was that, if our superhuman AI grows up watching us act as though most humans are dispensable and irrelevant, perhaps it will, in its adulthood, believe that we too are dispensable and irrelevant. On the other hand, perhaps, as

Eliezer says, it will grow up and understand that building it was the best thing the cyber-elite could do for humanity as a whole, and it will then proceed to spread joy and plenty throughout the land. Who knows for sure? However, the idea that we can manufacture a system which will protect ourselves from each other, while appealing, is very unlikely if we demonstrate in its infancy a profound disregard for life. Even if we could fabricate such a thing from "whole cloth" it would still be a particularly pathetic kind of irresponsibility to not teach both any AGI's we develop in the future, and the children we have now, to be humane and respect life, intelligence, and growth *by example*.

We find the motivation behind the Friendly AI concept admirable; but the Novamente digital mind design which embodies as many of our ideas in this book as we currently know how to formulate does not have quite so rigid a goal system as Eliezer suggested in his 2000 talk at Webmind Inc. We tend to agree with the view the Brazilian engineer expressed during Eliezer's talk -- that experience and education are going to make more of a difference to the Friendliness or otherwise of a seed AI, than any structure explicitly embedded in its goal system. Certainly, we'll be curious to see what kind of AI architecture Eliezer comes up with; no doubt his AI design will be more compatible with his own thoughts on goals and their management than anything either of us might design. He has recently written a paper on "Levels of Organization in General Intelligence" (Yudkowsky, 2002) which takes some serious steps in this direction – but there is still a lot further to go. In the spirit of both intellectual curiosity and cooperation, we hope that we'll see more developments from Yudkowsky regularly in the years to come.

A DIALOGUE ON HUMANIST TRANSHUMANISM

Though he's been chatting on the SL4 list regularly for the last few years Ben's last serious foray onto the Extropians e-mail list was in April 2001. Webmind Inc. had just folded, and in need of some distracting entertainment he thought it would be amusing to bring up the old social-consciousness theme again, though with a less adversarial approach than had been taken in the post years before.

There was a thread discussing how hard it was to get Extropian ideas accepted into mainstream culture. Suggesting that, as a counterbalance to the "scary" aspects of deep futurism it might be valuable for the Extropian community as a group to become involved in some kind of socially beneficial project, perhaps spreading technology to the disadvantaged, Ben had intended to suggest their participation in either the Brazilian net computing project or, more likely, a program he'd heard of that for approximately $14000 allowed anyone to sponsor a Cambodian elementary school.

Eliezer, who was active on the Extropians list at that time, responded firmly that the best thing he could to for the disadvantaged of the world was to focus all his time and effort on bringing about the Singularity, because the Singularity will help everyone. He said,

> "If you can't, on a deep emotional level, see the connection between my work and the starving people in the Sudan, then this - from my perspective - is an emotional peculiarity on your part, not mine."

Ben replied as follows:

> "I do perceive the connection, of course, both rationally and emotionally. Your work has a decent chance of increasing the probability that the Singularity is good for humans, and it's therefore a very important kind of work. I feel the same way about my own work. AI technology is going to do a lot of good for a lot of people, someday. I do feel AI will be a profoundly positive technology for humans, not a negative one like, say, nuclear weapons, which I wouldn't enjoy working on even if it were intellectually stimulating.
>
> But yet, for reasons that are still not easy for me to articulate, I feel a bit of discomfort with *solely focusing one's life* on this type of

> compassionate activity -- on "helping people by doing things that will help them in the future but don't affect them at all right now." This is a good kind of activity to be doing, for sure. But yet, I feel that, in general, this kind of long-term helping of others can be conducted better if it's harmonized with a short-term helping of others. “

Not surprisingly (and not too disappointingly – we love Eliezer’s work and don’t really want everyone to think exactly the same way we do), he wasn’t convinced. He asked:

> “How do you resolve issues like these? Split your efforts between both alternatives to maximize output. How much money is spent on attempts to actually ship food directly to the poor? Lots. How much money is spent on direct efforts to implement the Singularity? We can both personally attest, Ben, that there is not much.”

