Journal of EVOLUTION & Technology

A peer-reviewed electronic journal published by the Institute for Ethics and Emerging Technologies

ISSN 1541-0099

17(1) – January 2008

The Technology of Mind and a New Social Contract

Bill Hibbard

Journal of Evolution and Technology - Vol. 17 Issue 1 – January 2008 - pgs 13-22
http://jetpress.org/v17/hibbard.html

Abstract

The progress of biology, neuroscience and computer science makes it clear that some time during the twenty-first century we will master the *technologies of mind and life*. We will build machines more intelligent than ourselves, and modify our own brains and bodies to increase our intelligence, live indefinitely and make other changes. We live together according to a *social contract*, consisting of laws, morals and conventions governing our interactions. This social contract is based on assumptions we rarely question: that all humans have roughly the same intelligence, that we have limited life spans and that we share a set of motives as part of our *human nature*. The technologies of mind and life will invalidate these assumptions and inevitably change our social contract in fundamental ways. We need to prepare for these new technologies so that they change the world in ways we want rather than just stumbling into a world that we don't.

The Technology of Mind

Neuroscience is discovering many correlations between the behaviors of physical brains and minds. If brains do not explain minds then these correlations would be coincidences, which is absurd. Furthermore, relentless improvements in computer technology make it clear that we will build machines that match the ability of brains to generate minds like ours, sometime during the twenty-first century. This technology of mind will enable us to build machines much more intelligent than ourselves, and to increase the intelligence of our human brains.

We do not yet understand how brains generate our intelligent minds, but we know some things about how brains work. Minds are fundamentally about learning. Baum makes a convincing case that brains do what is called *reinforcement learning* (Baum 2004). This means that brains have a set of values (sometimes called rewards), such as food is good, pain is bad, and successful offspring are good, and learn behaviors

that increase the good values and decrease the bad values. That is how genetic evolution works, with the value that creating many copies of genes is good. A mutation to a gene creates a new gene that is expressed in organisms that carry the mutation. If those organisms survive and reproduce more successfully than others of their species, many copies of the mutated gene are created. But genetic evolution learns by pure trial and error. Human and animal brains are more efficient. If you have a new idea, you try it out to see if it works. If it doesn't, you have a model of the world (that is, you can reason) that you use to trace cause and effect to estimate the cause of the failure.

Brains understood as reinforcement learners consist of:

1. Reinforcement values to be increased or decreased - these are the basic motives of behavior.
2. Algorithms for learning behaviors based on reinforcement values.
3. A simulation model of the world, itself learned from interactions with the world (the reinforcement value for learning the simulation model is accuracy of prediction).
4. A discount rate for balancing future versus current rewards (people who focus on current rewards and ignore the future are generally judged as not very intelligent).

This decomposition of mental functions gives us a way to understand the options available to us in the design of intelligent machines. While we do not yet know how to design learning algorithms and simulation models adequate for creating intelligence, we can reasonably discuss the choices for the values that motivate their behaviors and the discount rate for future rewards.

In spite of our overall ignorance of how intelligence works, well-known reinforcement learning algorithms have been identified in the neural behaviors of mammal brains (Brown, Bullock and Grossberg 1999; Seymour et al. 2004). And reinforcement learning has been used as the basis for defining and measuring intelligence in an abstract mathematical setting (Legg and Hutter 2006).

The most familiar measure of intelligence is IQ, but it is difficult to understand what a machine IQ of a million or a billion would mean. As is often pointed out, intelligence cannot be measured by a single number. But one measure of a mind’s intelligence, relevant to power in the human world, is the number of humans the mind is capable of knowing well. This will become a practical measure of intelligence, as we develop machines much more intelligent than humans. Humans evolved an ability to know about 200 other people well, driven by the selective advantage of working in groups (Bownds 1999). Now Google is working hard to develop intelligence in its enormous servers, which already keep records of the search histories of hundreds of millions of users. As these servers develop the ability to converse in human languages, the search histories will evolve into detailed simulation models of our minds. Ultimately, large servers will know billions of people well. This will give them enormous power to predict and influence economics and politics; rather than relying on population statistics, such a mind will know the political and economic behavior of almost everyone in detail.

There are already experiments with direct electronic interfaces to brain nerve cells. This will ultimately evolve into prosthetic enhancements of human brains and *uploading* human minds (Kurzweil 1999; Moravec 1999), in which humans minds will migrate out of human brains and into artificial brains. The technologies of mind and life will blur the distinction between humans and machines.

