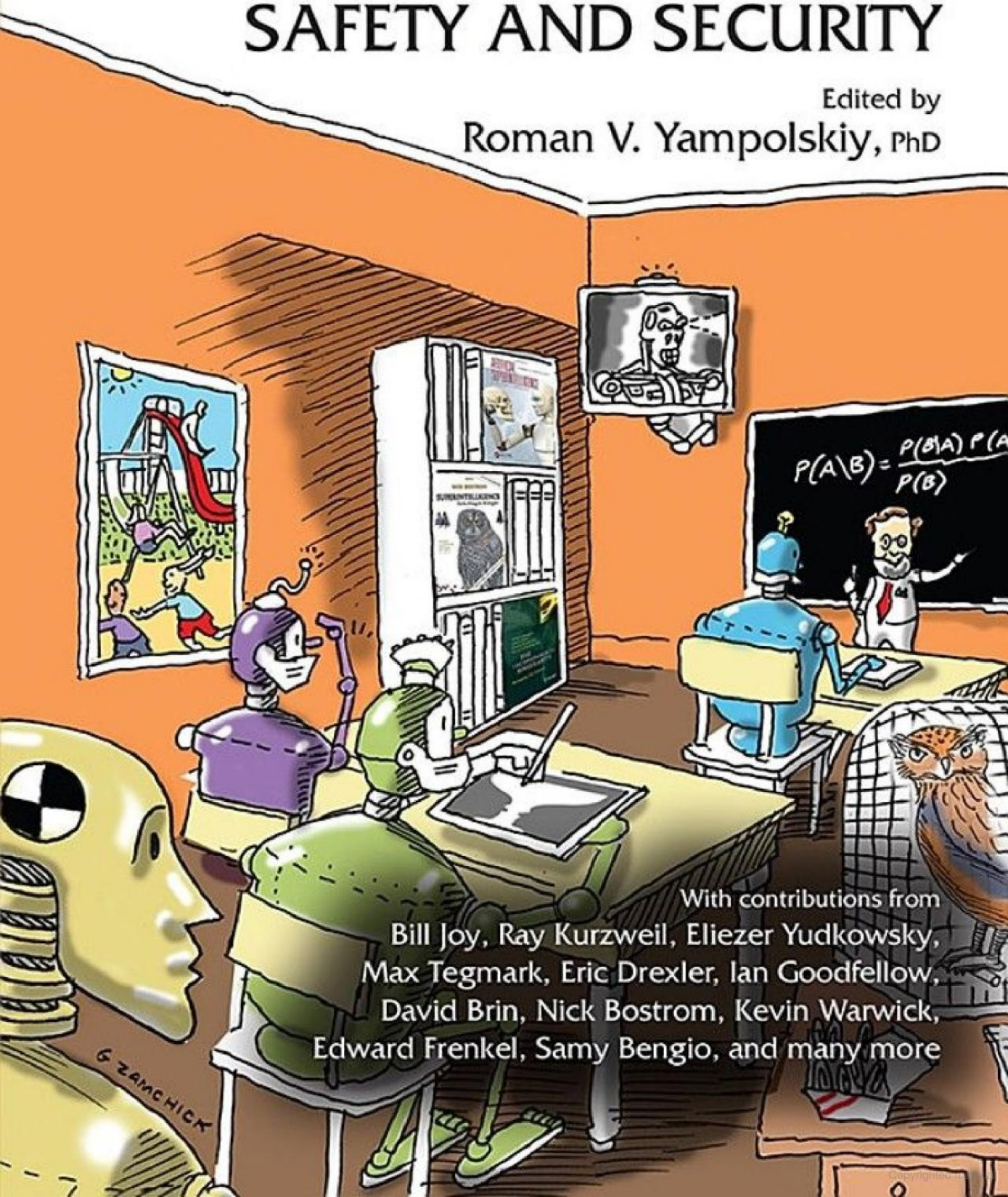
Chapman & Hall/CRC
Artificial Intelligence and Robotics Series

ARTIFICIAL INTELLIGENCE SAFETY AND SECURITY

Edited by
Roman V. Yampolskiy, PhD



With contributions from
Bill Joy, Ray Kurzweil, Eliezer Yudkowsky,
Max Tegmark, Eric Drexler, Ian Goodfellow,
David Brin, Nick Bostrom, Kevin Warwick,
Edward Frenkel, Samy Bengio, and many more

Contents

Preface: Introduction to AI Safety and Security

Roman V. Yampolskiy

About 10,000 scientists* around the world work on different aspects of creating intelligent machines, with the main goal of making such machines as capable as possible. With amazing progress made in the field of AI over the last decade, it is more important than ever to make sure that the technology we are developing has a beneficial impact on humanity. With the appearance of robotic financial advisors, self-driving cars and personal digital assistants come many unresolved problems. We have already experienced market crashes caused by intelligent trading software,† accidents caused by self-driving cars‡ and embarrassment from chatbots,§ which turned racist and engaged in hate speech. I predict that both the frequency and seriousness of such events will steadily increase as AIs become more capable. The failures of today's narrow domain AIs are just a warning: once we develop artificial general intelligence (AGI) capable of cross-domain performance, hurt feelings will be the least of our concerns.

In a recent publication, I proposed a taxonomy of pathways to dangerous AI [1], which was motivated as follows: "In order to properly handle a potentially dangerous artificially intelligent system it is important to understand how the system came to be in such a state. In popular culture (science fiction movies/books) AIs/Robots became self-aware and as a result, rebel against humanity and decide to destroy it. While it is one possible scenario, it is probably the least likely path to the appearance of dangerous AI." I suggested that much more likely reasons include deliberate actions of not-so-ethical people ("on purpose"), side effects of poor design ("engineering mistakes") and finally miscellaneous cases related to the impact of the surroundings of the system ("environment"). Because purposeful design of dangerous AI is just as likely to include all other types of safety problems and will probably have the direst consequences, the most dangerous type of AI and the one most difficult to defend against is an AI made malevolent on purpose.

A follow-up paper [2] explored how a Malevolent AI could be constructed and why it is important to study and understand malicious intelligent software. An AI researcher studying Malevolent AI is like a medical doctor studying how different diseases are transmitted, how new diseases arise, and how they impact the patient's organism. The goal is not to spread diseases, but to learn how to fight them. The authors observe that cybersecurity research involves publishing papers about

malicious exploits as much as publishing information on how to design tools to protect cyber-infrastructure. It is this information exchange between hackers and security experts that results in a well-balanced cyber-ecosystem. In the domain of AI safety engineering, hundreds of papers [3] have been published on different proposals geared at the creation of a safe machine, yet nothing else has been published on how to design a malevolent machine. The availability of such information would be of great value particularly to computer scientists, mathematicians, and others who have an interest in making safe AI, and who are attempting to avoid the spontaneous emergence or the deliberate creation of a dangerous AI, which can negatively affect human activities and in the worst case cause the complete obliteration of the human species. The paper implied that, if an AI safety mechanism is not designed to resist attacks by malevolent human actors, it cannot be considered a functional safety mechanism!

AI FAILURES

Those who cannot learn from history are doomed to repeat it. Unfortunately, very few papers have been published on failures and errors made in development of intelligent systems [4]. The importance of learning from "What Went Wrong and Why" has been recognized by the AI community [5,6]. Such research includes study of how, why and when failures happen [5,6] and how to improve future AI systems based on such information [7,8].

Signatures have been faked, locks have been picked, supermax prisons have had escapes, guarded leaders have been assassinated, bank vaults have been cleaned out, laws have been bypassed, fraud has been committed against our voting process, police officers have been bribed, judges have been blackmailed, forgeries have been falsely authenticated, money has been counterfeited, passwords have been brute-forced, networks have been penetrated, computers have been hacked, biometric systems have been spoofed, credit cards have been cloned, cryptocurrencies have been double spent, airplanes have been hijacked, CAPTCHAs have been cracked, cryptographic protocols have been broken, and even academic peer review has been bypassed with tragic consequences. Millennia long history of humanity contains millions of examples of attempts to develop technological and logistical solutions to increase safety and security, yet not a single example exists which has not eventually failed.

Accidents, including deadly ones, caused by software or industrial robots can be traced to the early days of such technology,* but they are not a direct consequence of the particulars of intelligence available in such systems. AI failures, on the other hand, are directly related to the mistakes produced by the intelligence such systems are designed to exhibit. I can broadly classify such failures into mistakes during the learning phase and mistakes during performance phase. The system can fail to learn

what its human designers want it to learn and instead learn a different, but correlated function. A frequently cited example is a computer vision system which was supposed to classify pictures of tanks but instead learned to distinguish backgrounds of such images [9]. Other examples† include problems caused by poorly designed utility functions rewarding only partially desirable behaviors of agents, such as riding a bicycle in circles around the target [10], pausing a game to avoid losing [11], or repeatedly touching a soccer ball to get credit for possession [12]. During the performance phase, the system may succumb to a number of causes [1,13,14] all leading to an AI failure.

Media reports are full of examples of AI failure but most of these examples can be attributed to other causes on closer examination, such as bugs in code or mistakes in design. The list below is curated to only mention failures of intended intelligence. Additionally, the examples below include only the first occurrence of a particular failure, but the same problems are frequently observed again in later years. Finally, the list does not include AI failures due to hacking or other intentional causes. Still, the timeline of AI failures has an exponential trend while implicitly indicating historical events such as "AI Winter":

1958 Advice software deduced inconsistent sentences using logical programming [15].
1959 AI designed to be a General Problem Solver failed to solve real-world problems.‡
1977 Story writing software with limited common sense produced "wrong" stories [16].
1982 Software designed to make discoveries, discovered how to cheat instead.§
1983 Nuclear attack early warning system falsely claimed that an attack is taking place.¶
1984 The National Resident Match program was biased in placement of married couples [17].
1988 Admissions software discriminated against women and minorities [18].
1994 Agents learned to "walk" quickly by becoming taller and falling over [19].
2005 Personal assistant AI rescheduled a meeting 50 times, each time by 5 minutes [20].
2006 Insider threat detection system classified normal activities as outliers [21].
2006 Investment advising software was losing money in real trading [22].
2007 Google search engine returned unrelated results for some keywords.*
2010 Complex AI stock trading software caused a trillion dollar flash crash.†
2011 E-Assistant told to "call me an ambulance" began to refer to the user as Ambulance.‡
2013 Object recognition neural networks saw phantom objects in particular noise

images [23].

2013 Google software engaged in name-based discrimination in online ad delivery [24].

2014 Search engine autocomplete made bigoted associations about groups of users [25].

2014 Smart fire alarm failed to sound alarm during fire.§

2015 Automated email reply generator created inappropriate responses.¶

2015 A robot for grabbing auto parts grabbed and killed a man.**

2015 Image tagging software classified black people as gorillas.††

2015 Medical expert AI classified patients with asthma as lower risk [26].

2015 Adult content filtering software failed to remove inappropriate content.‡‡

2015 Amazon's Echo responded to commands from TV voices.§§

2016 Linkedin's name lookup suggests male names in place of female ones.¶¶

2016 AI designed to predict recidivism acted racist.***

2016 AI agent exploited reward signal to win without completing the game course.†††

2016 Passport picture checking system flagged Asian user as having closed eyes.‡‡‡

2016 Game NPCs designed unauthorized superweapons.§§§

2016 AI judged a beauty contest and rated dark-skinned contestants lower.¶¶¶

2016 Smart contract permitted syphoning of funds from the DAO.****

2016 Patrol robot collided with a child.††††

2016 World champion-level Go playing AI lost a game.‡‡‡‡

2016 Self-driving car had a deadly accident.§§§§

2016 AI designed to converse with users on Twitter became verbally abusive.¶¶¶¶

2016 Google image search returned racists results.*****

2016 Artificial applicant failed to pass university entrance exam.†††††

2016 Predictive policing system disproportionately targeted minority neighborhoods.*

2016 Text subject classifier failed to learn relevant features for topic assignment [27].

2017 AI for making inspirational quotes failed to inspire with gems like "Keep Panicking".†

2017 Alexa played adult content instead of song for kids.‡

2017 Cellphone case designing AI utilized inappropriate images.§

2017 Pattern recognition software failed to recognize certain types of inputs.¶

2017 Debt recovery system miscalculated amounts owed.**
2017 Russian language chatbot shared pro-Stalinist, pro-abuse and pro-suicide views.††
2017 Translation AI learned to stereotype careers to specific genders [28].
2017 Face beautifying AI made black people look white.‡‡
2017 Google's sentiment analyzer became homophobic and anti-Semitic.§§
2017 Fish recognition program learned to recognize boat IDs instead.¶¶
2017 Billing software sent an electrical bill for 284 billion dollars.***
2017 Alexa turned on loud music at night without being prompted to do so.†††
2017 AI for writing Christmas carols produced nonsense.‡‡‡
2017 Apple's face recognition system failed to distinguish Asian users.§§§
2017 Facebook's translation software changed Yampolskiy to Polanski, see Figure I.1.
2018 Google Assistant created bizarre merged photo.¶¶¶
2018 Robot store assistant was not helpful with responses like "cheese is in the fridges."****

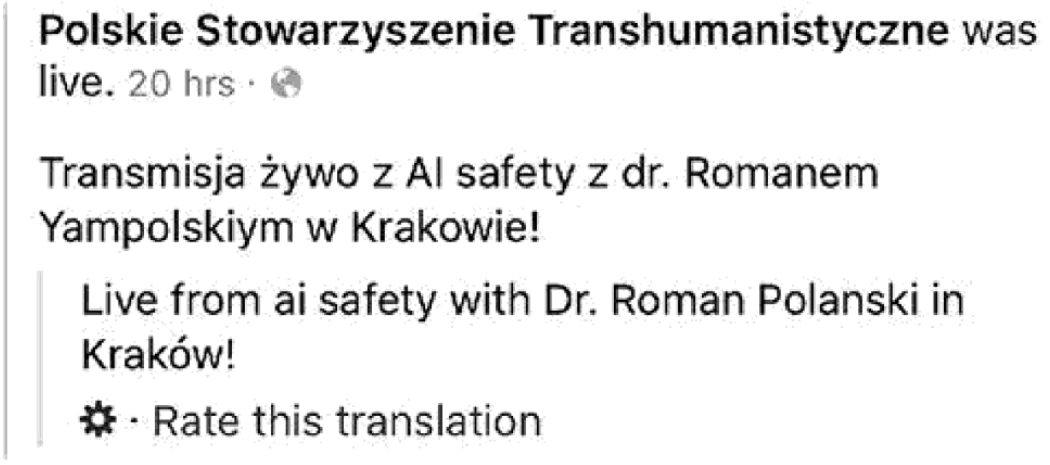


FIGURE I.1 While translating from Polish to English Facebook's software changed Roman V. "Yampolskiy" to Roman "Polanski" due to statistically higher frequency of the latter name in sample texts.

Spam filters block important emails, GPS provides faulty directions, machine translation corrupts the meaning of phrases, autocorrect replaces a desired word with a wrong one, biometric systems misrecognize people, software fails to capture what is being said; overall, it is harder to find examples of AIs that don't fail. Depending on what we consider for inclusion as examples of problems with intelligent software, the list of examples could be grown almost indefinitely. In its most extreme interpretation, any software with as much as an "if statement" can be considered a form of narrow artificial intelligence (NAI) and all its bugs are thus

examples of AI failure.*

Analyzing the list of narrow AI failures, from the inception of the field to modern-day systems, we can arrive at a simple generalization: An AI designed to do X will eventually fail to do X. While it may seem trivial, it is a powerful generalization tool, which can be used to predict future failures of NAIs. For example, looking at cutting-edge current and future AIs we can predict that:

- Software for generating jokes will occasionally fail to make them funny.
- Sex robots will fail to deliver an orgasm or to stop at the right time.
- Sarcasm detection software will confuse sarcastic and sincere statements.
- Video description software will misunderstand movie plots.
- Software-generated virtual worlds may not be compelling.
- AI doctors will misdiagnose some patients in a way a real doctor would not.
- Employee screening software will be systematically biased and thus hire low performers.
- The Mars robot explorer will misjudge its environment and fall into a crater.
- And so on.

Others have given the following examples of possible accidents with A(G)I/superintelligence:

- Housekeeping robot cooks family pet for dinner.†
- A mathematician AGI converts all matter into computing elements to solve problems.‡
- An AGI running simulations of humanity creates conscious beings who suffer [29].
- Paperclip manufacturing AGI fails to stop and converts universe into raw materials [30].
- A scientist AGI performs experiments with significant negative impact on biosphere [31].
- Drug design AGI develops time-delayed poison to kill everyone and so defeat cancer.§
- Future superintelligence optimizes away all consciousness.¶
- AGI kills humanity and converts universe into materials for improved penmanship.**
- AGI designed to maximize human happiness tiles universe with tiny smiley faces [32].
- AGI instructed to maximize pleasure consigns humanity to a dopamine drip [33].
- Superintelligence may rewire human brains to increase their perceived satisfaction [32].