To his:

> “There is absolutely *nothing* I could do that would help the rest of the world more than what I am already doing. ”
> Ben replied:
> “In my view, given the numerous uncertainties as to the timing and qualitative nature of the Singularity, it is irrational of you to hold to this view with such immense certainty.
> Actually, I honestly feel that if you spent a year teaching kids in the Sudan, you'd probably end up contributing MORE to the world than if you just kept doing exactly what you're doing now. You'd gain a different sort of understanding of human nature, and a different sort of connection with people, which would enrich your

> work in a lot of ways that you can hardly imagine at the moment. Not to mention a healthy respect for indoor plumbing!!"

Samantha Atkins, a long-time Extropian and a very thoughtful person, responded as follows, presenting a more compatible view to ours:

> "Perhaps there is a productive middle ground. Some of us could say more about precisely how the Singularity, and the technologies along the way, can be applied to solving many of the problems that beset real people right now. We can produce and spread the memes of technology generally and AI, NT and the Singularity in particular as answering the deepest needs, hopes and dreams of human beings.
>
> As part of this we also need more of a story about the steps up to Singularity as involves the actual lives and living conditions of people. That we will muddle along somehow while a few of the best and the brightest create a miracle is not very satisfying. What kind of world do we work toward in the meantime? What do we do about poverty, about technology obsoleting skills faster than new ones can be acquired, about creating workable visions including ethics and so on? What is our attitude toward humanity?
>
> The world we make along the way will shape the Singularity and may well determine whether it occurs at all."

Ben presented this parable: "Suppose you're stuck on a boat in the middle of the ocean with a bunch of people, and they're really hungry, and the boat is drifting away from the island that you know is to the east. Suppose you're the only person on the boat who can fish well, and also the only person who can paddle well. You may be helping the others most by ignoring their short-term needs (fish) in favor of their longer-term needs (getting to the island). If you get them to

the island, then eventually they'll get to an island with lots of food on it, a much better situation than being on a sinking boat with a slightly full stomach.

If the other people don't realize the boat is drifting in the wrong direction, though, because, they don't realize the island is there, then what? Then they'll just think you're a bit loopy for paddling so hard. And if they know you're a great fisherman, they'll be annoyed at you for not helping them get food.

What, is this little parable missing?

Sociality. If you feed the other people, they'll be stronger, and maybe they'll be more help in paddling the boat. Furthermore, if you maintain a friendly relationship with them by helping them out in ways that they perceive as well as ways that they do not, then they're more likely to collaborate creatively with you in figuring out ways to save the situation. Maybe because of their friendship with you, they'll take your notion that there's an island to the east more seriously, and they'll think about ways to get there faster, and they'll notice a current that you didn't notice, floating on which will allow you to get there faster with less paddling."

The difference here is between the following two attitudes:

1) Seeking to, as a lone and noble crusader, save the world in spite of itself, generally in conflict with others perceived short term goals for themselves

2) Seeking to cooperatively engage the world in the process of saving itself

To do 2, I pointed out, it's not enough to do things that you perceive are good for everyone in the long run. You have to gain the good will of others, and work with them together on things that both they, and you, feel are important. While you can influence them to change their goals to be more compatible with

your world view, history shows that forcing them rather than convincing them to behave in such a manner results in an unstable system.

Of course, it's impossible and undesirable to have a consensus among all humans as to what is good and what is bad. Like most things in the human world, the distinction between 1 and 2 is fuzzy rather than absolute. Leadership, which is the quality which allows someone to engage people in changing things when they'd much rather they stay the same, is the art of negotiating that fuzzy area between the two extremes. A good leader does this well, and a poor one gets stuck at one of the poles and guides his or her people into disaster – either through extreme action, or extreme inaction. Following up a discussion of the above parable, Ben replied to Eliezer:

> "I realize that you, Eli, are trying to cooperatively engage the world in the process of saving itself your way, by publishing your thoughts on Friendly AI. But I have an inkling that the way to cooperatively engage the world in the process of saving itself ISN'T to try to convince them to see them your way through rational argumentation. Rather, it's to try to enter into a real dialogue where each side (transhumanists vs. normal people in this case) makes a deep and genuine effort to understand the perspective of the other side."