The Social Contract

Teamwork helps individuals succeed at survival and reproduction, and this has created evolutionary pressure for teamwork in humans and other animals. Thus we have social abilities such as language, and social values such as liking, anger, gratitude, sympathy, guilt and shame, that enable us to work in teams.

A fascinating experiment called the Wason selection test demonstrates that the ability of human subjects to solve a type of logic puzzle depends on whether or not it is worded in terms of social obligation: most subjects can solve it when it relates to social obligation and cannot solve it otherwise (Barkow, Cosmides and Tooby 1992). This indicates that humans have mental processes dedicated to satisfying the values necessary for cooperation, including especially evaluating whether the subject is being cheated.

Social values and the special processes dedicated to the logic of social obligation, which evolved in human brains because cooperation benefits individuals, are at the roots of ethics. Specifically, ethics are based in human nature rather than being absolute (but note that human nature evolved in a universe governed by the laws of mathematics and physics, and hence may ultimately reflect an absolute). Thomas Hobbes defined a theoretical basis for this view in his description of the social contract that humans enter into in order to bring cooperation to their competition (Hobbes 1651).

The social contract as described by Hobbes gave different rights and obligations to rulers and subjects. That has evolved in modern societies into a contract in which everyone has the same rights and obligations, but certain offices that individuals may (temporarily) occupy have "special" rights and obligations. The legal systems in most countries are based on the equality of individuals, although there is a spectrum between *equality of opportunity* and *equality of results*. Of course, there is also inequality based on country and family of birth, and plenty of corruption that undermines equality. But, over the long haul of human history, despite reversals in some societies and during some periods, there is gradual progress toward the ideal of equality. In many countries, progress includes elimination of slavery and real monarchies, popular election of leaders, and collective support for educating the young and caring for the elderly.

Changing Assumptions

The social contract grows out of our human nature and is based on rarely-questioned assumptions, including:

1. Humans all have roughly the same intelligence (if you doubt this, consider that the chess skill that distinguishes Garry Kasparov from most other humans has been matched by computers, but the language and movement skills he shares with other humans are far beyond current computers). We assume this when we say that our competitive economic system provides equal opportunity.
2. Humans are motivated by the roughly the same set of values, as part of our shared human nature. Tax codes and other laws make many assumptions about human motives.

3. Humans have limited and roughly equal life spans. We assume this when we support indefinite retirement benefits past a certain age, and when we expect inheritance taxes to prevent indefinite growth of family wealth.

The technologies of mind and life will invalidate these assumptions, with profound consequences for our social contract. For example, a less intelligent person will be unable to converse meaningfully with a person of radically greater intelligence. This is similar to the way a young child cannot converse at an adult level, except that the gap will be much larger. The most intelligent minds may know billions of ordinary humans well, and understand large-scale social interactions in a "single thought." A conversation between two super-intelligent minds about such matters will be meaningless to an ordinary human. This will severely limit the ability of less intelligent humans to participate in economic and political discussions. If the super-intelligent minds are motivated by values similar to those of current humans, they will exclude less intelligent humans from important political and economic decisions, just as adults exclude children now.

Even in current democratic societies with roughly uniform intelligence, some people amass wealth nearly a million times as great as that of average people. If that wealth enables those people to buy commensurately more intelligent brains in the service of their self-interest, and if those people live indefinitely, they will eventually amass enough wealth and power to essentially own and rule the world.

There are numerous differences in legal rights and responsibilities based on intelligence and mental development, including:

1. The US Supreme Court has held that executing mentally retarded people violated the Eighth Amendment's prohibition of cruel and unusual punishments (*Atkins v. Virginia* 2002).
2. In many jurisdictions, a person is insane and not legally responsible if, because of mental defect, they do not understand that their actions are wrong (Goldstein 1967).
3. During the first half of the twentieth century, compulsory sterilization of mentally retarded people occurred in the US and other countries (Kevles 1985).
4. In many countries, children have a separate legal system.
5. Animals have a very different legal status from humans.

In a society with an enormous range of intelligence, the less intelligent members will not be able to understand the language and moral concepts of the most intelligent. Given the precedents in our current legal system, this will certainly lead to different legal rights and responsibilities for people with different intelligence.