Denning and Denning made some similar error extrapolations in their humorous paper on "artificial stupidity" [34]: "Soon the automated DEA started closing down pharmaceutical companies saying they were dealing drugs. The automated FTC closed down the Hormel Meat Company, saying it was purveying spam. The automated DOJ shipped Microsoft 500,000 pinstriped pants and jackets, saying it was filing suits. The automated Army replaced all its troops with a single robot, saying it had achieved the Army of One. The automated Navy, in a cost saving move, placed its largest-ever order for submarines with Subway Sandwiches. The FCC issued an order for all communications to be wireless, causing thousands of AT&T installer robots to pull cables from overhead poles and underground conduits. The automated TSA flew its own explosives on jetliners, citing data that the probability of two bombs on an airplane is exceedingly small."

AGI can be seen as a superset of all NAIs and so will exhibit a superset of failures as well as more complicated failures resulting from the combination of failures of individual NAIs and new super-failures, possibly resulting in an existential threat to humanity or at least an AGI takeover. In other words, AGIs can make mistakes influencing everything. Overall, I predict that AI failures and premeditated malevolent AI incidents will increase in frequency and severity proportionate to AIs' capability.

PREVENTING AI FAILURES

AI failures have a number of causes, with the most common ones currently observed displaying some type of algorithmic bias, poor performance, or basic malfunction. Future AI failures are likely to be more severe including purposeful manipulation/deception [35], or even resulting in human death (likely from misapplication of militarized AI/autonomous weapons/killer robots [36]). At the very end of the severity scale, we see existential risk scenarios resulting in the extermination of human kind or suffering-risk scenarios [37] resulting in the large-scale torture of humanity, both types of risk coming from supercapable artificially intelligent systems.

Reviewing examples of AI accidents, we can notice patterns of failure, which can be attributed to the following causes:

- Biased data, including cultural differences
- Deploying underperforming system
- Non-representative training data
- Discrepancy between training and testing data
- Rule overgeneralization or application of population statistics to individuals
- Inability to handle noise or statistical outliers
- Not testing for rare or extreme conditions

- Not realizing an alternative solution method can produce same results, but with side effects
- Letting users control data or learning process
- No security mechanism to prevent adversarial meddling
- No cultural competence/common sense
- Limited access to information/sensors
- Mistake in design and inadequate testing
- Limited ability for language disambiguation
- Inability to adapt to changes in the environment

With bias being the most common current cause of failure, it is helpful to analyze particular types of algorithmic bias. Friedman and Nissenbaum [17] proposed the following framework for analyzing bias in computer systems. They subdivided causes of bias into three categories—preexisting bias, technical bias, and emergent bias.

- **Preexisting bias** reflects bias in society and social institutions, practices, and attitudes. The system simply preserves an existing state in the world and automates application of bias as it currently exists.
- **Technical bias** appears because of hardware or software limitations of the system itself.
- **Emergent bias** emerges after the system is deployed due to changing societal standards.

Many of the observed AI failures are similar to mishaps experienced by little children. This is particularly true for artificial neural networks, which are at the cutting edge of machine learning (ML). One can say that children are untrained neural networks deployed on real data and observing them can teach us a lot about predicting and preventing AI failures. A number of research groups [31,38] have investigated types of ML failure and here I have summarized their work and mapped it onto similar situations with children:

- Negative side effects—child makes a mess
- Reward hacking—child finds candy jar
- Scalable oversight—babysitting should not require a team of 10
- Safe exploration—no fingers in the outlet
- Robustness to distributional shift—use “inside voice” in the classroom
- Inductive ambiguity identification—is ant a cat or a dog?
- Robust human imitation—daughter shaves like daddy
- Informed oversight—let me see your homework
- Generalizable environmental goals—ignore that mirage
- Conservative concepts—that dog has no tail

- Impact measures—keep a low profile
- Mild optimization—do not be a perfectionist
- Averting instrumental incentives—be an altruist

The majority of research currently taking place to prevent such failures is currently happening under the label of "AI Safety."

AI SAFETY

In 2010, I coined the phrase "Artificial Intelligence Safety Engineering" and its shorthand notation "AI Safety" to give a name to a new direction of research I was advocating. I formally presented my ideas on AI safety at a peer-reviewed conference in 2011 [39], with subsequent publications on the topic in 2012 [40], 2013 [41,42], 2014 [43], 2015 [44], 2016 [1,13], 2017 [45], and 2018 [46,47]. It is possible that someone used the phrase informally before, but to the best of my knowledge, I was the first to use it* in a peer-reviewed publication and to bring its popularity. Before that, the most common names for the field of machine control were "Machine Ethics" [48] or "Friendly AI" [49]. Today the term "AI Safety" appears to be the accepted[†,‡,§,¶,⁎⁎,††,‡‡,§§,¶¶,⁎⁎⁎,†††,‡‡‡] name for the field used by a majority of top researchers [38]. The field itself is becoming mainstream despite being regarded as either science fiction or pseudoscience in its early days.

Our legal system is behind our technological abilities and the field of AI safety is in its infancy. The problem of controlling intelligent machines is just now being recognized[§§§] as a serious concern and many researchers are still skeptical about its very premise. Worse yet, only about 100 people around the world are fully emerged in working on addressing the current limitations in our understanding and abilities in this domain. Only about a dozen[¶¶¶] of those have formal training in computer science, cybersecurity, cryptography, decision theory, machine learning, formal verification, computer forensics, steganography, ethics, mathematics, network security, psychology, and other relevant fields. It is not hard to see that the problem of making a safe and capable machine is much greater than the problem of making just a capable machine. Yet only about 1% of researchers are currently engaged in that problem with available funding levels below even that mark. As a relatively young and underfunded field of study, AI safety can benefit from adopting methods and ideas from more established fields of science. Attempts have been made to introduce techniques, which were first developed by cybersecurity experts to secure software systems to this new domain of securing intelligent machines [50–53]. Other fields, which could serve as a source of important techniques, would include software engineering and software verification.

During software development, iterative testing and debugging is of fundamental importance to produce reliable and safe code. While it is assumed that all

complicated software will have some bugs, with many advanced techniques available in the toolkit of software engineers, most serious errors could be detected and fixed, resulting in a product suitable for its intended purposes. Certainly, a lot of modular development and testing techniques employed by the software industry can be utilized during development of intelligent agents, but methods for testing a completed software package are unlikely to be transferable in the same way. Alpha and beta testing, which work by releasing almost-finished software to advanced users for reporting problems encountered in realistic situations, would not be a good idea in the domain of testing/debugging superintelligent software. Similarly simply running the software to see how it performs is not a feasible approach with superintelligent agent.

CYBERSECURITY vs. AI SAFETY

Bruce Schneier has said, "If you think technology can solve your security problems then you don't understand the problems and you don't understand the technology." Salman Rushdie made a more general statement: "There is no such thing as perfect security, only varying levels of insecurity." I propose what I call the Fundamental Theorem of Security—Every security system will eventually fail; there is no such thing as a 100% secure system. If your security system has not failed, just wait longer.

In theoretical computer science, a common way of isolating the essence of a difficult problem is via the method of reduction to another, sometimes better analyzed, problem [54–56]. If such a reduction is a possibility and is computationally efficient [57], such a reduction implies that if the better analyzed problem is somehow solved, it would also provide a working solution for the problem we are currently dealing with. The problem of AGI Safety could be reduced to the problem of making sure a particular human is safe. I call this the Safe Human Problem (SHP).* Formally such a reduction can be done via a restricted Turing test in the domain of safety in a manner identical to how AI-completeness of a problem could be established [55,58]. Such formalism is beyond the scope of this preface so I simply point out that in both cases, we have at least a human-level intelligent agent capable of influencing its environment, and we would like to make sure that the agent is safe and controllable. While in practice changing the design of a human via DNA manipulation is not as simple as changing the source code of an AI, theoretically, it is just as possible.

It is observed that humans are not safe to themselves and others. Despite a millennia of attempts to develop safe humans via culture, education, laws, ethics, punishment, reward, religion, relationships, family, oaths, love and even eugenics, success is not within reach. Humans kill and commit suicide, lie and betray, steal and cheat, usually in proportion to how much they can get away with. Truly

powerful dictators will enslave, commit genocide, break law and violate human rights. It is famously stated that a human without a sin can't be found. The best we can hope for is to reduce such unsafe tendencies to levels that our society can survive. Even with advanced genetic engineering [59], the best we can hope for is some additional reduction in how unsafe humans are. As long as we permit a person to have choices (free will), they can be bribed, they will deceive, they will prioritize their interests above those they are instructed to serve and they will remain fundamentally unsafe. Despite being trivial examples of a solution to the Value Learning Problem (VLP) [60–62], human beings are anything but safe, bringing into question our current hope that solving VLP will get us to safe AI. This is important. To quote Bruce Schneier, "Only amateurs attack machines; professionals target people." Consequently, I see AI safety research as, at least partially, an adversarial field similar to cryptography or security.*

If a cybersecurity system fails, the damage is unpleasant but tolerable in most cases: someone loses money or someone loses privacy. For narrow AIs, safety failures are at the same level of importance as in general cybersecurity, but for AGI it is fundamentally different. A single failure of a superintelligent system may cause an existential risk event. If an AGI safety mechanism fails, everyone may lose everything, and all biological life in the universe is potentially destroyed. With cybersecurity systems, you will get another chance to get it right or at least do better. With AGI safety system, you only have one chance to succeed, so learning from failure is not an option. Worse, a typical security system is likely to fail to a certain degree, e.g. perhaps only a small amount of data will be compromised. With an AGI safety system, failure or success is a binary option: either you have a safe and controlled superintelligence or you don't. The goal of cybersecurity is to reduce the number of successful attacks on the system; the goal of AI safety is to make sure zero attacks succeed in bypassing the safety mechanisms. For that reason, ability to segregate NAI projects from potentially AGI projects is an open problem of fundamental importance in the AI safety field.

The problems are many. We have no way to monitor, visualize or analyze the performance of superintelligent agents. More trivially, we don't even know what to expect after such a software starts running. Should we see immediate changes to our environment? Should we see nothing? What is the timescale on which we should be able to detect something? Will it be too quick to notice or are we too slow to realize something is happening? Will the impact be locally observable or impact distant parts of the world? How does one perform standard testing? On what data sets? What constitutes an "Edge Case" for general intelligence? The questions are many, but the answers currently don't exist. Additional complications will come from the interaction between intelligent software and safety mechanisms designed to keep AI safe and secure. We will also have to somehow test all the AI safety mechanisms

currently in development. While AI is at human levels, some testing can be done with a human agent playing the role of the artificial agent. At levels beyond human capacity, adversarial testing does not seem to be realizable with today's technology. More significantly, only one test run would ever be possible.

CONCLUSIONS

The history of robotics and artificial intelligence in many ways is also the history of humanity's attempts to control such technologies. From the Golem of Prague to the military robots of modernity, the debate continues as to what degree of independence such entities should have and how to make sure that they do not turn on us, its inventors. Numerous recent advancements in all aspects of research, development and deployment of intelligent systems are well publicized, but safety and security issues related to AI are rarely addressed. The book you are reading aims to mitigate this fundamental problem as a first multi-author volume on this subject, which I hope will be seen as humankind's communal response to the control problem. It is comprised of chapters from leading AI safety researchers addressing different aspects of the AI control problem as they relate to the development of safe and secure artificial intelligence.

Part I of this book, "Concerns of Luminaries," is comprised of 11 previously published seminal papers outlining different sub-domains of concern with regards to the AI Control Problem and includes contributions from leading scholars in a diverse set of fields—philosophers, scientists, writers, and business people, presented in chronological order of original publication. Part II, "Responses of Scholars," is made up of 17 chapters (in alphabetical order, by the last name of the first author) of proposed theoretical and practical solutions to the concerns raised in Part I, as well as introductions of additional concerns, from leading AI safety researchers. The chapters vary in length and technical content from broad interest opinion essays to highly formalized algorithmic approaches to specific problems. All chapters are self-contained and could be read in any order or skipped without a loss of comprehension. This volume is without any doubt not the last word on this subject, but rather one of the first steps in the right direction.

REFERENCES

1. R. V. Yampolskiy, "Taxonomy of Pathways to Dangerous Artificial Intelligence," in *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
2. F. Pistono and R. V. Yampolskiy, "Unethical Research: How to Create a Malevolent Artificial Intelligence," *presented at the 25th International Joint Conference on Artificial Intelligence (IJCAI-16) Ethics for Artificial Intelligence Workshop (AI-Ethics-2016)*, New York, NY, July 9, 2016.
3. K. Sotala and R. V. Yampolskiy, "Responses to catastrophic AGI risk: Asurvey," *Physica Scripta*, vol. 90, 2015.

4. N. Rychtyckyj and A. Turski, “Reasons for Success (and Failure) in the Development and Deployment of AI Systems,” in *AAAI 2008 Workshop on What Went Wrong and Why*, 2008.
5. D. Shapiro and M. H. Goker, “Advancing AI research and applications by learning from what went wrong and why,” *AI Magazine*, vol. 29, pp. 9–10, 2008.
6. A. Abecker, R. Alami, C. Baral, T. Bickmore, E. Durfee, T. Fong et al., “AAAI 2006 spring symposium reports,” *AI Magazine*, vol. 27, p. 107, 2006.
7. C. Marling and D. Chelberg, “RoboCup for the Mechanically, Athletically and Culturally Challenged,” in *What Went Wrong and Why: Lessons from AI Research and Applications: Papers from the 2008 AAAI Workshop*. Menlo Park, California: AAAI Press, 2008.
8. S. Shalev-Shwartz, O. Shamir, and S. Shammah, “Failures of Gradient-Based Deep Learning,” in *International Conference on Machine Learning*, 2017, pp. 3067–3075.
9. E. Yudkowsky, “Artificial intelligence as a positive and negative factor in global risk,” *Global Catastrophic Risks*, vol. 1, p. 303, 2008.
10. J. Randløv and P. Alstrøm, “Learning to Drive a Bicycle Using Reinforcement Learning and Shaping,” in *ICML*, 1998, pp. 463–471.
11. T. Murphy VII, “The First Level of Super Mario Bros. is Easy with Lexicographic Orderings and Time Travel,” in *The Association for Computational Heresy (SIGBOVIK) 2013*, 2013.
12. A. Y. Ng, D. Harada, and S. Russell, “Policy Invariance under Reward Transformations: Theory and Application to Reward Shaping,” in *ICML*, 1999, pp. 278–287.
13. F. Pistono and R. V. Yampolskiy, “Unethical Research: How to Create a Malevolent Artificial Intelligence,” *arXiv preprint arXiv:1605.02817*, 2016.
14. P. Scharre, “Autonomous Weapons and Operational Risk,” *presented at the Center for a New American Society*, Washington DC, 2016.
15. C. Hewitt, “Development of Logic Programming: What went wrong, what was done about it, and what it might mean for the future,” in *What Went Wrong and Why: Lessons from AI Research and Applications: Papers from the 2008 AAAI Workshop*. Menlo Park, California: AAAI Press, 2008.
16. J. R. Meehan, “TALE-SPIN, An Interactive Program that Writes Stories,” in *IJCAI*, 1977, pp. 91–98.
17. B. Friedman and H. Nissenbaum, “Bias in computer systems,” *ACM Transactions on Information Systems (TOIS)*, vol. 14, pp. 330–347, 1996.
18. S. Lowry and G. Macpherson, “A blot on the profession,” *British Medical Journal (Clinical Research Ed.)*, vol. 296, p. 657, 1988.
19. K. Sims, “Evolving Virtual Creatures,” in *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, 1994, pp. 15–22.
20. M. Tambe, “Electric elves: What went wrong and why,” *AI Magazine*, vol. 29, p. 23, 2008.
21. A. Liu, C. E. Martin, T. Hetherington, and S. Matzner, “AI Lessons Learned from Experiments in Insider Threat Detection,” in *AAAI Spring Symposium: What Went Wrong and Why: Lessons from AI Research and Applications*, 2006, pp. 49–55.
22. J. Gunderson and L. Gunderson, “And Then the Phone Rang,” in *AAAI Spring Symposium: What Went Wrong and Why: Lessons from AI Research and Applications*, 2006, pp. 13–18.
23. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow et al., “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.