Eliezer's reply was both well-thought-out in its details, and predictable in its overall course:

> "*Ben: f the other people don't **realize** the boat is drifting in the wrong direction, though, because, they don't realize the island is there, then what? Then they'll just think you're a bit loopy for paddling so hard. And if they know you're a great fisherman, they'll be annoyed at you for not helping them get food....*

Except I'm *not* a great fisherman. I am a far, far better paddler than I am a fisherman. There are *lots* and *lots* of people fishing, and nobody paddling. That is the situation we are currently in.

Ben: What is my answer missing? Sociality.

Very well, then, let's look at the social aspects of this.

Your answer makes sense for a small boat. Your answer even scales for a hunter-gatherer tribe of 200 people. But we don't live in a hunter-gatherer tribe. We live in a world with six billion people. From a "logical" perspective, that means that it takes something like AI to get the leverage to benefit that many people. From a "social" perspective, it means that at least some of those people will always be ticked off, and hopefully some of them will sign on.

Plans can be divided into three types. There are plans like Bill Joy's, that work only if everyone on the planet signs on, and which get hosed if even 1% disagree. Such plans are unworkable. There are plans like the thirteen colonies' War for Independence, which work only if a *lot* of people - i.e., 30% or 70% or whatever - sign on. Such plans require tremendous effort, and pre-existing momentum, to build up to the requisite number of people.

And there are plans like building a seed AI, which require only a finite number of people to sign on, but which benefit the whole world. The third class of plan requires only that a majority *not* get ticked off enough to shut you down, which is a more achievable goal than proselytizing a majority of the entire planet.

Plans of the third type are far less tenuous than plans of the second type.

And the fact is that a majority of the world isn't about to knock on my door and complain that I'm doing all this useless paddling instead of fishing. The fall-off-the-edge-of-the-world types might

knock and complain about my *evil* paddling, but *no way* is a *majority* going to complain about my paddling instead of fishing. Certainly not here in the US, where going your own way is a well-established tradition, and most people are justifiably impressed if you spend a majority of your time doing *anything* for the public benefit."

As Brian Atkins10 said:

"The moral of the story, when it comes to actually having a large effect on the world: the more advanced technology you have access to, the more likely that the "lone crusader" approach makes more sense to take compared to the traditional "start a whole movement" path. Advanced technologies like AI give huge power to the individual/small org, and it is an utter waste of time (and lives per day) to miss this fact."

Ben's response to Eli was:

"Eli… here is my sense of things, which I know is different than yours.

10 Brian Atkins, another Extropian, was for many years Eliezer's patron – meaning that he was the primary source of funding for the Singularity Institute, whose primary practical function was the financial support of Eliezer Yudkowsky. (In late 2002, for personal-finance reasons, Brian decreased his support of Eliezer's fellowship, causing Eliezer to seek alternate sources of financing.) Brian is not an AI wizard, but he does have a quick mind, a broad knowledge base, and a good sense for future technology. Not surprisingly, he is pretty close to "true believer" in Eliezer – not that he's sure Eliezer's work will certainly save the world, but that there's a enough of a chance to merit some investment in this work. He and his wife also have a lot of personal affection for Eliezer, and in some sense took Eliezer under their wing. When Eliezer began to work for the Singularity Institute, funded by Brian, he moved to Atlanta where the Atkins live (more recently he's relocated to the San Francisco region). His defense of the "lone crusader" approach was thusly as much a practical defense as a philosophical one.

There's the seed AI, and then there's the "global brain" -- the network of computing and communication systems and humans that increasingly acts as a whole system.