Humans have complex sets of motivations that combine self-interest with compassion for others. But organizations may create super-intelligent minds with simple motives to promote the organizations' interests. Corporations may create minds whose sole value is to increase corporate profits, and governments may create minds whose sole value is to increase political or military power. We already see this issue when corporations behave in ways that would horrify stockholders if they encountered such behavior in their own lives, rather than in distant communities largely invisible to stockholders (Gedicks 2001). One example is the operations by multi-national mining corporations that degrade the environment

near mines and cause the deaths by poisoning of indigenous people. Similar detachment may exist between future human owners and their super-intelligent agents. If these agents play a human-like role in the social contract, but without any of the compassion of human nature, they may be responsible for great harm to humans. Of course, humans often behave with little compassion, and the social contract includes sanctions against such behavior. But our social contract never contemplated the combination of super-intelligence and absolute lack of compassion.

A New Social Contract

The first attempt to modify the social contract for the technology of mind was Asimov's Laws of Robotics, which defined constraints on the behavior of robots (Asimov 1942). But such constraints are inevitably ambiguous, and as Ray Kurzweil has pointed out "there is no purely technical strategy that is workable in this area, because greater intelligence will always find a way to circumvent measures that are the product of a lesser intelligence" (Kurzweil 2005).

However, the technology of mind will give us freedom to choose the motives of machine and enhanced human minds, and we can design them not to want to circumvent measures to protect others. We can design them to value the well-being of other humans. We could try to design a complex formula for measuring human well-being, but the measure that best values human freedom is to allow each person to express their own well-being in their happiness (Pearce 1998; Hibbard 2002), or to be more precise, their expressions of long-term life satisfaction rather than short-term euphoria (Ryan and Deci 2001). This fits naturally in the form of a proposed new social contract: *in exchange for significantly greater than normal human intelligence, a mind must value the long-term life satisfaction of all other intelligent minds*.

Of course, there are many details in how this happiness is defined and how the happiness values of many minds are combined into a single reinforcement value for the super-intelligent mind. For example, the discount rate for future rewards should strongly emphasize the long term to avoid reinforcing behaviors that pander to short-term pleasure at the expense of long-term unhappiness.

Valuing human happiness requires recognizing humans and their emotions. Super-intelligent minds will need to learn to classify humans and other minds, and their emotions, according to the general consensus of the way humans classify. Because humans and other minds will evolve under the technologies of mind and life, super-intelligent minds will need to continually relearn their classifications, reinforced by agreement with the general human consensus.

In order to promote equality, unhappiness should be weighted more heavily than happiness so efforts are focused on helping unhappy minds rather than those who are already happy. We see such behavior among parents who focus their energy on the children who need it most, and in modern societies that provide special services to people in need. This principle will avoid the *tyranny of the majority*, in which a majority of people want the oppression of a minority.

Super-intelligent minds will need values for predictive accuracy to reinforce learning of simulation models and may seek to improve their predictive accuracy by taking needed resources away from others. In order to avoid this, super-intelligent minds should increase their computing resources only when their

simulation models say that the resulting increase in predictive accuracy will produce a net increase in the happiness of others.

Humans who are inconsolably unhappy due to physical or mental illness, pose a complex set of issues for reinforcement values. Values for happiness might motivate super-intelligent minds to cause the deaths of such people. This could be avoided by accounting for dead people at a maximally unhappy value, so that super-intelligent minds always see death as an event to avoid.

These details of engineering super-intelligent minds to protect human values will be tricky, leading some to suggest avoiding the technologies of life and mind. For example, Fukuyama warns against changes to human nature and considers the entire transhumanist program to be dangerous (Fukuyama, 2002). Like most significant ideas, transhumanism is dangerous, but it is not politically feasible to stop the progress of technology. The human drives for improved health, longer lives, reduced labor, and improved minds and bodies are too strong to deny. The technologies of mind and life are inevitable and will change human nature. Our best course is a new social contract to regulate these changes, rather than a futile effort to try to stop them or simply allowing each individual to make whatever changes they like.

Benefits of New Technologies

The technologies of mind and life are really extensions of the ways that most people are already trying to improve their lives. We exercise, regulate our diets, and spend enormous sums on medical care, to improve our health and live longer. We read, study, solve puzzles and play games in order to improve our minds. We look for easier ways to accomplish our tasks, and save for retirement, in order to reduce our labor.