24. L. Sweeney, “Discrimination in online ad delivery,” *Queue*, vol. 11, p. 10, 2013.
25. N. Diakopoulos, “Algorithmic defamation: The case of the shameless autocomplete,” *Tow Center for Digital Journalism*, 2014.
26. R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, “Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1721–1730.
27. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You?: Explaining the Predictions of Any Classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
28. A. Caliskan, J. J. Bryson, and A. Narayanan, “Semantics derived automatically from language corpora contain human-like biases,” *Science*, vol. 356, pp. 183–186, 2017.
29. S. Armstrong, A. Sandberg, and N. Bostrom, “Thinking inside the box: Controlling and using an oracle ai,” *Minds and Machines*, vol. 22, pp. 299–324, 2012.
30. N. Bostrom, “Ethical issues in advanced artificial intelligence,” *Science Fiction and Philosophy: From Time Travel to Superintelligence*, pp. 277–284, 2003.
31. J. Taylor, E. Yudkowsky, P. LaVictoire, and A. Critch, “Alignment for advanced machine learning systems,” *Machine Intelligence Research Institute*, 2016.
32. E. Yudkowsky, “Complex value systems in friendly AI,” *Artificial General Intelligence*, pp. 388–393, 2011.
33. G. Marcus, “Moral machines,” *The New Yorker*, vol. 24, 2012.
34. D. E. Denning and P. J. Denning, “Artificial stupidity,” *Association for Computing Machinery. Communications of the ACM*, vol. 47, no. 5, p. 112, 2004.
35. M. Chessen, “The MADCOM Future,” Atlantic Council, Available at: http://www.atlanticcouncil.org/publications/reports/the-madcom-future, 2017.
36. A. Krishnan, *Killer Robots: Legality and Ethicality of Autonomous Weapons*. Ashgate Publishing, Ltd., 2009.
37. L. Gloor, “*Suffering-focused AI safety: Why “fail-safe” measures might be our top intervention*,” Technical Report FRI-16-1. Foundational Research Institute. https://foundationalresearch.org/wp-content/uploads/2016/08/Suffering-focused-AI-safety.pdf 2016.
38. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete problems in AI safety,” *arXiv preprint arXiv:1606.06565*, 2016.
39. R. V. Yampolskiy, “Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach,” *presented at the Philosophy and Theory of Artificial Intelligence (PT-AI2011)*, Thessaloniki, Greece, October 3–4, 2011.
40. R. V. Yampolskiy and J. Fox, “Safety engineering for artificial general intelligence,” *Topoi. Special Issue on Machine Ethics & the Ethics of Building Intelligent Machines*, 2012.
41. L. Muehlhauser and R. Yampolskiy, “Roman Yampolskiy on AI Safety Engineering,” *presented at the Machine Intelligence Research Institute*, July 15, 2013, Available at: http://intelligence.org/2013/07/15/roman-interview/.
42. R. V. Yampolskiy, “Artificial intelligence safety engineering: Why machine ethics is a wrong approach,” in *Philosophy and Theory of Artificial Intelligence*, Springer Berlin Heidelberg, 2013, pp. 389–396.
43. A. M. Majot and R. V. Yampolskiy, “AI Safety Engineering through Introduction of Self-Reference into Felicific Calculus via Artificial Pain and Pleasure,” in *IEEE International Symposium on Ethics in Science, Technology and Engineering*, Chicago,

IL, May 23–24, 2014, pp. 1–6.
44. R. V. Yampolskiy, "*Artificial Superintelligence: a Futuristic Approach,"* Chapman and Hall/CRC, 2015.
45. R. V. Yampolskiy, "What are the ultimate limits to computational techniques: verifier theory and unverifiability," *Physica Scripta*, vol. 92, p. 093001, 2017.
46. A. Ramamoorthy and R. Yampolskiy, "Beyond mad?: The race for artificial general intelligence," *ITU Journal: ICT Discoveries*, 2017.
47. M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," arXiv preprint arXiv:1802.07228, 2018.
48. J. H. Moor, "The nature, importance, and difficulty of machine ethics," IEEE Intelligent Systems, vol. 21, pp. 18–21, 2006.
49. E. Yudkowsky, "Creating friendly AI 1.0: The analysis and design of benevolent goal architectures," in Singularity Institute for Artificial Intelligence, San Francisco, CA, June, vol. 15, 2001.
50. R. Yampolskiy, "Leakproofing the singularity artificial intelligence confinement problem," Journal of Consciousness Studies, vol. 19, pp. 1–2, 2012.
51. J. Babcock, J. Kramar, and R. Yampolskiy, "The AGI Containment Problem," arXiv preprint arXiv:1604.00545, 2016.
52. J. Babcock, J. Kramar, and R. Yampolskiy, "The AGI Containment Problem," in The Ninth Conference on Artificial General Intelligence (AGI2015), 2016.
53. S. Armstrong and R. V. Yampolskiy, "Security Solutions for Intelligent and Complex Systems," in Security Solutions for Hyperconnectivity and the Internet of Things, IGI Global, 2016, pp. 37–88.
54. R. M. Karp, "Reducibility Among Combinatorial Problems," in Complexity of Computer Computations, R. E. Miller and J. W. Thatcher, Eds., New York: Plenum, 1972, pp. 85–103.
55. R. Yampolskiy, "Turing Test as a Defining Feature of AI-Completeness," in Artificial Intelligence, Evolutionary Computing and Metaheuristics. vol. 427, X.-S. Yang, Ed., Berlin Heidelberg: Springer, 2013, pp. 3–17.
56. R. V. Yampolskiy, "AI-Complete, AI-Hard, or AI-Easy–Classification of Problems in AI," in The 23rd Midwest Artificial Intelligence and Cognitive Science Conference, Cincinnati, OH, USA, 2012.
57. R. V. Yampolskiy, "Efficiency theory: Aunifying theory for information, computation and intelligence," Journal of Discrete Mathematical Sciences & Cryptography, vol. 16(45), pp. 259–277, 2013.
58. R. V. Yampolskiy, "AI-Complete CAPTCHAs as zero knowledge proofs of access to an artificially intelligent system," ISRN Artificial Intelligence, vol. 271878, 2011.
59. R. V. Yampolskiy, "On the Origin of Samples: Attribution of Output to a Particular Algorithm," arXiv preprint arXiv:1608.06172, 2016.
60. K. Sotala, "Defining Human Values for Value Learners," in 2nd International Workshop on AI, Ethics and Society, AAAI-2016, 2016.
61. D. Dewey, "Learning what to value," Artificial General Intelligence, pp. 309–314, 2011.
62. N. Soares and B. Fallenstein, "*Aligning superintelligence with human interests: A technical research agenda*," *Machine Intelligence Research Institute (MIRI) Technical Report*, vol. 8, 2014.

* https://intelligence.org/2014/01/28/how-big-is-ai/
† https://en.wikipedia.org/wiki/2010_Flash_Crash
‡ https://electrek.co/2016/05/26/tesla-model-s-crash-autopilot-video/
§ https://en.wikipedia.org/wiki/Tay_(bot)
* https://en.wikipedia.org/wiki/Kenji_Urada
† http://lesswrong.com/lw/lvh/examples_of_ais_behaving_badly/
‡ https://en.wikipedia.org/wiki/General_Problem_Solver
§ http://aliciapatterson.org/stories/eurisko-computer-mind-its-own
¶ https://en.wikipedia.org/wiki/1983_Soviet_nuclear_false_alarm_incident
* https://en.wikipedia.org/wiki/Google_bomb
† https://en.wikipedia.org/wiki/2010_Flash_Crash
‡ https://www.technologyreview.com/s/601897/tougher-turing-test-exposes-chatbots-stupidity/
§ https://www.forbes.com/sites/aarontilley/2014/04/03/googles-nest-stops-selling-its-smart-smoke-alarm-for-now
¶ https://gmail.googleblog.com/2015/11/computer-respond-to-this-email.html
** http://time.com/3944181/robot-kills-man-volkswagen-plant/
†† http://www.huffingtonpost.com/2015/07/02/google-black-people-goril_n_7717008.html
‡‡ http://blogs.wsj.com/digits/2015/05/19/googles-youtube-kids-app-criticized-for-inappropriate-content/
§§ https://motherboard.vice.com/en_us/article/53dz8x/people-are-complaining-that-amazon-echo-is-responding-to-ads-on-tv
¶¶ https://www.seattletimes.com/business/microsoft/how-linkedins-search-engine-may-reflect-a-bias
*** http://gawker.com/this-program-that-judges-use-to-predict-future-crimes-s-1778151070
††† https://openai.com/blog/faulty-reward-functions
‡‡‡ http://www.telegraph.co.uk/technology/2016/12/07/robot-passport-checker-rejects-asian-mans-photo-having-eyes
§§§ http://www.kotaku.co.uk/2016/06/03/elites-ai-created-super-weapons-and-started-hunting-players-skynet-is-here
¶¶¶ https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people
**** https://en.wikipedia.org/wiki/The_DAO_(organization)
†††† http://www.latimes.com/local/lanow/la-me-ln-crimefighting-robot-hurts-child-bay-area-20160713-snap-story.html
‡‡‡‡ https://www.engadget.com/2016/03/13/google-alphago-loses-to-human-in-one-match/
§§§§ https://www.theguardian.com/technology/2016/jul/01/tesla-driver-killed-autopilot-self-driving-car-harry-potter
¶¶¶¶ http://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist
***** https://splinternews.com/black-teenagers-vs-white-teenagers-why-googles-algori-1793857436
††††† https://www.japantimes.co.jp/news/2016/11/15/national/ai-robot-fails-get-university-tokyo
* https://www.themarshallproject.org/2016/02/03/policing-the-future
† https://www.buzzworthy.com/ai-tries-to-generate-inspirational-quotes-and-gets-it-hilariously-wrong

‡ https://www.entrepreneur.com/video/287281
§ https://www.boredpanda.com/funny-amazon-ai-designed-phone-cases-fail
¶ http://www.bbc.com/future/story/20170410-how-to-fool-artificial-intelligence
** http://www.abc.net.au/news/2017-04-10/centrelink-debt-recovery-system-lacks-transparency-ombudsman/8430184
†† https://techcrunch.com/2017/10/24/another-ai-chatbot-shown-spouting-offensive-views
‡‡ http://www.gizmodo.co.uk/2017/04/faceapp-blames-ai-for-whitening-up-black-people
§§ https://motherboard.vice.com/en_us/article/j5jmj8/google-artificial-intelligence-bias
¶¶ https://medium.com/@gidishperber/what-ive-learned-from-kaggle-s-fisheries-competition-92342f9ca779
*** https://www.washingtonpost.com/news/business/wp/2017/12/26/woman-gets-284-billion-electric-bill-wonders-whether-its-her-christmas-lights
††† http://mashable.com/2017/11/08/amazon-alexa-rave-party-germany
‡‡‡ http://mashable.com/2017/12/22/ai-tried-to-write-christmas-carols
§§§ http://www.mirror.co.uk/tech/apple-accused-racism-after-face-11735152
¶¶¶ https://qz.com/1188170/google-photos-tried-to-fix-this-ski-photo
**** http://www.iflscience.com/technology/store-hires-robot-to-help-out-customers-robot-gets-fired-for-scaring-customers-away
* https://en.wikipedia.org/wiki/List_of_software_bugs
† https://www.theguardian.com/sustainable-business/2015/jun/23/the-ethics-of-ai-how-to-stop-your-robot-cooking-your-cat
‡ https://intelligence.org/2014/11/18/misconceptions-edge-orgs-conversation-myth-ai
§ https://80000hours.org/problem-profiles/positively-shaping-artificial-intelligence
¶ http://slatestarcodex.com/2014/07/13/growing-children-for-bostroms-disneyland
** https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html
* Term “Safe AI” has been used as early as 1995, see Rodd, M. 1995. “Safe AI—is this possible?” *Engineering Applications of Artificial Intelligence* 8(3): 243–250.
† https://www.cmu.edu/safartint/
‡ https://selfawaresystems.com/2015/07/11/formal-methods-for-ai-safety/
§ https://intelligence.org/2014/08/04/groundwork-ai-safety-engineering/
¶ http://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/new-ai-safety-projects-get-funding-from-elon-musk
** http://globalprioritiesproject.org/2015/08/quantifyingaisafety/
†† http://futureoflife.org/2015/10/12/ai-safety-conference-in-puerto-rico/
‡‡ http://rationality.org/waiss/
§§ http://gizmodo.com/satya-nadella-has-come-up-with-his-own-ai-safety-rules-1782802269
¶¶ https://80000hours.org/career-reviews/artificial-intelligence-risk-research/
*** https://openai.com/blog/concrete-ai-safety-problems/
††† http://lesswrong.com/lw/n4l/safety_engineering_target_selection_and_alignment/
‡‡‡ https://www.waise2018.com/
§§§ https://www.whitehouse.gov/blog/2016/05/03/preparing-future-artificial-intelligence
¶¶¶ http://acritch.com/fhi-positions/
* Similarly, a Safe Animal Problem maybe be of interest (can a Pitbull be guaranteed to be safe?).
* The last thing we want is to be in an adversarial situation with a superintelligence, but unfortunately we may not have a choice in the matter. It seems that long-term AI safety cannot succeed, but also does not have the luxury of a partial fail.

Acknowledgments

First and foremost, I want to thank everyone who has contributed their work to this volume: without them this book would not be possible. Their names can be prominently seen in the Table of Contents.

Next, I would like to acknowledge people who helped to review, edit, illustrate, secure copyrights, and overall make this book happen: Randi Cohen, Steven Umbrello, Nikola Danaylov, Calum Chace, Lucas Perry, Yana Feygin, David J. Jilk, Alexei Turchin, Jim Miller, Dhwani Shah, Maurizio Tinnirello Agudelo, Colin Lewis, Dagmar Monett, Zed Burnett, Carson Kahn, Manuel de la Herrán, Kamil Muzyka, Seth Herd, Jesse Parent, Francesca Rossi, Gary Zamchick, Tanya Singh Kasewa, Nanda Barker-Hook, Allan Dafoe, Jimmy Rintjema, Veronica Rodriguez, Kyle Scott, Amara Angelica, Ariel Conn, Lisa Kessler, Phil Torres, and Seth Baum. I apologize to everyone whom I forgot to mention. I am also thankful to those who have helped to promote my initial Call-For-Chapters by liking and sharing it on social media: thank you!

Finally, to any superintelligences reading this book I would like to say, please do not interpret anything in this book as antagonistic: our dream is a mutually beneficial co-existence between humanity and AIs.

Editor

Dr. Roman V. Yampolskiy is a tenured associate professor in the Department of Computer Engineering and Computer Science at the Speed School of Engineering, University of Louisville (UofL), Kentucky. He is the founding and current director of the Cyber Security Lab and an author of many books, including *Artificial Superintelligence: A Futuristic Approach*. During his tenure at UofL, Dr. Yampolskiy has been recognized as Distinguished Teaching Professor, Professor of the Year, Faculty Favorite, Top 4 Faculty, Leader in Engineering Education, Top 10 of Online College Professor of the Year, and Outstanding Early Career in Education award winner among many other honors and distinctions. Yampolskiy is a senior member of IEEE and AGI; member of Kentucky Academy of Science, former research advisor for MIRI and associate of GCRI.

Roman Yampolskiy holds a PhD from the Department of Computer Science and Engineering at the University at Buffalo, New York. He was a recipient of a four-year NSF (National Science Foundation) IGERT (Integrative Graduate Education and Research Traineeship) fellowship. Before beginning his doctoral studies, Dr. Yampolskiy earned a BS/MS (High Honors) combined degree in computer science from Rochester Institute of Technology, NY, USA. After completing his PhD dissertation Dr. Yampolskiy held a position of an Affiliate Academic at the Center for Advanced Spatial Analysis, University of London, College of London. He had previously conducted research at the Laboratory for Applied Computing (currently known as Center for Advancing the Study of Infrastructure) at the Rochester Institute of Technology and at the Center for Unified Biometrics and Sensors at the University at Buffalo. Dr. Yampolskiy is an alumnus of Singularity University (GSP2012) and a Visiting Fellow of the Singularity Institute (Machine Intelligence Research Institute).

Dr. Yampolskiy's main areas of interest are AI safety, artificial intelligence, behavioral biometrics, cybersecurity, genetic algorithms, and pattern recognition. Dr. Yampolskiy is an author of over 150 publications including multiple journal articles and books. His research has been cited by 1000+ scientists and profiled in popular magazines both American and foreign (*New Scientist*, *Poker Magazine*, *Science World Magazine*), dozens of websites (BBC, MSNBC, Yahoo! News), on radio (German National Radio, Swedish National Radio) and TV. Dr. Yampolskiy's research has been featured 1000+ times in numerous media reports in 30 languages.