For the seed AI to be useful to humans rather than indifferent or hostile to them, what we need in my view is NOT an artificially-rigged Friendliness goal system, but rather, an organic integration of the seed AI with the global brain.

And this, I suspect, is a plan of the second type, according to your categorization....

*Eli: And the fact is that a majority of the world isn't about to knock on my door and.complain that I'm doing all this useless paddling instead of fishing. The fall-off-the-edge-of-the-world types might knock and complain about my *evil* paddling, but *no.way* is a *majority* going to complain about my paddling instead of fishing. Certainly not here in the US, where going your own way is a well-established tradition, and most people are justifiably impressed if you spend a majority of your time doing *anything* for the public benefit.*

My belief is that one will work toward Friendly AI better if one spends a bit of one's time actually engaged in directly Friendly (compassionate, helpful) activities toward humans in real-time. This is because such activities help give one a much richer intuition for the nuances of what helping people really means.

This is an age-old philosophical dispute, of course. Your lifestyle and approach to work are what Nietzsche called "ascetic", and he railed against ascetisicm mercilessly while practicing it himself. I'm fairly close to an ascetic by most standards -- I spend most of my time working on abstract stuff, and otherwise I don't do all that much else aside from play with my kids -- but, yes, I admit it, I spend some of my time indulging myself in the various pleasures of the real world ... and some of my time doing stuff like teaching

> in my kids' schools, which is fun and useful to the kids, but doesn't use my unique talents as fully as working on AI. I think my work is the better, not the worse, for these "diversions".... But perhaps it wouldn't be so for you.... Perhaps the philosophical dispute over the merits of asceticism just comes down to individual differences in personality."

All in all, Eliezer didn't convince us, and Ben didn't convince him, but some headway was made in terms of understanding each others' points of view. We are happy that Eliezer's altruistic attitude exists; it's a great counterbalance to the more draconianly elitist strains of Extropian thought even if it is an extremely abstract kind of altruism focused solely on ultra-long-term benefit. It is important that such things be discussed, even if the discussions are at times rambling and silly. Just because we in the deep-futurist camp don't sympathize much with the ethical concerns of the mainstream media (Is human cloning somehow intrinsically immoral? Give us a break! Is creating humans by any other means intrinsically immoral? The difference scientifically is trivial, if you actually understand reproduction starting at the cellular level.), that doesn't mean ethical issues are irrelevant to our thinking and our work.

Pursuing the "Lone Crusader" approach to social change, minor or massive, can lead to a situation where one believes one's own myth. The Lone Crusader can easily become so detached from their social context that their perspectives on what will make the world better can become useless or dangerous, because the tend towards engaging wtih only themselves and their few like-minded supporters. Without a broader social context, critique and refinement of ideas just doesn't happen, and an idea which may have had initial merit can devolve into a perverted form that is propped-up by cronyist totalitarianism.

On the other hand, an unwavering devotion to the "Consensus Builder" approach can lead to endless, pointless bureaucratism in which all effort is spent on the frictional costs of the mythical absolute consensus, and can itself run an interesting idea aground on the Plutonian shores of the dictatorship of mob rule in

which every unusual, challenging, independent idea is crushed. As with so many things, the key to success is not absolute devotion to a formulaic ideal but rather an ethically guided flexibility in which two competing approaches, either of which is independently unviable, are synthesized into a morally guided but practically workable approach.

In part, we agree with Eliezer's perspective that not enough energy is spent trying to ensure a positive Transcension of humanity beyond our current state. More resources should be spent not only on the philosophy of such things, but on the research itself into AGI, genetics and biotechnology, renewable energy, space travel, and all the other things we will need to evolve to the next level of sentient potential. He is correct that more people are trying to solve the problems of poverty and starvation directly than are trying to bring about their end by lifting-up all of humanity to a Transcendent state in a material, non-supernatural way. However, relative to the number of people in the world, and the percentage of those in poverty, his claim that *lots and lots* of people are is perhaps overstated.