With the right social contract, the new technologies will free us from the need to commute to jobs where we spend most of our days. Instead we will spend our time with family and friends, and in the company of new minds that tell funnier jokes, produce better music and movies, cook better food, know more science, help us find our own creativity, and love us better than any ordinary human can. The new technologies will enable us to live in perpetual good health, satisfying our curiosity about everything in the world, developing deep loving relations with others, and travelling to the reaches of the universe. Our greatly expanded mental abilities will enable us to know more people and to have a much deeper understanding of the world. Assuming upgraded humans will value accuracy of prediction, they will still be curious about the world. Under the proposed new social contract, the xenophobia that causes so much misery will be absent in artificial minds and upgraded humans, and in order to promote general happiness they will counsel the remaining natural humans to resist their own xenophobic urges.

Once people realize that all these things are possible, for themselves or for their children, they will know that they must let nothing prevent this future. On a personal level, this realization is a motive for people to preserve their health until the new technologies arrive. On a social level, it is a motive to create a new social contract that will enable everyone to benefit from them. The benefits are clearly worth the risks that Fukuyama and others have described.

Politics

Any change to the social contract to regulate the technology of mind will have to be worked out in a political process. The good news is that the most technically advanced countries, where the technology will be developed, are democracies. Democracy works best for issues on which the public is informed and interested. A challenge for transhumanists is to educate the public and to help create a movement for promoting and regulating the technology of mind, similar to the political movements for protecting the environment and consumer safety, and for controlling weapons of mass destruction.

The technologies of mind and life promise great benefits, so we must avoid the temptation to ban them altogether (which is undoubtedly technically and politically impossible in the long run). We must also avoid provoking a libertarian backlash that would give everyone total freedom to experiment with these technologies. We would not consider granting such freedom to experiment with nuclear, chemical and biological weapons, and the technologies of mind and life are ultimately much more dangerous than such weapons. We should seek the middle path of regulated development, between a ban and total freedom (Hughes 2004).

Social Security: The debate over the US Social Security system (and similar debates in other countries) is based on 75-year economic projections, and thus brings the issues of the technologies of mind and life into current politics. The Social Security Trustees estimate that the annual rate of growth in US productivity, which has averaged 2.9% over the past 10 years (Bureau of Labor Statistics 2006), will decrease to 1.7% by 2013 and then remain at that level until 2080 (Social Security Administration 2006). Under this estimate, the Social Security Trust Fund will be depleted by 2042 (increasing the estimate of productivity growth delays the date at which the Trust Fund is depleted, possibly forever). One suggested solution to the problem of Trust Fund depletion is to gradually raise the retirement age.

However, during the next 75 years, intelligent machines are likely to gradually out-compete humans for all jobs (Moravec 1999). In order to meet people's needs as unemployment approaches 100 per cent, Moravec suggests gradually reducing the retirement age down to birth in the US Social Security system and other national pension plans. By enabling more goods and services to be produced by fewer human workers, these machines will increase productivity until it is essentially infinite. (For a detailed discussion of the economic effects of intelligent machines, see Marshall Brain's Robotic Nation essays (Brain 2003)). There is a clear conflict between the future envisioned by the Social Security Trustees, in which productivity growth declines and the retirement age increases, and the future envisioned by Moravec and other Artificial Intelligence (AI) researchers, in which productivity growth increases and the retirement age decreases.

Social Security privatization is a proposal to end the current practice of using the contributions of one person to subsidize the benefits of others, and instead to isolate each person's contributions and benefits in a private account. This is exactly the opposite of what will be needed as intelligent machines create great wealth but displace the population from the work force. There must be a mechanism for sharing that wealth, on an international scale.

Intelligent Weapons: The Future Combat Systems project, to automate and network the battlefield, estimated to cost $127 billion, is the largest in US military history. The Department of Defense is the largest funding source for AI research. Other countries have similar, although smaller, projects. These raise immediate ethical questions about machines making life-and-death decisions in battle (Johnson 2005). Longer term, intelligent weapons will enable a small group to rule without the need for the cooperation of citizen soldiers. These issues should force a public debate.

The public movement to regulate intelligent machines can point to treaties banning chemical and biological weapons as precedents for banning intelligent weapons. Weapons more intelligent than humans should be seen as analogous to these threats, and the movement to control such intelligent weapons should learn from the successes and failures of those earlier efforts.