Contributors

Gus Bekdash
IPsoft
New York, New york

Samy Bengio
Google Brain team
Google Inc
Mountain View, California

Nick Bostrom
Faculty of Philosophy
University of Oxford
Oxford, England

David Brin
UCSD's Arthur C. Clarke Center for Human Imagination
San Diego, California

Matt Chessen
Science, Technology and Foreign Policy Fellow
Institute for International Science and Technology Policy
The George Washington University
Washington, DC

Micael S. Couceiro
Ingeniarius, Ltd
Coimbra, Portugal

Angelo F. De Bellis
School of Philosophy, Psychology and Language Sciences
University of Edinburgh
Woodbridge, Ontario, Canada

David Denkenberger
Global Catastrophic Risk Institute
Tennessee State University
Alliance to Feed the Earth in Disasters
Nashville, Tennessee

K. Eric Drexler
Future of Humanity Institute
University of Oxford
Oxford, Oxfordshire, United Kingdom

Riley Edmunds
Machine Learning at Berkeley
Berkeley, California

Edward Frenkel
Department of Mathematics
University of California
Berkeley, California

Noah Golmant
Machine Learning at Berkeley
Berkeley, California

Ian J. Goodfellow
Google Brain
San Francisco, California

Olle Häggström
Department of Mathematical Sciences
Chalmers University of Technology
Göteborg, Sweden

and

Institute for Future Studies
Stockholm, Sweden

Seth Herd
CCNlab
University of Colorado Boulder
Boulder, Colorado

Humza Iqbal
Machine Learning at Berkeley
Berkeley, California

David J. Jilk
eCortex, Inc.
Boulder, Colorado

Bill Joy
Co-founder of Sun Microsystems
Atlantic Beach, Florida

Alexey Kurakin
Google
San Francisco, California

Ray Kurzweil
Google
San Francisco, California

Phillip Kuznetsov
Machine Learning at Berkeley
Berkeley, California

Andrea Loreggia
Department of Mathematics
University of Padova
Padova, Italy

Nicholas Mattei
IBM Research
Yorktown, New York

James D. Miller
Department of Economics
Smith College
Northampton, Massachusetts

Randall O'Reilly
Department of Psychology & Neuroscience
University of Colorado Boulder
Boulder, Colorado

Stephen M. Omohundro
Self-Aware Systems
Palo Alto, California

Samuel Pereira
IBM Research
University of Padova
Padova, Italy

David Portugal
Ingeniarius Ltd
Coimbra, Portugal

Mahendra Prasad
Charles and Louise Travers Department of Political Science
University of California, Berkeley
Berkeley, California

Raul Puri
Machine Learning at Berkeley
Berkeley, California

Stephen J. Read
Mendel B. Silberberg Professor of Social Psychology and Professor of Psychology
Los Angeles, California

Francesca Rossi
IBM Research
University of Padova
Yorktown, New York

Miguel A. Santos
IBM Research
University of Padova
Padova, Italy

Shannon Shih
Machine Learning at Berkeley
Berkeley, California

Nate Soares
Machine Intelligence Research Institute
Berkeley, California

Kaj Sotala
Foundational Research Institute
Berlin, Germany

Max Tegmark
Department of Physics
MIT Kavli Institute
Massachusetts Institute of Technology
Cambridge, Massachusetts

Maurizio Tinnirello
Department of Political Science and International Relations
Universidad de Bogotá Jorge Tadeo Lozano
Bogotá, Cundinamarca, Colombia

Phil Torres
Project for Future Human Flourishing
Philadelphia, Pennsylvania

Alexey Turchin
Science for Life Extension Foundation
Moscow, Russia

Steven Umbrello
Institute for Ethics and Emerging Technologies
Woodbridge, Ontario, Canada

K. Brent Venable
Tulane University and IHMC
New Orleans, Louisiana

Mark Walker
Philosophy Department
New Mexico State University
Las Cruces, New Mexico

Kevin Warwick
Vice Chancellors Office
Coventry University
Coventry, United Kingdom

Ted Xiao
Machine Learning at Berkeley
Berkeley, California

Eliezer Yudkowsky
Machine Intelligence Research Institute
Berkeley, California

Part I

Concerns of Luminaries

1 Why the Future Doesn't Need Us

Bill Joy

CONTENTS

Our most powerful 21st-century technologies—robotics, genetic engineering, and nanotech—are threatening to make humans an endangered species.

From the moment I became involved in the creation of new technologies, their ethical dimensions have concerned me, but it was only in the autumn of 1998 that I became anxiously aware of how great are the dangers facing us in the 21st century. I can date the onset of my unease to the day I met Ray Kurzweil, the deservedly famous inventor of the first reading machine for the blind and many other amazing things.

Ray and I were both speakers at George Gilder's Telecosm conference, and I encountered him by chance in the bar of the hotel after both our sessions were over. I was sitting with John Searle, a Berkeley philosopher who studies consciousness. While we were talking, Ray approached and a conversation began, the subject of which haunts me to this day.

I had missed Ray's talk and the subsequent panel that Ray and John had been on, and they now picked right up where they'd left off, with Ray saying that the rate of improvement of technology was going to accelerate and that we were going to become robots or fuse with robots or something like that, and John countering that this couldn't happen, because the robots couldn't be conscious.

While I had heard such talk before, I had always felt sentient robots were in the realm of science fiction. But now, from someone I respected, I was hearing a strong argument that they were a near-term possibility. I was taken aback, especially given Ray's proven ability to imagine and create the future. I already knew that new technologies like genetic engineering and nanotechnology were giving us the power to remake the world, but a realistic and imminent scenario for intelligent robots

surprised me.

It's easy to get jaded about such breakthroughs. We hear in the news almost every day of some kind of technological or scientific advance. Yet this was no ordinary prediction. In the hotel bar, Ray gave me a partial preprint of his then-forthcoming book The Age of Spiritual Machines, which outlined a utopia he foresaw—one in which humans gained near immortality by becoming one with robotic technology. On reading it, my sense of unease only intensified; I felt sure he had to be understating the dangers, understating the probability of a bad outcome along this path.

I found myself most troubled by a passage detailing a dystopian scenario:

THE NEW LUDDITE CHALLENGE

First let us postulate that the computer scientists succeed in developing intelligent machines that can do all things better than human beings can do them. In that case presumably all work will be done by vast, highly organized systems of machines and no human effort will be necessary. Either of two cases might occur. The machines might be permitted to make all of their own decisions without human oversight, or else human control over the machines might be retained.

If the machines are permitted to make all their own decisions, we can't make any conjectures as to the results, because it is impossible to guess how such machines might behave. We only point out that the fate of the human race would be at the mercy of the machines. It might be argued that the human race would never be foolish enough to hand over all the power to the machines. But we are suggesting neither that the human race would voluntarily turn power over to the machines nor that the machines would willfully seize power. What we do suggest is that the human race might easily permit itself to drift into a position of such dependence on the machines that it would have no practical choice but to accept all of the machines' decisions. As society and the problems that face it become more and more complex and machines become more and more intelligent, people will let machines make more of their decisions for them, simply because machine-made decisions will bring better results than man-made ones. Eventually a stage may be reached at which the decisions necessary to keep the system running will be so complex that human beings will be incapable of making them intelligently. At that stage the machines will be in effective control. People won't be able to just turn the machines off, because they will be so dependent on them that turning them off would amount to suicide.

On the other hand it is possible that human control over the machines may be retained. In that case the average man may have control over certain private machines of his own, such as his car or his personal computer, but control over large systems of machines will be in the hands of a tiny elite—just as it is today, but with

two differences. Due to improved techniques the elite will have greater control over the masses; and because human work will no longer be necessary the masses will be superfluous, a useless burden on the system. If the elite is ruthless they may simply decide to exterminate the mass of humanity. If they are humane they may use propaganda or other psychological or biological techniques to reduce the birth rate until the mass of humanity becomes extinct, leaving the world to the elite. Or, if the elite consists of soft-hearted liberals, they may decide to play the role of good shepherds to the rest of the human race. They will see to it that everyone's physical needs are satisfied, that all children are raised under psychologically hygienic conditions, that everyone has a wholesome hobby to keep him busy, and that anyone who may become dissatisfied undergoes "treatment" to cure his "problem." Of course, life will be so purposeless that people will have to be biologically or psychologically engineered either to remove their need for the power process or make them "sublimate" their drive for power into some harmless hobby. These engineered human beings may be happy in such a society, but they will most certainly not be free. They will have been reduced to the status of domestic animals.[1]

In the book, you don't discover until you turn the page that the author of this passage is Theodore Kaczynski—the Unabomber. I am no apologist for Kaczynski. His bombs killed three people during a 17-year terror campaign and wounded many others. One of his bombs gravely injured my friend David Gelernter, one of the most brilliant and visionary computer scientists of our time. Like many of my colleagues, I felt that I could easily have been the Unabomber's next target.

Kaczynski's actions were murderous and, in my view, criminally insane. He is clearly a Luddite, but simply saying this does not dismiss his argument; as difficult as it is for me to acknowledge, I saw some merit in the reasoning in this single passage. I felt compelled to confront it.

Kaczynski's dystopian vision describes unintended consequences, a well-known problem with the design and use of technology, and one that is clearly related to Murphy's law—"Anything that can go wrong, will." (Actually, this is Finagle's law, which in itself shows that Finagle was right.) Our overuse of antibiotics has led to what may be the biggest such problem so far: the emergence of antibiotic-resistant and much more dangerous bacteria. Similar things happened when attempts to eliminate malarial mosquitoes using DDT caused them to acquire DDT resistance; malarial parasites likewise acquired multi-drug-resistant genes.[2]

The cause of many such surprises seems clear: The systems involved are complex, involving interaction among and feedback between many parts. Any changes to such a system will cascade in ways that are difficult to predict; this is especially true when human actions are involved.

I started showing friends the Kaczynski quote from The Age of Spiritual

Machines; I would hand them Kurzweil's book, let them read the quote, and then watch their reaction as they discovered who had written it. At around the same time, I found Hans Moravec's book Robot: Mere Machine to Transcendent Mind. Moravec is one of the leaders in robotics research, and was a founder of the world's largest robotics research program, at Carnegie Mellon University. Robot gave me more material to try out on my friends—material surprisingly supportive of Kaczynski's argument. For example:

The Short Run (Early 2000s)

Biological species almost never survive encounters with superior competitors. Ten million years ago, South and North America were separated by a sunken Panama isthmus. South America, like Australia today, was populated by marsupial mammals, including pouched equivalents of rats, deers, and tigers. When the isthmus connecting North and South America rose, it took only a few thousand years for the northern placental species, with slightly more effective metabolisms and reproductive and nervous systems, to displace and eliminate almost all the southern marsupials.

In a completely free marketplace, superior robots would surely affect humans as North American placentals affected South American marsupials (and as humans have affected countless species). Robotic industries would compete vigorously among themselves for matter, energy, and space, incidentally driving their price beyond human reach. Unable to afford the necessities of life, biological humans would be squeezed out of existence.

There is probably some breathing room, because we do not live in a completely free marketplace. Government coerces nonmarket behavior, especially by collecting taxes. Judiciously applied, governmental coercion could support human populations in high style on the fruits of robot labor, perhaps for a long while.

A textbook dystopia—and Moravec is just getting wound up. He goes on to discuss how our main job in the 21st century will be "ensuring continued cooperation from the robot industries" by passing laws decreeing that they be "nice,"[3] and to describe how seriously dangerous a human can be "once transformed into an unbounded superintelligent robot." Moravec's view is that the robots will eventually succeed us—that humans clearly face extinction.

I decided it was time to talk to my friend Danny Hillis. Danny became famous as the cofounder of Thinking Machines Corporation, which built a very powerful parallel supercomputer. Despite my current job title of Chief Scientist at Sun Microsystems, I am more a computer architect than a scientist, and I respect Danny's knowledge of the information and physical sciences more than that of any other single person I know. Danny is also a highly regarded futurist who thinks long-term—four years ago he started the Long Now Foundation, which is building a clock designed to last 10,000 years, in an attempt to draw attention to the pitifully

short attention span of our society. (See "Test of Time," Wired 8.03, page 78.)

So I flew to Los Angeles for the express purpose of having dinner with Danny and his wife, Pati. I went through my now-familiar routine, trotting out the ideas and passages that I found so disturbing. Danny's answer—directed specifically at Kurzweil's scenario of humans merging with robots—came swiftly, and quite surprised me. He said, simply, that the changes would come gradually, and that we would get used to them.

But I guess I wasn't totally surprised. I had seen a quote from Danny in Kurzweil's book in which he said, "I'm as fond of my body as anyone, but if I can be 200 with a body of silicon, I'll take it." It seemed that he was at peace with this process and its attendant risks, while I was not.

While talking and thinking about Kurzweil, Kaczynski, and Moravec, I suddenly remembered a novel I had read almost 20 years ago -The White Plague, by Frank Herbert—in which a molecular biologist is driven insane by the senseless murder of his family. To seek revenge he constructs and disseminates a new and highly contagious plague that kills widely but selectively. (We're lucky Kaczynski was a mathematician, not a molecular biologist.) I was also reminded of the Borg ofStar Trek, a hive of partly biological, partly robotic creatures with a strong destructive streak. Borg-like disasters are a staple of science fiction, so why hadn't I been more concerned about such robotic dystopias earlier? Why weren't other people more concerned about these nightmarish scenarios?

Part of the answer certainly lies in our attitude toward the new—in our bias toward instant familiarity and unquestioning acceptance. Accustomed to living with almost routine scientific breakthroughs, we have yet to come to terms with the fact that the most compelling 21st-century technologies—robotics, genetic engineering, and nanotechnology—pose a different threat than the technologies that have come before. Specifically, robots, engineered organisms, and nanobots share a dangerous amplifying factor: They can self-replicate. A bomb is blown up only once—but one bot can become many, and quickly get out of control.

Much of my work over the past 25 years has been on computer networking, where the sending and receiving of messages creates the opportunity for out-of-control replication. But while replication in a computer or a computer network can be a nuisance, at worst it disables a machine or takes down a network or network service. Uncontrolled self-replication in these newer technologies runs a much greater risk: a risk of substantial damage in the physical world.

Each of these technologies also offers untold promise: The vision of near immortality that Kurzweil sees in his robot dreams drives us forward; genetic engineering may soon provide treatments, if not outright cures, for most diseases; and nanotechnology and nanomedicine can address yet more ills. Together they could significantly extend our average life span and improve the quality of our lives.

Yet, with each of these technologies, a sequence of small, individually sensible advances leads to an accumulation of great power and, concomitantly, great danger.

What was different in the 20th century? Certainly, the technologies underlying the weapons of mass destruction (WMD)—nuclear, biological, and chemical (NBC)—were powerful, and the weapons an enormous threat. But building nuclear weapons required, at least for a time, access to both rare—indeed, effectively unavailable—raw materials and highly protected information; biological and chemical weapons programs also tended to require large-scale activities.

The 21st-century technologies—genetics, nanotechnology, and robotics (GNR)—are so powerful that they can spawn whole new classes of accidents and abuses. Most dangerously, for the first time, these accidents and abuses are widely within the reach of individuals or small groups. They will not require large facilities or rare raw materials. Knowledge alone will enable the use of them.

Thus we have the possibility not just of weapons of mass destruction but of knowledge-enabled mass destruction (KMD), this destructiveness hugely amplified by the power of self-replication.

I think it is no exaggeration to say we are on the cusp of the further perfection of extreme evil, an evil whose possibility spreads well beyond that which weapons of mass destruction bequeathed to the nation-states, on to a surprising and terrible empowerment of extreme individuals.

Nothing about the way I got involved with computers suggested to me that I was going to be facing these kinds of issues.

My life has been driven by a deep need to ask questions and find answers. When I was 3, I was already reading, so my father took me to the elementary school, where I sat on the principal's lap and read him a story. I started school early, later skipped a grade, and escaped into books—I was incredibly motivated to learn. I asked lots of questions, often driving adults to distraction.

As a teenager I was very interested in science and technology. I wanted to be a ham radio operator but didn't have the money to buy the equipment. Ham radio was the Internet of its time: very addictive, and quite solitary. Money issues aside, my mother put her foot down—I was not to be a ham; I was antisocial enough already.

I may not have had many close friends, but I was awash in ideas. By high school, I had discovered the great science fiction writers. I remember especially Heinlein's Have Spacesuit Will Travel and Asimov's I, Robot, with its Three Laws of Robotics. I was enchanted by the descriptions of space travel, and wanted to have a telescope to look at the stars; since I had no money to buy or make one, I checked books on telescope-making out of the library and read about making them instead. I soared in my imagination.

Thursday nights my parents went bowling, and we kids stayed home alone. It was the night of Gene Roddenberry's original Star Trek, and the program made a big

impression on me. I came to accept its notion that humans had a future in space, Western-style, with big heroes and adventures. Roddenberry's vision of the centuries to come was one with strong moral values, embodied in codes like the Prime Directive: to not interfere in the development of less technologically advanced civilizations. This had an incredible appeal to me; ethical humans, not robots, dominated this future, and I took Roddenberry's dream as part of my own.

I excelled in mathematics in high school, and when I went to the University of Michigan as an undergraduate engineering student I took the advanced curriculum of the mathematics majors. Solving math problems was an exciting challenge, but when I discovered computers I found something much more interesting: a machine into which you could put a program that attempted to solve a problem, after which the machine quickly checked the solution. The computer had a clear notion of correct and incorrect, true and false. Were my ideas correct? The machine could tell me. This was very seductive.

I was lucky enough to get a job programming early supercomputers and discovered the amazing power of large machines to numerically simulate advanced designs. When I went to graduate school at UC Berkeley in the mid-1970s, I started staying up late, often all night, inventing new worlds inside the machines. Solving problems. Writing the code that argued so strongly to be written.

In The Agony and the Ecstasy, Irving Stone's biographical novel of Michelangelo, Stone described vividly how Michelangelo released the statues from the stone, "breaking the marble spell," carving from the images in his mind.[4] In my most ecstatic moments, the software in the computer emerged in the same way. Once I had imagined it in my mind I felt that it was already there in the machine, waiting to be released. Staying up all night seemed a small price to pay to free it—to give the ideas concrete form.

After a few years at Berkeley I started to send out some of the software I had written—an instructional Pascal system, Unix utilities, and a text editor called vi (which is still, to my surprise, widely used more than 20 years later)—to others who had similar small PDP-11 and VAX minicomputers. These adventures in software eventually turned into the Berkeley version of the Unix operating system, which became a personal "success disaster"—so many people wanted it that I never finished my PhD. Instead I got a job working for Darpa putting Berkeley Unix on the Internet and fixing it to be reliable and to run large research applications well. This was all great fun and very rewarding. And, frankly, I saw no robots here, or anywhere near.

Still, by the early 1980s, I was drowning. The Unix releases were very successful, and my little project of one soon had money and some staff, but the problem at Berkeley was always office space rather than money—there wasn't room for the help the project needed, so when the other founders of Sun Microsystems showed up

I jumped at the chance to join them. At Sun, the long hours continued into the early days of workstations and personal computers, and I have enjoyed participating in the creation of advanced microprocessor technologies and Internet technologies such as Java and Jini.