We choose, instead, to spend most of our resources on frivolous pursuits. Currently, we have decided, as a society, to spend more of our resources developing a few dozen automobiles, or a few thousand articles of clothing, each year which are trivially different from last year's – or any one of thousands of similarly trivial pursuits. Rather than focusing on important things that will elevate all of us to a higher level of being, whether it be the sciences of such things, or philosophy and art which will inspire the science to reach further and be more humane, we allow ourselves to be seduced by the short-term excitement of trends and gossip. Such distractions are amusing in moderation, but we've allowed them to crowd-out any long term thinking about things that are beautiful and transcendent such as the arts, science, mathematics, and philosophy. Furthermore, in the aggregate we spend more time acquiring, counting, and figuring out new ways to acquire and count resource tokens – money in industrialized societies – than actually doing anything with those tokens, at this point probably even more time than feeding our basal desires for food and sex, from which our desire for money and power biologically originated.

We agree with Eliezer that we as a race are predominantly wasting our time on frivolous pursuits, but not that we should take the time to pursue Transcendent developments away from family, art, or social compassion. Rather, we ought to draw more resources away from the frivolous, no matter how much money and power we might gain from them *in the short-term*, and apply those societal resources towards the things which will make life better. All of the barriers to humanity moving to its next level of mental and social development are now economic, that is, they are a matter of how we choose to allocate our resources. All the compelling environmental reasons to do so are becoming increasingly blatant – we need to repair and improve our environment, but we seem not to be willing to do that as our current selves – and all the technological and scientific progress we need to guide our own evolution instead of leaving it solely up to chance is proceeding at a rate limited primarily by funding (which represents the social value of a pursuit in our society, and thus the desirability of pursuing that thing amongst people in our society, and thus the quality and dedication of talent that a pursuit can attract). You may believe that scientific barriers to true immortality are insurmountable, that Tipler and others are just repackaged mysticism, and medical immortality is another Alchemy – but those are the big leaps. We are close enough now to take the next step to the level of being that is one-beyond what we are currently, if we're willing to embrace self-determination and make the effort rather than waiting for random or preordained intervention from outside ourselves. In essence, we can either chose to reformulate Coca-Cola another dozen times (and a million other similar pursuits), or to try to improve the condition of all humanity.

Like Eliezer, we believe that the Singularity, in some form which may or may not resemble anything we can currently imagine, can be brought about in a way that benefits everyone, or nearly everyone. We're not sure that the path to this conclusion is as *simple* as the creation of a human-friendly seed AI, however. I think this a laudable goal, but I also think it's important to bring as much of the world as possible into the process of creating the Singularity, not just proceed apace with massive technological and social changes – changes to life itself –

with all the decision making power about those changes concentrated in a brotherhood of Elites (some of whom, like Moravec, are already itching to cast aside the unbelievers who dare question their infinite wisdom about such things) . The Global Brain and Mindplex ideas may be critical here (see Chapters 2, 3 and 14).

If the first AGI doesn't achieve superintelligence locked in a box, but rather through ongoing interaction with humans in all nations across the world, then its mind stands a good chance of being intrinsically human-focused and human-friendly as a consequence of its upbringing. It will be both a separate being, an individual AI mind, and part of a symbiotic mind of sorts involving little bits of millions of people. As it invents new technologies, it will want to invent not only technologies to make itself smarter, but also technologies to improve the human component of the symbiotic AI-human-internet mind. Having a human-friendly goal system is fine, but in any system flexible enough to be an intelligence, goals and motivations are going to shift over time (and with the assumption that self-rewriting is a prerequisite for AGI evolution, meta-abstract self-analysis and reformulation of even low-level goal encodings would be part and parcel of its existence).

To be really meaningful and stable, a human-friendly goal system must be allowed to evolve and mature through intensive mutually rewarding interactions with the mass of human beings, and the Global Brain path to superintelligent AI would seem to have the potential to accomplish this -- *if* we can carry it off in a positive manner.