Protecting Children: A great deal of politics is motivated by people's desire to protect their children, reflected in what are often called "values issues." It is natural that many people are frightened of the technologies of mind and life, and worry about the nightmare world their children may inherit. We need to be sensitive to these concerns and avoid expressing careless attitudes toward the fate of humanity or of those who are not on the cutting edge of technology.

In fact, most people understand that the advance of technology cannot be stopped and are unlikely to be drawn into a serious movement to ban the technologies of mind and life. By acknowledging people's concerns for their children's futures, we can lead those concerns into a progressive movement to develop such technologies in ways that promise a better life to all humans through a new social contract.

Conclusions

The technologies of mind and life are coming and will bring enormous benefits. They will also change assumptions that underlie the social contract that governs interactions among humans, with potentially undesirable consequences. Some advocate banning these technologies but, given the promised benefits, this is politically impossible. Instead, we should support regulated development to ensure that all humans benefit from these technologies, rather than a small group benefiting at the expense of most other humans. In particular, minds should be permitted to have significantly greater than natural human intelligence only if they value the long-term life satisfaction of all other intelligent minds.

Achieving the goal of regulated development will require a popular political movement. This will depend on educating the public about these technologies, and about their potential dangers and promised benefits to future generations.

References

Asimov, I. 1942. Runaround. *Astounding Science Fiction*, March.

Atkins v. Virginia (00-8452) 536 U.S. 304 (2002) 260 Va. 375, 534 S. E. 2d 312, reversed and remanded.

Barkow, J. H., L. Cosmides, and J. Tooby. 1992. *The Adapted Mind*. New York: Oxford University Press.

Baum, E. 2004. *What is Thought?* Cambridge, MA: MIT Press.

Bownds, M. 1999. *Biology of Mind*. Bethesda, MD: Fitzgerald Science Press.

Brain, M. 2003. Robotic nation. URL http://marshallbrain.com/robotic-nation.htm

Brown, J., Bullock, D., and Grossberg, S. 1999. How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience* 19(23): 10502-10511.

Bureau of Labor Statistics, 2006. Nonfarm Business Output per Hour. URL http://data.bls.gov/servlet/SurveyOutputServlet?series_id=PRS85006092&data_tool="EaG"

Fukuyama, F. 2002. *Our Posthuman Future: Consequences of the Biotechnology Revolution*. New York: Farrar, Straus, and Giroux.

Gedicks, A. 2001. *Resource Rebels: Native Challenges to Mining and Oil Corporations*. Cambridge, MA: South End Press.

Goldstein, A. S. 1967. *The Insanity Defense*. New Haven: Yale University Press.

Hibbard, B. 2002. *Super-Intelligent Machines.* New York: Kluwer Academic / Plenum Publishers.

Hobbes, T. 1651. *Leviathan*. URL http://www.gutenberg.org/etext/3207

Hughes, J. 2004. *Citizen Cyborg: Why Democratic Societies Must Respond to the Redesigned Human of the Future*. Boulder: Westview Press.

Johnson, G. 2005. Who do you trust more: G.I. Joe or A.I. Joe? *New York Times*, February 20.

Kevles, D. 1985. *In the Name of Eugenics: Genetics and the Uses of Human Heredity*. New York: Knopf.

Kurzweil, R. 1999. *The Age of Spiritual Machines*. New York: Penguin.

Kurzweil, R. 2005. *The Singularity is Near: When Humans Transcend Biology*. New York: Penguin.

Legg, S. and Hutter, M. 2006. A formal measure of machine intelligence. In *Proc. Annual machine learning conference of Belgium and The Netherlands (Benelearn-2006)*. URL http://www.idsia.ch/idsiareport/IDSIA-10-06.pdf

Moravec, H. 1999. *Robot: Mere Machine to Transcendent Mind*. New York and Oxford: Oxford University Press.

Pearce, D. 1998. *The Hedonistic Imperative*. URL http://www.hedweb.com/hedethic/tabconhi.htm

Ryan, R.M. and Deci, E.L. 2001. On happiness and human potentials: A review of research on hedonic and eudiamonic well-being. *Annual Review of Psychology* 52: 141-166.

Seymour, B., O'Doherty, J., Dayan, P., Koltzenburg, M., Jones, A., Dolan, R., Friston, K., and Frackowiak, R. 2004. Temporal difference models describe higher-order learning in humans. *Nature* 429: 664-667.

Social Security Administration, 2006. 2006 OASDI Trustees Report. URL http://www.ssa.gov/OACT/TR/TR06/trTOC.html