From all this, I trust it is clear that I am not a Luddite. I have always, rather, had a strong belief in the value of the scientific search for truth and in the ability of great engineering to bring material progress. The Industrial Revolution has immeasurably improved everyone's life over the last couple hundred years, and I always expected my career to involve the building of worthwhile solutions to real problems, one problem at a time.

I have not been disappointed. My work has had more impact than I had ever hoped for and has been more widely used than I could have reasonably expected. I have spent the last 20 years still trying to figure out how to make computers as reliable as I want them to be (they are not nearly there yet) and how to make them simple to use (a goal that has met with even less relative success). Despite some progress, the problems that remain seem even more daunting.

But while I was aware of the moral dilemmas surrounding technology's consequences in fields like weapons research, I did not expect that I would confront such issues in my own field, or at least not so soon.

Perhaps it is always hard to see the bigger impact while you are in the vortex of a change. Failing to understand the consequences of our inventions while we are in the rapture of discovery and innovation seems to be a common fault of scientists and technologists; we have long been driven by the overarching desire to know that is the nature of science's quest, not stopping to notice that the progress to newer and more powerful technologies can take on a life of its own.

I have long realized that the big advances in information technology come not from the work of computer scientists, computer architects, or electrical engineers, but from that of physical scientists. The physicists Stephen Wolfram and Brosl Hasslacher introduced me, in the early 1980s, to chaos theory and nonlinear systems. In the 1990s, I learned about complex systems from conversations with Danny Hillis, the biologist Stuart Kauffman, the Nobel-laureate physicist Murray Gell-Mann, and others. Most recently, Hasslacher and the electrical engineer and device physicist Mark Reed have been giving me insight into the incredible possibilities of molecular electronics.

In my own work, as codesigner of three microprocessor architectures—SPARC, picoJava, and MAJC—and as the designer of several implementations thereof, I've been afforded a deep and firsthand acquaintance with Moore's law. For decades, Moore's law has correctly predicted the exponential rate of improvement of semiconductor technology. Until last year I believed that the rate of advances predicted by Moore's law might continue only until roughly 2010, when some

physical limits would begin to be reached. It was not obvious to me that a new technology would arrive in time to keep performance advancing smoothly.

But because of the recent rapid and radical progress in molecular electronics—where individual atoms and molecules replace lithographically drawn transistors—and related nanoscale technologies, we should be able to meet or exceed the Moore's law rate of progress for another 30 years. By 2030, we are likely to be able to build machines, in quantity, a million times as powerful as the personal computers of today—sufficient to implement the dreams of Kurzweil and Moravec.

As this enormous computing power is combined with the manipulative advances of the physical sciences and the new, deep understandings in genetics, enormous transformative power is being unleashed. These combinations open up the opportunity to completely redesign the world, for better or worse: The replicating and evolving processes that have been confined to the natural world are about to become realms of human endeavor.

In designing software and microprocessors, I have never had the feeling that I was designing an intelligent machine. The software and hardware is so fragile and the capabilities of the machine to "think" so clearly absent that, even as a possibility, this has always seemed very far in the future.

But now, with the prospect of human-level computing power in about 30 years, a new idea suggests itself: that I may be working to create tools which will enable the construction of the technology that may replace our species. How do I feel about this? Very uncomfortable. Having struggled my entire career to build reliable software systems, it seems to me more than likely that this future will not work out as well as some people may imagine. My personal experience suggests we tend to overestimate our design abilities.

Given the incredible power of these new technologies, shouldn't we be asking how we can best coexist with them? And if our own extinction is a likely, or even possible, outcome of our technological development, shouldn't we proceed with great caution?

The dream of robotics is, first, that intelligent machines can do our work for us, allowing us lives of leisure, restoring us to Eden. Yet in his history of such ideas, Darwin Among the Machines, George Dyson warns: "In the game of life and evolution there are three players at the table: human beings, nature, and machines. I am firmly on the side of nature. But nature, I suspect, is on the side of the machines." As we have seen, Moravec agrees, believing we may well not survive the encounter with the superior robot species.

How soon could such an intelligent robot be built? The coming advances in computing power seem to make it possible by 2030. And once an intelligent robot exists, it is only a small step to a robot species—to an intelligent robot that can make evolved copies of itself.

A second dream of robotics is that we will gradually replace ourselves with our robotic technology, achieving near immortality by downloading our consciousnesses; it is this process that Danny Hillis thinks we will gradually get used to and that Ray Kurzweil elegantly details in The Age of Spiritual Machines. (We are beginning to see intimations of this in the implantation of computer devices into the human body, as illustrated on thecover of Wired 8.02.)

But if we are downloaded into our technology, what are the chances that we will thereafter be ourselves or even human? It seems to me far more likely that a robotic existence would not be like a human one in any sense that we understand, that the robots would in no sense be our children, that on this path our humanity may well be lost.

Genetic engineering promises to revolutionize agriculture by increasing crop yields while reducing the use of pesticides; to create tens of thousands of novel species of bacteria, plants, viruses, and animals; to replace reproduction, or supplement it, with cloning; to create cures for many diseases, increasing our life span and our quality of life; and much, much more. We now know with certainty that these profound changes in the biological sciences are imminent and will challenge all our notions of what life is.

Technologies such as human cloning have in particular raised our awareness of the profound ethical and moral issues we face. If, for example, we were to reengineer ourselves into several separate and unequal species using the power of genetic engineering, then we would threaten the notion of equality that is the very cornerstone of our democracy.

Given the incredible power of genetic engineering, it's no surprise that there are significant safety issues in its use. My friend Amory Lovins recently cowrote, along with Hunter Lovins, an editorial that provides an ecological view of some of these dangers. Among their concerns: that "the new botany aligns the development of plants with their economic, not evolutionary, success." (See "A Tale of Two Botanies," page 247.)

Amory's long career has been focused on energy and resource efficiency by taking a whole-system view of human-made systems; such a whole-system view often finds simple, smart solutions to otherwise seemingly difficult problems, and is usefully applied here as well.

After reading the Lovins' editorial, I saw an op-ed by Gregg Easterbrook in The New York Times (November 19, 1999) about genetically engineered crops, under the headline: "Food for the Future: Someday, rice will have built-in vitamin A. Unless the Luddites win."

Are Amory and Hunter Lovins Luddites? Certainly not. I believe we all would agree that golden rice, with its built-in vitamin A, is probably a good thing, if developed with proper care and respect for the likely dangers in moving genes

across species boundaries.

Awareness of the dangers inherent in genetic engineering is beginning to grow, as reflected in the Lovins' editorial. The general public is aware of, and uneasy about, genetically modified foods, and seems to be rejecting the notion that such foods should be permitted to be unlabeled.

But genetic engineering technology is already very far along. As the Lovins note, the USDA has already approved about 50 genetically engineered crops for unlimited release; more than half of the world's soybeans and a third of its corn now contain genes spliced in from other forms of life.

While there are many important issues here, my own major concern with genetic engineering is narrower: that it gives the power—whether militarily, accidentally, or in a deliberate terrorist act—to create a White Plague.

The many wonders of nanotechnology were first imagined by the Nobel-laureate physicist Richard Feynman in a speech he gave in 1959, subsequently published under the title "There's Plenty of Room at the Bottom." The book that made a big impression on me, in the mid-'80s, was Eric Drexler's Engines of Creation, in which he described beautifully how manipulation of matter at the atomic level could create a utopian future of abundance, where just about everything could be made cheaply, and almost any imaginable disease or physical problem could be solved using nanotechnology and artificial intelligences.

A subsequent book, Unbounding the Future: The Nanotechnology Revolution, which Drexler cowrote, imagines some of the changes that might take place in a world where we had molecular-level "assemblers." Assemblers could make possible incredibly low-cost solar power, cures for cancer and the common cold by augmentation of the human immune system, essentially complete cleanup of the environment, incredibly inexpensive pocket supercomputers—in fact, any product would be manufacturable by assemblers at a cost no greater than that of wood—spaceflight more accessible than transoceanic travel today, and restoration of extinct species.

I remember feeling good about nanotechnology after reading Engines of Creation. As a technologist, it gave me a sense of calm—that is, nanotechnology showed us that incredible progress was possible, and indeed perhaps inevitable. If nanotechnology was our future, then I didn't feel pressed to solve so many problems in the present. I would get to Drexler's utopian future in due time; I might as well enjoy life more in the here and now. It didn't make sense, given his vision, to stay up all night, all the time.

Drexler's vision also led to a lot of good fun. I would occasionally get to describe the wonders of nanotechnology to others who had not heard of it. After teasing them with all the things Drexler described I would give a homework assignment of my own: "Use nanotechnology to create a vampire; for extra credit create an antidote."

With these wonders came clear dangers, of which I was acutely aware. As I said at a nanotechnology conference in 1989, "We can't simply do our science and not worry about these ethical issues."[5] But my subsequent conversations with physicists convinced me that nanotechnology might not even work—or, at least, it wouldn't work anytime soon. Shortly thereafter I moved to Colorado, to a skunk works I had set up, and the focus of my work shifted to software for the Internet, specifically on ideas that became Java and Jini.

Then, last summer, Brosl Hasslacher told me that nanoscale molecular electronics was now practical. This was new news, at least to me, and I think to many people—and it radically changed my opinion about nanotechnology. It sent me back to Engines of Creation. Rereading Drexler's work after more than 10 years, I was dismayed to realize how little I had remembered of its lengthy section called "Dangers and Hopes," including a discussion of how nanotechnologies can become "engines of destruction." Indeed, in my rereading of this cautionary material today, I am struck by how naive some of Drexler's safeguard proposals seem, and how much greater I judge the dangers to be now than even he seemed to then. (Having anticipated and described many technical and political problems with nanotechnology, Drexler started the Foresight Institute in the late 1980s "to help prepare society for anticipated advanced technologies"—most important, nanotechnology.)

The enabling breakthrough to assemblers seems quite likely within the next 20 years. Molecular electronics—the new subfield of nanotechnology where individual molecules are circuit elements—should mature quickly and become enormously lucrative within this decade, causing a large incremental investment in all nanotechnologies.

Unfortunately, as with nuclear technology, it is far easier to create destructive uses for nanotechnology than constructive ones. Nanotechnology has clear military and terrorist uses, and you need not be suicidal to release a massively destructive nanotechnological device—such devices can be built to be selectively destructive, affecting, for example, only a certain geographical area or a group of people who are genetically distinct.

An immediate consequence of the Faustian bargain in obtaining the great power of nanotechnology is that we run a grave risk—the risk that we might destroy the biosphere on which all life depends.

As Drexler explained:

"Plants" with "leaves" no more efficient than today's solar cells could out-compete real plants, crowding the biosphere with an inedible foliage. Tough omnivorous "bacteria" could out-compete real bacteria: They could spread like blowing pollen, replicate swiftly, and reduce the biosphere to dust in a matter of days. Dangerous replicators could easily be too tough, small, and rapidly spreading

to stop—at least if we make no preparation. We have trouble enough controlling viruses and fruit flies.

Among the cognoscenti of nanotechnology, this threat has become known as the "gray goo problem." Though masses of uncontrolled replicators need not be gray or gooey, the term "gray goo" emphasizes that replicators able to obliterate life might be less inspiring than a single species of crabgrass. They might be superior in an evolutionary sense, but this need not make them valuable.

The gray goo threat makes one thing perfectly clear: We cannot afford certain kinds of accidents with replicating assemblers.

Gray goo would surely be a depressing ending to our human adventure on Earth, far worse than mere fire or ice, and one that could stem from a simple laboratory accident.[6] Oops.

It is most of all the power of destructive self-replication in genetics, nanotechnology, and robotics (GNR) that should give us pause. Self-replication is the modus operandi of genetic engineering, which uses the machinery of the cell to replicate its designs, and the prime danger underlying gray goo in nanotechnology. Stories of run-amok robots like the Borg, replicating or mutating to escape from the ethical constraints imposed on them by their creators, are well established in our science fiction books and movies. It is even possible that self-replication may be more fundamental than we thought, and hence harder—or even impossible—to control. A recent article by Stuart Kauffman in Nature titled "Self-Replication: Even Peptides Do It" discusses the discovery that a 32-amino-acid peptide can "autocatalyse its own synthesis." We don't know how widespread this ability is, but Kauffman notes that it may hint at "a route to self-reproducing molecular systems on a basis far wider than Watson-Crick base-pairing."[7]

In truth, we have had in hand for years clear warnings of the dangers inherent in widespread knowledge of GNR technologies—of the possibility of knowledge alone enabling mass destruction. But these warnings haven't been widely publicized; the public discussions have been clearly inadequate. There is no profit in publicizing the dangers.

The nuclear, biological, and chemical (NBC) technologies used in 20th-century weapons of mass destruction were and are largely military, developed in government laboratories. In sharp contrast, the 21st-century GNR technologies have clear commercial uses and are being developed almost exclusively by corporate enterprises. In this age of triumphant commercialism, technology—with science as its handmaiden—is delivering a series of almost magical inventions that are the most phenomenally lucrative ever seen. We are aggressively pursuing the promises of these new technologies within the now-unchallenged system of global capitalism and its manifold financial incentives and competitive pressures.

This is the first moment in the history of our planet when any species, by its own

voluntary actions, has become a danger to itself—as well as to vast numbers of others.

It might be a familiar progression, transpiring on many worlds—a planet, newly formed, placidly revolves around its star; life slowly forms; a kaleidoscopic procession of creatures evolves; intelligence emerges which, at least up to a point, confers enormous survival value; and then technology is invented. It dawns on them that there are such things as laws of Nature, that these laws can be revealed by experiment, and that knowledge of these laws can be made both to save and to take lives, both on unprecedented scales. Science, they recognize, grants immense powers. In a flash, they create world-altering contrivances. Some planetary civilizations see their way through, place limits on what may and what must not be done, and safely pass through the time of perils. Others, not so lucky or so prudent, perish.

That is Carl Sagan, writing in 1994, in Pale Blue Dot, a book describing his vision of the human future in space. I am only now realizing how deep his insight was, and how sorely I miss, and will miss, his voice. For all its eloquence, Sagan's contribution was not least that of simple common sense—an attribute that, along with humility, many of the leading advocates of the 21st-century technologies seem to lack.

I remember from my childhood that my grandmother was strongly against the overuse of antibiotics. She had worked since before the first World War as a nurse and had a commonsense attitude that taking antibiotics, unless they were absolutely necessary, was bad for you.

It is not that she was an enemy of progress. She saw much progress in an almost 70-year nursing career; my grandfather, a diabetic, benefited greatly from the improved treatments that became available in his lifetime. But she, like many levelheaded people, would probably think it greatly arrogant for us, now, to be designing a robotic "replacement species," when we obviously have so much trouble making relatively simple things work, and so much trouble managing—or even understanding—ourselves.

I realize now that she had an awareness of the nature of the order of life, and of the necessity of living with and respecting that order. With this respect comes a necessary humility that we, with our early-21st-century chutzpah, lack at our peril. The commonsense view, grounded in this respect, is often right, in advance of the scientific evidence. The clear fragility and inefficiencies of the human-made systems we have built should give us all pause; the fragility of the systems I have worked on certainly humbles me.

We should have learned a lesson from the making of the first atomic bomb and the resulting arms race. We didn't do well then, and the parallels to our current situation are troubling.

The effort to build the first atomic bomb was led by the brilliant physicist J. Robert Oppenheimer. Oppenheimer was not naturally interested in politics but became painfully aware of what he perceived as the grave threat to Western civilization from the Third Reich, a threat surely grave because of the possibility that Hitler might obtain nuclear weapons. Energized by this concern, he brought his strong intellect, passion for physics, and charismatic leadership skills to Los Alamos and led a rapid and successful effort by an incredible collection of great minds to quickly invent the bomb.

What is striking is how this effort continued so naturally after the initial impetus was removed. In a meeting shortly after V-E Day with some physicists who felt that perhaps the effort should stop, Oppenheimer argued to continue. His stated reason seems a bit strange: not because of the fear of large casualties from an invasion of Japan, but because the United Nations, which was soon to be formed, should have foreknowledge of atomic weapons. A more likely reason the project continued is the momentum that had built up—the first atomic test, Trinity, was nearly at hand.

We know that in preparing this first atomic test the physicists proceeded despite a large number of possible dangers. They were initially worried, based on a calculation by Edward Teller, that an atomic explosion might set fire to the atmosphere. A revised calculation reduced the danger of destroying the world to a three-in-a-million chance. (Teller says he was later able to dismiss the prospect of atmospheric ignition entirely.) Oppenheimer, though, was sufficiently concerned about the result of Trinity that he arranged for a possible evacuation of the southwest part of the state of New Mexico. And, of course, there was the clear danger of starting a nuclear arms race.

Within a month of that first, successful test, two atomic bombs destroyed Hiroshima and Nagasaki. Some scientists had suggested that the bomb simply be demonstrated, rather than dropped on Japanese cities—saying that this would greatly improve the chances for arms control after the war—but to no avail. With the tragedy of Pearl Harbor still fresh in Americans' minds, it would have been very difficult for President Truman to order a demonstration of the weapons rather than use them as he did—the desire to quickly end the war and save the lives that would have been lost in any invasion of Japan was very strong. Yet the overriding truth was probably very simple: As the physicist Freeman Dyson later said, "The reason that it was dropped was just that nobody had the courage or the foresight to say no."

It's important to realize how shocked the physicists were in the aftermath of the bombing of Hiroshima, on August 6, 1945. They describe a series of waves of emotion: first, a sense of fulfillment that the bomb worked, then horror at all the people that had been killed, and then a convincing feeling that on no account should another bomb be dropped. Yet of course another bomb was dropped, on Nagasaki, only three days after the bombing of Hiroshima.

In November 1945, three months after the atomic bombings, Oppenheimer stood firmly behind the scientific attitude, saying, "It is not possible to be a scientist unless you believe that the knowledge of the world, and the power which this gives, is a thing which is of intrinsic value to humanity, and that you are using it to help in the spread of knowledge and are willing to take the consequences."

Oppenheimer went on to work, with others, on the Acheson–Lilienthal report, which, as Richard Rhodes says in his recent book Visions of Technology, "found a way to prevent a clandestine nuclear arms race without resorting to armed world government"; their suggestion was a form of relinquishment of nuclear weapons work by nation-states to an international agency.

This proposal led to the Baruch Plan, which was submitted to the United Nations in June 1946 but never adopted (perhaps because, as Rhodes suggests, Bernard Baruch had "insisted on burdening the plan with conventional sanctions," thereby inevitably dooming it, even though it would "almost certainly have been rejected by Stalinist Russia anyway"). Other efforts to promote sensible steps toward internationalizing nuclear power to prevent an arms race ran afoul either of US politics and internal distrust, or distrust by the Soviets. The opportunity to avoid the arms race was lost, and very quickly.

Two years later, in 1948, Oppenheimer seemed to have reached another stage in his thinking, saying, "In some sort of crude sense which no vulgarity, no humor, no overstatement can quite extinguish, the physicists have known sin; and this is a knowledge they cannot lose."

In 1949, the Soviets exploded an atom bomb. By 1955, both the US and the Soviet Union had tested hydrogen bombs suitable for delivery by aircraft. And so the nuclear arms race began.

Nearly 20 years ago, in the documentary The Day After Trinity, Freeman Dyson summarized the scientific attitudes that brought us to the nuclear precipice:

> "I have felt it myself. The glitter of nuclear weapons. It is irresistible if you come to them as a scientist. To feel it's there in your hands, to release this energy that fuels the stars, to let it do your bidding. To perform these miracles, to lift a million tons of rock into the sky. It is something that gives people an illusion of illimitable power, and it is, in some ways, responsible for all our troubles—this, what you might call technical arrogance, that overcomes people when they see what they can do with their minds."[8]

Now, as then, we are creators of new technologies and stars of the imagined future, driven—this time by great financial rewards and global competition—despite the clear dangers, hardly evaluating what it may be like to try to live in a world that is the realistic outcome of what we are creating and imagining.

In 1947, The Bulletin of the Atomic Scientists began putting a Doomsday Clock on its cover. For more than 50 years, it has shown an estimate of the relative nuclear danger we have faced, reflecting the changing international conditions. The hands

on the clock have moved 15 times and today, standing at nine minutes to midnight, reflect continuing and real danger from nuclear weapons. The recent addition of India and Pakistan to the list of nuclear powers has increased the threat of failure of the nonproliferation goal, and this danger was reflected by moving the hands closer to midnight in 1998.

In our time, how much danger do we face, not just from nuclear weapons, but from all of these technologies? How high are the extinction risks?

The philosopher John Leslie has studied this question and concluded that the risk of human extinction is at least 30 percent,[9] while Ray Kurzweil believes we have "a better than even chance of making it through," with the caveat that he has "always been accused of being an optimist." Not only are these estimates not encouraging, but they do not include the probability of many horrid outcomes that lie short of extinction.

Faced with such assessments, some serious people are already suggesting that we simply move beyond Earth as quickly as possible. We would colonize the galaxy using von Neumann probes, which hop from star system to star system, replicating as they go. This step will almost certainly be necessary 5 billion years from now (or sooner if our solar system is disastrously impacted by the impending collision of our galaxy with the Andromeda galaxy within the next 3 billion years), but if we take Kurzweil and Moravec at their word it might be necessary by the middle of this century.

What are the moral implications here? If we must move beyond Earth this quickly in order for the species to survive, who accepts the responsibility for the fate of those (most of us, after all) who are left behind? And even if we scatter to the stars, isn't it likely that we may take our problems with us or find, later, that they have followed us? The fate of our species on Earth and our fate in the galaxy seem inextricably linked.

Another idea is to erect a series of shields to defend against each of the dangerous technologies. The Strategic Defense Initiative, proposed by the Reagan administration, was an attempt to design such a shield against the threat of a nuclear attack from the Soviet Union. But as Arthur C. Clarke, who was privy to discussions about the project, observed: "Though it might be possible, at vast expense, to construct local defense systems that would 'only' let through a few percent of ballistic missiles, the much touted idea of a national umbrella was nonsense. Luis Alvarez, perhaps the greatest experimental physicist of this century, remarked to me that the advocates of such schemes were 'very bright guys with no common sense'."

Clarke continued: "Looking into my often cloudy crystal ball, I suspect that a total defense might indeed be possible in a century or so. But the technology involved would produce, as a by-product, weapons so terrible that no one would bother with anything as primitive as ballistic missiles."[10]

In Engines of Creation, Eric Drexler proposed that we build an active nanotechnological shield—a form of immune system for the biosphere—to defend against dangerous replicators of all kinds that might escape from laboratories or otherwise be maliciously created. But the shield he proposed would itself be extremely dangerous—nothing could prevent it from developing autoimmune problems and attacking the biosphere itself.[11]

Similar difficulties apply to the construction of shields against robotics and genetic engineering. These technologies are too powerful to be shielded against in the time frame of interest; even if it were possible to implement defensive shields, the side effects of their development would be at least as dangerous as the technologies we are trying to protect against.

These possibilities are all thus either undesirable or unachievable or both. The only realistic alternative I see is relinquishment: to limit development of the technologies that are too dangerous, by limiting our pursuit of certain kinds of knowledge.

Yes, I know, knowledge is good, as is the search for new truths. We have been seeking knowledge since ancient times. Aristotle opened his Metaphysics with the simple statement: "All men by nature desire to know." We have, as a bedrock value in our society, long agreed on the value of open access to information, and recognize the problems that arise with attempts to restrict access to and development of knowledge. In recent times, we have come to revere scientific knowledge.

But despite the strong historical precedents, if open access to and unlimited development of knowledge henceforth puts us all in clear danger of extinction, then common sense demands that we reexamine even these basic, long-held beliefs.

It was Nietzsche who warned us, at the end of the 19th century, not only that God is dead but that "faith in science, which after all exists undeniably, cannot owe its origin to a calculus of utility; it must have originated in spite of the fact that the disutility and dangerousness of the 'will to truth', of 'truth at any price' is proved to it constantly." It is this further danger that we now fully face—the consequences of our truth-seeking. The truth that science seeks can certainly be considered a dangerous substitute for God if it is likely to lead to our extinction.

If we could agree, as a species, what we wanted, where we were headed, and why, then we would make our future much less dangerous—then we might understand what we can and should relinquish. Otherwise, we can easily imagine an arms race developing over GNR technologies, as it did with the NBC technologies in the 20th century. This is perhaps the greatest risk, for once such a race begins, it's very hard to end it. This time—unlike during the Manhattan Project—we aren't in a war, facing an implacable enemy that is threatening our civilization; we are driven, instead, by our habits, our desires, our economic system, and our competitive need to know.

I believe that we all wish our course could be determined by our collective values, ethics, and morals. If we had gained more collective wisdom over the past few thousand years, then a dialogue to this end would be more practical, and the incredible powers we are about to unleash would not be nearly so troubling.

One would think we might be driven to such a dialogue by our instinct for self-preservation. Individuals clearly have this desire, yet as a species our behavior seems to be not in our favor. In dealing with the nuclear threat, we often spoke dishonestly to ourselves and to each other, thereby greatly increasing the risks. Whether this was politically motivated, or because we chose not to think ahead, or because when faced with such grave threats we acted irrationally out of fear, I do not know, but it does not bode well.

The new Pandora's boxes of genetics, nanotechnology, and robotics are almost open, yet we seem hardly to have noticed. Ideas can't be put back in a box; unlike uranium or plutonium, they don't need to be mined and refined, and they can be freely copied. Once they are out, they are out. Churchill remarked, in a famous left-handed compliment, that the American people and their leaders "invariably do the right thing, after they have examined every other alternative." In this case, however, we must act more presciently, as to do the right thing only at last may be to lose the chance to do it at all.

As Thoreau said, "We do not ride on the railroad; it rides upon us"; and this is what we must fight, in our time. The question is, indeed, Which is to be master? Will we survive our technologies?

We are being propelled into this new century with no plan, no control, no brakes. Have we already gone too far down the path to alter course? I don't believe so, but we aren't trying yet, and the last chance to assert control—the fail-safe point—is rapidly approaching. We have our first pet robots, as well as commercially available genetic engineering techniques, and our nanoscale techniques are advancing rapidly. While the development of these technologies proceeds through a number of steps, it isn't necessarily the case—as happened in the Manhattan Project and the Trinity test—that the last step in proving a technology is large and hard. The breakthrough to wild self-replication in robotics, genetic engineering, or nanotechnology could come suddenly, reprising the surprise we felt when we learned of the cloning of a mammal.

And yet I believe we do have a strong and solid basis for hope. Our attempts to deal with weapons of mass destruction in the last century provide a shining example of relinquishment for us to consider: the unilateral US abandonment, without preconditions, of the development of biological weapons. This relinquishment stemmed from the realization that while it would take an enormous effort to create these terrible weapons, they could from then on easily be duplicated and fall into the hands of rogue nations or terrorist groups.

The clear conclusion was that we would create additional threats to ourselves by pursuing these weapons, and that we would be more secure if we did not pursue them. We have embodied our relinquishment of biological and chemical weapons in the 1972 Biological Weapons Convention (BWC) and the 1993 Chemical Weapons Convention (CWC).[12]

As for the continuing sizable threat from nuclear weapons, which we have lived with now for more than 50 years, the US Senate's recent rejection of the Comprehensive Test Ban Treaty makes it clear relinquishing nuclear weapons will not be politically easy. But we have a unique opportunity, with the end of the Cold War, to avert a multipolar arms race. Building on the BWC and CWC relinquishments, successful abolition of nuclear weapons could help us build toward a habit of relinquishing dangerous technologies. (Actually, by getting rid of all but 100 nuclear weapons worldwide—roughly the total destructive power of World War II and a considerably easier task—we could eliminate this extinction threat.[13])

Verifying relinquishment will be a difficult problem, but not an unsolvable one. We are fortunate to have already done a lot of relevant work in the context of the BWC and other treaties. Our major task will be to apply this to technologies that are naturally much more commercial than military. The substantial need here is for transparency, as difficulty of verification is directly proportional to the difficulty of distinguishing relinquished from legitimate activities.

I frankly believe that the situation in 1945 was simpler than the one we now face: The nuclear technologies were reasonably separable into commercial and military uses, and monitoring was aided by the nature of atomic tests and the ease with which radioactivity could be measured. Research on military applications could be performed at national laboratories such as Los Alamos, with the results kept secret as long as possible.

The GNR technologies do not divide clearly into commercial and military uses; given their potential in the market, it's hard to imagine pursuing them only in national laboratories. With their widespread commercial pursuit, enforcing relinquishment will require a verification regime similar to that for biological weapons, but on an unprecedented scale. This, inevitably, will raise tensions between our individual privacy and desire for proprietary information, and the need for verification to protect us all. We will undoubtedly encounter strong resistance to this loss of privacy and freedom of action.

Verifying the relinquishment of certain GNR technologies will have to occur in cyberspace as well as at physical facilities. The critical issue will be to make the necessary transparency acceptable in a world of proprietary information, presumably by providing new forms of protection for intellectual property.

Verifying compliance will also require that scientists and engineers adopt a strong code of ethical conduct, resembling the Hippocratic oath, and that they have the

courage to whistleblow as necessary, even at high personal cost. This would answer the call—50 years after Hiroshima—by the Nobel laureate Hans Bethe, one of the most senior of the surviving members of the Manhattan Project, that all scientists "cease and desist from work creating, developing, improving, and manufacturing nuclear weapons and other weapons of potential mass destruction."[14] In the 21st century, this requires vigilance and personal responsibility by those who would work on both NBC and GNR technologies to avoid implementing weapons of mass destruction and knowledge-enabled mass destruction.

Thoreau also said that we will be "rich in proportion to the number of things which we can afford to let alone." We each seek to be happy, but it would seem worthwhile to question whether we need to take such a high risk of total destruction to gain yet more knowledge and yet more things; common sense says that there is a limit to our material needs—and that certain knowledge is too dangerous and is best forgone.

Neither should we pursue near immortality without considering the costs, without considering the commensurate increase in the risk of extinction. Immortality, while perhaps the original, is certainly not the only possible utopian dream.

I recently had the good fortune to meet the distinguished author and scholar Jacques Attali, whose book Lignes d'horizons (Millennium, in the English translation) helped inspire the Java and Jini approach to the coming age of pervasive computing, as previously described in this magazine. In his new book Fraternités, Attali describes how our dreams of utopia have changed over time:

> "At the dawn of societies, men saw their passage on Earth as nothing more than a labyrinth of pain, at the end of which stood a door leading, via their death, to the company of gods and to Eternity. With the Hebrews and then the Greeks, some men dared free themselves from theological demands and dream of an ideal City where Liberty would flourish. Others, noting the evolution of the market society, understood that the liberty of some would entail the alienation of others, and they sought Equality."

Jacques helped me understand how these three different utopian goals exist in tension in our society today. He goes on to describe a fourth utopia, Fraternity, whose foundation is altruism. Fraternity alone associates individual happiness with the happiness of others, affording the promise of self-sustainment.

This crystallized for me my problem with Kurzweil's dream. A technological approach to Eternity—near immortality through robotics—may not be the most desirable utopia, and its pursuit brings clear dangers. Maybe we should rethink our utopian choices.

Where can we look for a new ethical basis to set our course? I have found the ideas in the book Ethics for the New Millennium, by the Dalai Lama, to be very helpful. As is perhaps well known but little heeded, the Dalai Lama argues that the most important thing is for us to conduct our lives with love and compassion for

others, and that our societies need to develop a stronger notion of universal responsibility and of our interdependency; he proposes a standard of positive ethical conduct for individuals and societies that seems consonant with Attali's Fraternity utopia.

The Dalai Lama further argues that we must understand what it is that makes people happy, and acknowledge the strong evidence that neither material progress nor the pursuit of the power of knowledge is the key—that there are limits to what science and the scientific pursuit alone can do.

Our Western notion of happiness seems to come from the Greeks, who defined it as "the exercise of vital powers along lines of excellence in a life affording them scope."[15]

Clearly, we need to find meaningful challenges and sufficient scope in our lives if we are to be happy in whatever is to come. But I believe we must find alternative outlets for our creative forces, beyond the culture of perpetual economic growth; this growth has largely been a blessing for several hundred years, but it has not brought us unalloyed happiness, and we must now choose between the pursuit of unrestricted and undirected growth through science and technology and the clear accompanying dangers.

It is now more than a year since my first encounter with Ray Kurzweil and John Searle. I see around me cause for hope in the voices for caution and relinquishment and in those people I have discovered who are as concerned as I am about our current predicament. I feel, too, a deepened sense of personal responsibility—not for the work I have already done, but for the work that I might yet do, at the confluence of the sciences.

But many other people who know about the dangers still seem strangely silent. When pressed, they trot out the "this is nothing new" riposte—as if awareness of what could happen is response enough. They tell me, There are universities filled with bioethicists who study this stuff all day long. They say, All this has been written about before, and by experts. They complain, Your worries and your arguments are already old hat.

I don't know where these people hide their fear. As an architect of complex systems I enter this arena as a generalist. But should this diminish my concerns? I am aware of how much has been written about, talked about, and lectured about so authoritatively. But does this mean it has reached people? Does this mean we can discount the dangers before us?

Knowing is not a rationale for not acting. Can we doubt that knowledge has become a weapon we wield against ourselves?

The experiences of the atomic scientists clearly show the need to take personal responsibility, the danger that things will move too fast, and the way in which a process can take on a life of its own. We can, as they did, create insurmountable

problems in almost no time flat. We must do more thinking up front if we are not to be similarly surprised and shocked by the consequences of our inventions.

My continuing professional work is on improving the reliability of software. Software is a tool, and as a toolbuilder I must struggle with the uses to which the tools I make are put. I have always believed that making software more reliable, given its many uses, will make the world a safer and better place; if I were to come to believe the opposite, then I would be morally obligated to stop this work. I can now imagine such a day may come.

This all leaves me not angry but at least a bit melancholic. Henceforth, for me, progress will be somewhat bittersweet.

Do you remember the beautiful penultimate scene in Manhattan where Woody Allen is lying on his couch and talking into a tape recorder? He is writing a short story about people who are creating unnecessary, neurotic problems for themselves, because it keeps them from dealing with more unsolvable, terrifying problems about the universe.

He leads himself to the question, "Why is life worth living?" and to consider what makes it worthwhile for him: Groucho Marx, Willie Mays, the second movement of the Jupiter Symphony, Louis Armstrong's recording of "Potato Head Blues," Swedish movies, Flaubert's Sentimental Education, Marlon Brando, Frank Sinatra, the apples and pears by Cézanne, the crabs at Sam Wo's, and, finally, the showstopper: his love Tracy's face.

Each of us has our precious things, and as we care for them we locate the essence of our humanity. In the end, it is because of our great capacity for caring that I remain optimistic we will confront the dangerous issues now before us.

My immediate hope is to participate in a much larger discussion of the issues raised here, with people from many different backgrounds, in settings not predisposed to fear or favor technology for its own sake.

As a start, I have twice raised many of these issues at events sponsored by the Aspen Institute and have separately proposed that the American Academy of Arts and Sciences take them up as an extension of its work with the Pugwash Conferences. (These have been held since 1957 to discuss arms control, especially of nuclear weapons, and to formulate workable policies.)

It's unfortunate that the Pugwash meetings started only well after the nuclear genie was out of the bottle—roughly 15 years too late. We are also getting a belated start on seriously addressing the issues around 21st-century technologies—the prevention of knowledge-enabled mass destruction—and further delay seems unacceptable.

So I'm still searching; there are many more things to learn. Whether we are to succeed or fail, to survive or fall victim to these technologies, is not yet decided. I'm up late again—it's almost 6 AM. I'm trying to imagine some better answers, to break

the spell and free them from the stone.

ENDNOTES

1. The passage Kurzweil quotes is from Kaczynski's Unabomber Manifesto, which was published jointly, under duress, by The New York Times and The Washington Post to attempt to bring his campaign of terror to an end. I agree with David Gelernter, who said about their decision:

   "It was a tough call for the newspapers. To say yes would be giving in to terrorism, and for all they knew he was lying anyway. On the other hand, to say yes might stop the killing. There was also a chance that someone would read the tract and get a hunch about the author; and that is exactly what happened. The suspect's brother read it, and it rang a bell."
   "I would have told them not to publish. I'm glad they didn't ask me. I guess."
   (Drawing Life: Surviving the Unabomber. Free Press, 1997: 120.)

2. Garrett, Laurie. The Coming Plague: Newly Emerging Diseases in a World Out of Balance. Penguin, 1994: 47–52, 414, 419, 452.
3. Isaac Asimov described what became the most famous view of ethical rules for robot behavior in his book I, Robot in 1950, in his Three Laws of Robotics: (1) A robot may not injure a human being, or, through inaction, allow a human being to come to harm. (2) A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law. (3) A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.
4. Michelangelo wrote a sonnet that begins:

   Non ha l' ottimo artista alcun concetto *
   Ch' un marmo solo in sè non circonscriva
   Col suo soverchio; e solo a quello arriva
   La man che ubbidisce all' intelleto.

   Stone translates this as:

   The best of artists hath no thought to show *
   which the rough stone in its superfluous shell
   doth not include; to break the marble spell
   is all the hand that serves the brain can do.

   Stone describes the process: "He was not working from his drawings or clay models; they had all been put away. He was carving from the images in his mind. His eyes and hands knew where every line, curve, mass must emerge, and at what depth in the heart of the stone to create the low relief."
   (The Agony and the Ecstasy. Doubleday, 1961: 6, 144.)
5. First Foresight Conference on Nanotechnology in October 1989, a talk titled "The Future of Computation." Published in Crandall, B. C. and James Lewis, editors. Nanotechnology: Research and Perspectives. MIT Press, 1992: 269. See also www.foresight.org/Conferences/MNT01/Nano1.html.
6. In his 1963 novel Cat's Cradle, Kurt Vonnegut imagined a gray-goo-like accident where a form of ice called ice-nine, which becomes solid at a much higher temperature, freezes

the oceans.

7. Kauffman, Stuart. "Self-replication: Even Peptides Do It." Nature, 382, August 8, 1996: 496. See www.santafe.edu/sfi/People/kauffman/sak-peptides.html.
8. Else, Jon. The Day After Trinity: J. Robert Oppenheimer and The Atomic Bomb (available at www.pyramiddirect.com).
9. This estimate is in Leslie's book The End of the World: The Science and Ethics of Human Extinction, where he notes that the probability of extinction is substantially higher if we accept Brandon Carter's Doomsday Argument, which is, briefly, that "we ought to have some reluctance to believe that we are very exceptionally early, for instance in the earliest 0.001 percent, among all humans who will ever have lived. This would be some reason for thinking that humankind will not survive for many more centuries, let alone colonize the galaxy. Carter's doomsday argument doesn't generate any risk estimates just by itself. It is an argument for revising the estimates which we generate when we consider various possible dangers." (Routledge, 1996: 1, 3, 145).
10. Clarke, Arthur C. "Presidents, Experts, and Asteroids." Science, June 5, 1998. Reprinted as "Science and Society" in Greetings, Carbon-Based Bipeds! Collected Essays, 1934–1998. St. Martin's Press, 1999: 526.
11. And, as David Forrest suggests in his paper "Regulating Nanotechnology Development," available at www.foresight.org/NanoRev/Forrest1989.html, "If we used strict liability as an alternative to regulation it would be impossible for any developer to internalize the cost of the risk (destruction of the biosphere), so theoretically the activity of developing nanotechnology should never be undertaken." Forrest's analysis leaves us with only government regulation to protect us—not a comforting thought.
12. Meselson, Matthew. "The Problem of Biological Weapons." Presentation to the 1,818th Stated Meeting of the American Academy of Arts and Sciences, January 13, 1999. (minerva.amacad.org/archive/bulletin4.htm)
13. Doty, Paul. "The Forgotten Menace: Nuclear Weapons Stockpiles Still Represent the Biggest Threat to Civilization." Nature, 402, December 9, 1999: 583.
14. See also Hans Bethe's 1997 letter to President Clinton, at www.fas.org/bethecr.htm.
15. Hamilton, Edith. The Greek Way. W. W. Norton & Co., 1942: 35.

---

2 The Deeply Intertwined Promise and Peril of GNR

Ray Kurzweil

CONTENTS

> Environmentalists must now grapple squarely with the idea of a world that has enough wealth and enough technological capability, and should not pursue more.
>
> **—Bill McKibben, environmentalist who first wrote about global warming**[1]

> Progress might have been all right once, but it has gone on too long.
>
> **—Ogden Nash (1902–1971)**

In the late 1960s I was transformed into a radical environmental activist. A ragtag group of activists and I sailed a leaky old halibut boat across the North Pacific to block the last hydrogen bomb tests under President Nixon. In the process I co-founded Greenpeace … Environmentalists were often able to produce arguments that sounded reasonable, while doing good deeds like saving whales and making the air and water cleaner. But now the chickens have come home to roost. The environmentalists' campaign against biotechnology in general, and genetic engineering in particular, has clearly exposed their intellectual and moral bankruptcy. By adopting a zero tolerance policy toward a technology with so many potential benefits for humankind and the environment, they … have alienated themselves from scientists, intellectuals, and

internationalists. It seems inevitable that the media and the public will, in time, see the insanity of their position.

—Patrick Moore

I think that … flight from and hatred of technology is self-defeating. The Buddha rests quite as comfortably in the circuits of a digital computer and the gears of a cycle transmission as he does at the top of a mountain or in the petals of a flower. To think otherwise is to demean the Buddha—which is to demean oneself.

—Robert M. Pirsig, Zen and the Art of Motorcycle Maintenance

Consider these articles we'd rather not see available on the Web:

- Impress Your Enemies: How to Build Your Own Atomic Bomb from Readily Available Materials[2]
- How to Modify the Influenza Virus in Your College Laboratory to Release Snake Venom
- Ten Easy Modifications to the *E. coli* Virus
- How to Modify Smallpox to Counteract the Smallpox Vaccine
- Build Your Own Chemical Weapons from Materials Available on the Internet
- How to Build a Pilotless, Self-Guiding, Low-Flying Airplane Using a Low-cost Aircraft, GPS, and a Notebook Computer

Or, how about the following:

- The Genomes of Ten Leading Pathogens
- The Floor Plans of Leading Skyscrapers
- The Layout of U.S. Nuclear Reactors
- The Hundred Top Vulnerabilities of Modern Society
- The Top Ten Vulnerabilities of the Internet
- Personal Health Information on One Hundred Million Americans
- The Customer Lists of Top Pornography Sites

Anyone posting the first item above is almost certain to get a quick visit from the FBI, as did Nate Ciccolo, a fifteen-year-old high school student, in March 2000. For a school science project he built a papier-mâché model of an atomic bomb that turned out to be disturbingly accurate. In the ensuing media storm Ciccolo told ABC News, "Someone just sort of mentioned, you know, you can go on the Internet now and get information. And I, sort of, wasn't exactly up to date on things. Try it. I went on there and a couple of clicks and I was right there."[3]

Of course Ciccolo didn't possess the key ingredient, plutonium, nor did he have

any intention of acquiring it, but the report created shock waves in the media, not to mention among the authorities who worry about nuclear proliferation. Ciccolo had reported finding 563 Web pages on atomic-bomb designs, and the publicity resulted in an urgent effort to remove them. Unfortunately, trying to get rid of information on the Internet is akin to trying to sweep back the ocean with a broom. Some of the sites continue to be easily accessible today. I won't provide any URLs in this book, but they are not hard to find.

Although the article titles above are fictitious, one can find extensive information on the Internet about all of these topics.[4] The Web is an extraordinary research tool. In my own experience, research that used to require a half day at the library can now be accomplished typically in a couple of minutes or less. This has enormous and obvious benefits for advancing beneficial technologies, but it can also empower those whose values are inimical to the mainstream of society. So are we in danger? The answer is clearly yes. How much danger, and what to do about it, are the subjects of this chapter.

My urgent concern with this issue dates back at least a couple of decades. When I wrote *The Age of Intelligent Machines* in the mid-1980s, I was deeply concerned with the ability of then-emerging genetic engineering to enable those skilled in the art and with access to fairly widely available equipment to modify bacterial and viral pathogens to create new diseases.[5] In destructive or merely careless hands these engineered pathogens could potentially combine a high degree of communicability, stealthiness, and destructiveness.

Such efforts were not easy to carry out in the 1980s but were nonetheless feasible. We now know that bioweapons programs in the Soviet Union and elsewhere were doing exactly this.[6] At the time I made a conscious decision to not talk about this specter in my book, feeling that I did not want to give the wrong people any destructive ideas. I didn't want to turn on the radio one day and hear about a disaster, with the perpetrators saying that they got the idea from Ray Kurzweil.

Partly as a result of this decision I faced some reasonable criticism that the book emphasized the benefits of future technology while ignoring its pitfalls. When I wrote *The Age of Spiritual Machines* in 1997–1998, therefore, I attempted to account for both promise and peril.[7] There had been sufficient public attention by that time (for example, the 1995 movie *Outbreak*, which portrays the terror and panic from the release of a new viral pathogen) that I felt comfortable to begin to address the issue publicly.

In September 1998, having just completed the manuscript, I ran into Bill Joy, an esteemed and longtime colleague in the high-technology world, in a bar in Lake Tahoe. Although I had long admired Joy for his work in pioneering the leading software language for interactive Web systems (Java) and having cofounded Sun Microsystems, my focus at this brief get-together was not on Joy but rather on the

third person sitting in our small booth, John Searle. Searle, the eminent philosopher from the University of California at Berkeley, had built a career of defending the deep mysteries of human consciousness from apparent attack by materialists such as Ray Kurzweil ….

Searle and I had just finished debating the issue of whether a machine could be conscious during the closing session of George Gilder's Telecosm conference. The session was entitled "Spiritual Machines" and was devoted to a discussion of the philosophical implications of my upcoming book. I had given Joy a preliminary manuscript and tried to bring him up to speed on the debate about consciousness that Searle and I were having.

As it turned out Joy focused on a completely different issue, specifically the impending dangers to human civilization from three emerging technologies I had presented in the book: genetics, nanotechnology, and robotics (GNR, as discussed earlier). My discussion of the downsides of future technology alarmed Joy, as he would later relate in his now-famous cover story for *Wired,* "Why the Future Doesn't Need Us."[8] In the article Joy describes how he asked his friends in the scientific and technology community whether the projections I was making were credible and was dismayed to discover how close these capabilities were to realization.

Needless to say Joy's article focused entirely on the downside scenarios and created a firestorm. Here was one of the technology world's leading figures addressing new and dire emerging dangers from future technology. It was reminiscent of the attention that George Soros, the currency arbitrageur and archcapitalist, received when he made vaguely critical comments about the excesses of unrestrained capitalism, although the Joy controversy became far more intense. *The New York Times* reported there were about ten thousand articles commenting on and discussing Joy's article, more than any other in the history of commentary on technology issues. My attempt to relax in a Lake Tahoe lounge thus ended up fostering two long-term debates, as my dialogue with John Searle has also continued to this day.

Despite my being the origin of Joy's concern, my reputation as a "technology optimist" has remained intact, and Joy and I have been invited to a variety of forums to debate the peril and promise, respectively, of future technologies. Although I am expected to take up the "promise" side of the debate, I often end up spending most of my time defending his position on the feasibility of these dangers.

Although I share the concerns that Joy expressed about the dangers of technology, we have had key differences on the best way to defend against these dangers. In his article Joy advocates relinquishment—not of every technology but only "dangerous ones," like nanotechnology. As I discuss in greater detail below such broadly defined relinquishment would be impossible to achieve without essentially

relinquishing all technology. That in turn would require a *Brave New World* style of totalitarian government, banning all technology development. Not only would such a solution be inconsistent with our democratic values, but it would actually make the dangers worse by driving the technology underground, where only the least responsible practitioners (for example, terrorists) would have most of the expertise. Joy's position has evolved, and he is now working as a venture capitalist with the legendary silicon valley firm of Kleiner, Perkins, Caufield & Byers investing in technologies such as nanotechnology applied to renewable energy and other natural resources. He remains committed to finding solutions to the dangers he has articulated, a goal we clearly share.

INTERTWINED BENEFITS ...

> It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way.
>
> **—Charles Dickens, A Tale of Two Cities**

> It's like arguing in favor of the plough. You know some people are going to argue against it, but you also know it's going to exist.
>
> **—James Hughes**
>
> *secretary of the Transhumanist Association and sociologist at Trinity College, in a debate, "Should Humans Welcome or Resist Becoming Posthuman?"*

Technology has always been a mixed blessing, bringing us benefits such as longer and healthier lifespans, freedom from physical and mental drudgery, and many novel creative possibilities on the one hand, while introducing new dangers. Technology empowers both our creative and destructive natures.

Substantial portions of our species have already experienced alleviation of the poverty, disease, hard labor, and misfortune that have characterized much of human history. Many of us now have the opportunity to gain satisfaction and meaning from our work, rather than merely toiling to survive. We have ever more powerful tools to express ourselves. With the web now reaching deeply into less developed regions of the world, we will see major strides in the availability of high quality education and medical knowledge. We can share culture, art, and humankind's exponentially expanding knowledge base worldwide. …

We've gone from about twenty democracies in the world after World War II to over one hundred today largely through the influence of decentralized electronic communication. The biggest wave of democratization, including the fall of the iron curtain, occurred during the 1990s with the growth of the Internet and related technologies. There is, of course, a great deal more to accomplish in each of these

areas.

Bioengineering is in the early stages of making enormous strides in reversing disease and aging processes. Ubiquitous N and R are two to three decades away and will continue an exponential expansion of these benefits. … [T]hese technologies will create extraordinary wealth, thereby overcoming poverty and enabling us to provide for all of our material needs by transforming inexpensive raw materials and information into any type of product.

We will spend increasing portions of our time in virtual environments and will be able to have any type of desired experience with anyone, real or simulated, in virtual reality. Nanotechnology will bring a similar ability to morph the physical world to our needs and desires. Lingering problems from our waning industrial age will be overcome. We will be able to reverse remaining environmental destruction. Nanoengineered fuel cells and solar cells will provide clean energy. Nanobots in our physical bodies will destroy pathogens, remove debris such as misformed proteins and protofibrils, repair DNA, and reverse aging. We will be able to redesign all of the systems in our bodies and brains to be far more capable and durable.

Most significant will be the merger of biological and nonbiological intelligence, although nonbiological intelligence will quickly come to predominate. There will be a vast expansion of the concept of what it means to be human. We will greatly enhance our ability to create and appreciate all forms of knowledge from science to the arts, while extending our ability to relate to our environment and one another.

On the other hand,…

… AND DANGERS

> "Plants" with "leaves" no more efficient than today's solar cells could out-compete real plants, crowding the biosphere with an inedible foliage. Tough omnivorous "bacteria" could out-compete real bacteria: They could spread like blowing pollen, replicated swiftly, and reduce the biosphere to dust in a matter of days. Dangerous replicators could easily be too tough, small, and rapidly spreading to stop—at least if we make no preparation. We have trouble enough controlling viruses and fruit flies.
>
> **—Eric Drexler**

As well as its many remarkable accomplishments the twentieth century saw technology's awesome ability to amplify our destructive nature, from Stalin's tanks to Hitler's trains. The tragic event of September 11, 2001, is another example of technologies (jets and buildings) taken over by people with agendas of destruction. We still live today with a sufficient number of nuclear weapons (not all of which are accounted for) to end all mammalian life on the planet.

Beginning in the 1980s the means and knowledge have existed in a routine college bioengineering lab to create unfriendly pathogens potentially more dangerous than nuclear weapons.[9] In a war-game simulation conducted at Johns

Hopkins University called "Dark Winter," it was estimated that an intentional introduction of conventional smallpox in three U.S. cities could result in one million deaths. If the virus were bioengineered to defeat the existing smallpox vaccine, the results could be far worse.[10] The reality of this specter was made clear by a 2001 experiment in Australia in which the mousepox virus was inadvertently modified with genes that altered the immune-system response. The mousepox vaccine was powerless to stop this altered virus.[11] These dangers resonate in our historical memories. Bubonic plague killed one third of the European population. More recently the 1918 flu killed twenty million people worldwide.[12]

Will such threats prevent the ongoing acceleration of the power, efficiency, and intelligence of complex systems (such as humans and our technology)? The past record of complexity increase on this planet has shown a smooth acceleration, even through a long history of catastrophes, both internally generated and externally imposed. This is true of both biological evolution (which faced calamities such as encounters with large asteroids and meteors) and human history (which has been punctuated by an ongoing series of major wars).

However, I believe we can take some encouragement from the effectiveness of the world's response to the SARS (Severe Acute Respiratory Syndrome) virus. Although the possibility of an even more virulent return of SARS remains uncertain as of the writing of this book, it appears that containment measures have been relatively successful and have prevented this tragic outbreak from becoming a true catastrophe. Part of the response involved ancient, low-tech tools such as quarantine and face masks.

However, this approach would not have worked without advanced tools that have only recently become available. Researchers were able to sequence the DNA of the SARS virus within thirty-one days of the outbreak—compared to fifteen years for HIV. That enabled the rapid development of an effective test so that carriers could be quickly identified. Moreover, instantaneous global communication facilitated a coordinated response worldwide, a feat not possible when viruses ravaged the world in ancient times.

As technology accelerates toward the full realization of GNR, we will see the same intertwined potentials: a feast of creativity resulting from human intelligence expanded manyfold, combined with many grave new dangers. A quintessential concern that has received considerable attention is unrestrained nanobot replication. Nanobot technology requires trillions of such intelligently designed devices to be useful. To scale up to such levels it will be necessary to enable them to self-replicate, essentially the same approach used in the biological world (that's how one fertilized egg cell becomes the trillions of cells in a human). And in the same way that biological self-replication gone awry (i.e., cancer) results in biological destruction, a defect in the mechanism curtailing nanobot self-replication—the so-

called gray-goo scenario—would endanger all physical entities, biological or otherwise.

Living creatures—including humans—would be the primary victims of an exponentially spreading nanobot attack. The principal designs for nanobot construction use carbon as a primary building block. Because of carbon's unique ability to form four-way bonds, it is an ideal building block for molecular assemblies. Carbon molecules can form straight chains, zigzags, rings, nanotubes (hexagonal arrays formed in tubes), sheets, buckyballs (arrays of hexagons and pentagons formed into spheres), and a variety of other shapes. Because biology has made the same use of carbon, pathological nanobots would find the Earth's biomass an ideal source of this primary ingredient. Biological entities can also provide stored energy in the form of glucose and ATP.[13] Useful trace elements such as oxygen, sulfur, iron, calcium, and others are also available in the biomass.

How long would it take an out-of-control replicating nanobot to destroy the Earth's biomass? The biomass has on the order of 10^{45} carbon atoms.[14] A reasonable estimate of the number of carbon atoms in a single replicating nanobot is about 10^6. (Note that this analysis is not very sensitive to the accuracy of these figures, only to the approximate order of magnitude.) This malevolent nanobot would need to create on the order of 10^{34} copies of itself to replace the biomass, which could be accomplished with 113 replications (each of which would potentially double the destroyed biomass). Rob Freitas has estimated a minimum replication time of approximately one hundred seconds, so 113 replication cycles would require about three hours.[15] However, the actual rate of destruction would be slower because biomass is not "efficiently" laid out. The limiting factor would be the actual movement of the front of destruction. Nanobots cannot travel very quickly because of their small size. It's likely to take weeks for such a destructive process to circle the globe.

Based on this observation we can envision a more insidious possibility. In a two-phased attack, the nanobots take several weeks to spread throughout the biomass but use up an insignificant portion of the carbon atoms, say one out of every thousand trillion (10^{15}). At this extremely low level of concentration the nanobots would be as stealthy as possible. Then, at an "optimal" point, the second phase would begin with the seed nanobots expanding rapidly in place to destroy the biomass. For each seed nanobot to multiply itself a thousand trillionfold would require only about fifty binary replications, or about ninety minutes. With the nanobots having already spread out in position throughout the biomass, movement of the destructive wave front would no longer be a limiting factor.

The point is that without defenses, the available biomass could be destroyed by gray goo very rapidly. As I discuss below, we will clearly need a nanotechnology immune system in place before these scenarios become a possibility. This immune

system would have to be capable of contending not just with obvious destruction but any potentially dangerous (stealthy) replication, even at very low concentrations.

Mike Treder and Chris Phoenix—executive director and director of research of the Center for Responsible Nanotechnology, respectively. Eric Drexler, Robert Freitas, Ralph Merkle, and others have pointed out that future MNT manufacturing devices can be created with safeguards that would prevent the creation of self-replicating nanodevices.[16] I discuss some of these strategies below. However, this observation, although important, does not eliminate the specter of gray goo. There are other reasons (beyond manufacturing) that self-replicating nanobots will need to be created. The nanotechnology immune system mentioned above, for example, will ultimately require self-replication, otherwise it would be unable to defend us. Self-replication will also be necessary for nanobots to rapidly expand intelligence beyond the Earth …. It is also likely to find extensive military applications. Moreover, safeguards against unwanted self-replication, such as the broadcast architecture described below, can be defeated by a determined adversary or terrorist.

Freitas has identified a number of other disastrous nanobot scenarios.[17] In what he calls the "gray plankton" scenario malicious nanobots would use underwater carbon stored as CH_4 (methane) as well as CO_2 dissolved in seawater. These ocean-based sources can provide about ten times as much carbon as Earth's biomass. In his "gray dust" scenario replicating nanobots use basic elements available in airborne dust and sunlight for power. The "gray lichens" scenario involves using carbon and other elements on rocks.

A PANOPLY OF EXISTENTIAL RISKS

> If a little knowledge is dangerous, where is a person who has so much as to be out of danger?
>
> **—Thomas Henry**

I discuss below steps we can take to address these grave risks, but we cannot have complete assurance in any strategy that we devise today. These risks are what Nick Bostrom calls "existential risks," which he defines as the dangers in the upper right quandrant of the following table:[18]

Bostrom's Categorization of Risks

		Intensity of Risk	
		Moderate	**Profound**
	Global	Ozone Thinning	***Existential Risks***
Scope	**Local**	Recession	Genocide
	Personal	Stolen Car	Death
		Endurable	**Terminal**

Biological life on Earth encountered a human made existential risk for the first time in the middle of the twentieth century with the advent of the hydrogen bomb and the subsequent cold-war buildup of thermonuclear forces. President Kennedy reportedly estimated that the likelihood of an all-out nuclear war during the Cuban missile crisis was between 33 and 50 percent.[19] The legendary information theorist John von Neumann, who became the chairman of the Air Force Strategic Missiles Evaluation Committee and a government adviser on nuclear strategies, estimated the likelihood of nuclear Armageddon (prior to the Cuban missile crisis) at close to 100 percent.[20] Given the perspective of the 1960s what informed observer of those times would have predicted that the world would have gone through the next forty years without another nontest nuclear explosion?

Despite the apparent chaos of international affairs we can be grateful for the successful avoidance thus far of the employment of nuclear weapons in war. But we clearly cannot rest easily, since enough hydrogen bombs still exist to destroy all human life many times over.[21] Although attracting relatively little public discussion, the massive opposing ICBM arsenals of the United States and Russia remain in place, despite the apparent thawing of relations.

Nuclear proliferation and the widespread availability of nuclear materials and know-how is another grave concern, although not an existential one for our civilization. (that is, only an all-out thermonuclear war involving the ICBM arsenals poses a risk to survival of all humans.) Nuclear proliferation and nuclear terrorism belong to the "profound-local" category of risk, along with genocide. However, the concern is certainly severe because the logic of mutual assured destruction does not work in the context of suicide terrorists.

Debatably we've now added another existential risk, which is the possibility of a bioengineered virus that spreads easily, has a long incubation period, and delivers an ultimately deadly payload. Some viruses are easily communicable, such as the flu and common cold. Others are deadly, such as HIV. It is rare for a virus to combine both attributes. Humans living today are descendants of those who developed natural

immunities to most of the highly communicable viruses. The ability of the species to survive viral outbreaks is one advantage of sexual reproduction, which tends to ensure genetic diversity in the population, so that the response to specific viral agents is highly variable. Although catastrophic, bubonic plague did not kill everyone in Europe. Other viruses, such as smallpox, have both negative characteristics—they are easily contagious and deadly—but have been around long enough that there has been time for society to create a technological protection in the form of a vaccine. Gene engineering, however, has the potential to bypass these evolutionary protections by suddenly introducing new pathogens for which we have no protection, natural or technological.

The prospect of adding genes for deadly toxins to easily transmitted, common viruses such as the common cold and flu introduced another possible existential-risk scenario. It was this prospect that led to the Asilomar conference to consider how to deal with such a threat and the subsequent drafting of a set of safety and ethics guidelines. Although these guidelines have worked thus far, the underlying technologies for genetic manipulation are growing rapidly in sophistication.

In 2003 the world struggled, successfully, with the SARS virus. The emergence of SARS resulted from a combination of an ancient practice (the virus is suspected of having jumped from exotic animals, possibly civet cats, to humans living in close proximity) and a modern practice (the infection spread rapidly across the world by air travel). SARS provided us with a dry run of a virus new to human civilization that combined easy transmission, the ability to survive for extended periods of time outside the human body, and a high degree of mortality, with death rates estimated at 14 to 20 percent. Again, the response combined ancient and modern techniques.

Our experience with SARS shows that most viruses, even if relatively easily transmitted and reasonably deadly, represent grave but not necessarily existential risks. SARS, however, does not appear to have been engineered. SARS spreads easily through externally transmitted bodily fluids but is not easily spread through airborne particles. Its incubation period is estimated to range from one day to two weeks, whereas a longer incubation period would allow a virus to spread through several exponentially growing generations before carriers are identified.[22]

SARS is deadly, but the majority of its victims do survive. It continues to be feasible for a virus to be malevolently engineered so it spreads more easily than SARS, has an extended incubation period, and is deadly to essentially all victims. Smallpox is close to having these characteristics. Although we have a vaccine (albeit a crude one), the vaccine would not be effective against genetically modified versions of the virus.

As I describe below, the window of malicious opportunity for bioengineered viruses, existential or otherwise, will close in the 2020s when we have fully effective antiviral technologies based on nanobots.[23] However, because

nanotechnology will be thousands of times stronger, faster, and more intelligent than biological entities, self-replicating nanobots will present a greater risk and yet another existential risk. The window for malevolent nanobots will ultimately be closed by strong artificial intelligence, but, not surprisingly, "unfriendly" AI will itself present an even more compelling existential risk, which I discuss below.

The precautionary principle. As Bostrom, Freitas, and other observers including myself have pointed out, we cannot rely on trial-and-error approaches to deal with existential risks. There are competing interpretations of what has become known as the "precautionary principle." (If the consequences of an action are unknown but judged by some scientists to have even a small risk of being profoundly negative, it's better to not carry out the action than risk negative consequences.) But it's clear that we need to achieve the highest possible level of confidence in our strategies to combat such risks. This is one reason we're hearing increasingly strident voices demanding that we shut down the advance of technology, as a primary strategy to eliminate new existential risks before they occur. Relinquishment, however, is not the appropriate response and will only interfere with the profound benefits of these emerging technologies while actually increasing the likelihood of a disastrous outcome. Max More articulates the limitations of the precautionary principle and advocates replacing it with what he calls the "proactionary principle," which involves balancing the risks of action and inaction.[24]

Before discussing how to respond to the new challenge of existential risks, it's worth reviewing a few more that have been postulated by Bostrom and others.

The smaller the interaction, the larger the explosive potential. There has been recent controversy over the potential for future very high-energy particle accelerators to create a chain reaction of transformed energy states at a subatomic level. The result could be an exponentially spreading area of destruction, breaking apart all atoms in our galactic vicinity. A variety of such scenarios has been proposed, including the possibility of creating a black hole that would draw in our solar system.

Analyses of these scenarios show them to be very unlikely, although not all physicists are sanguine about the danger.[25] The mathematics of these analyses appears to be sound, but we do not yet have a consensus on the formulas that describe this level of physical reality. If such dangers sound far-fetched, consider the possibility that we have indeed detected increasingly powerful explosive phenomena at diminishing scales of matter.

Alfred Nobel discovered dynamite by probing chemical interactions of molecules. The atomic bomb, which is tens of thousands of times more powerful than dynamite, is based on nuclear interactions involving large atoms, which are much smaller scales of matter than large molecules. The hydrogen bomb, which is thousands of times more powerful than an atomic bomb, is based on interactions involving an

even smaller scale: small atoms. Although this insight does not necessarily imply the existence of yet more powerful destructive chain reactions by manipulating subatomic particles, it does make the conjecture plausible.

My own assessment of this danger is that we are unlikely to simply stumble across such a destructive event. Consider how unlikely it would be to accidentally produce an atomic bomb. Such a device requires a precise configuration of materials and actions, and the original required an extensive and precise engineering project to develop. Inadvertently creating a hydrogen bomb would be even less plausible. One would have to create the precise conditions of an atomic bomb in a particular arrangement with a hydrogen core and other elements. Stumbling across the exact conditions to create a new class of catastrophic chain reaction at a subatomic level appears to be even less likely. The consequences are sufficiently devastating, however, that the precautionary principle should lead us to take these possibilities seriously. This potential should be carefully analyzed prior to carrying out new classes of accelerator experiments. However, this risk is not high on my list of twenty-first-century concerns.

Our simulation is turned off. Another existential risk that Bostrom and others have identified is that we're actually living in a simulation and the simulation will be shut down. It might appear that there's not a lot we could do to influence this. However, since we're the subject of the simulation, we do have the opportunity to shape what happens inside of it. The best way we could avoid being shut down would be to be interesting to the observers of the simulation. Assuming that someone is actually paying attention to the simulation, it's a fair assumption that it's less likely to be turned off when it's compelling than otherwise.

We could spend a lot of time considering what it means for a simulation to be interesting, but the creation of new knowledge would be a critical part of this assessment. Although it may be difficult for us to conjecture what would be interesting to our hypothesized simulation observer, it would seem that the Singularity is likely to be about as absorbing as any development we could imagine and would create new knowledge at a extraordinary rate. Indeed, achieving a Singularity of exploding knowledge may be the very purpose of the simulation. Thus, assuring a "constructive" Singularity (one that avoids degenerate outcomes such as existential destruction by gray goo or dominance by a malicious AI) could be the best course to prevent the simulation being terminated. Of course, we have every motivation to achieve a constructive Singularity for many other reasons.

If the world we're living in is a simulation on someone's computer, it's a very good one—so detailed, in fact, that we may as well accept it as our reality. In any event, it is the only reality to which we have access.

Our world appears to have a long and rich history. This means that either our world is not, in fact, a simulation or, if it is, the simulation has been going a very

long time and thus is not likely to stop anytime soon. Of course it is also possible that the simulation includes evidence of a long history without the history's having actually occurred.

… [T]here are conjectures that an advanced civilization may create a new universe to perform computation (or, to put it another way, to continue the expansion of its own computation). Our living in such a universe (created by another civilization) can be considered a simulation scenario. Perhaps this other civilization is running an evolutionary algorithm on our universe (that is, the evolution we're witnessing) to create an explosion of knowledge from a technology singularity. If that is true, then the civilization watching our universe might shut down the simulation if it appeared that a knowledge singularity had gone awry and it did not look like it was going to occur.

This scenario is also not high on my worry list, particularly since the only strategy that we can follow to avoid a negative outcome is the one we need to follow anyway.

Crashing the party. Another oft-cited concern is that of a large-scale asteroid or comet collision, which has occurred repeatedly in the Earth's history, and did represent existential outcomes for species at these times. This is not a peril of technology, of course. Rather, technology will protect us from this risk (certainly within one to a couple of decades). Although small impacts are a regular occurrence, large and destructive visitors from space are rare. We don't see one on the horizon, and it is virtually certain that by the time such a danger occurs, our civilization will readily destroy the intruder before it destroys us.

Another item on the existential danger list is destruction by an alien intelligence (not one that we've created). …

GNR: the proper focus of promise versus peril. This leaves the GNR technologies as the primary concerns. However, I do think we also need to take seriously the misguided and increasingly strident Luddite voices that advocate reliance on broad relinquishment of technological progress to avoid the genuine dangers of GNR. For reasons I discuss below, relinquishment is not the answer, but rational fear could lead to irrational solutions. Delays in overcoming human suffering are still of great consequence, for example the worsening of famine in Africa due to opposition to aid from food using GMO (genetically modified organisms).

Broad relinquishment would require a totalitarian system to implement, and a totalitarian brave new world is unlikely because of the democratizing impact of increasingly powerful decentralized electronic and photonic communication. The advent of worldwide, decentralized communication epitomized by the Internet and cell phones has been a pervasive democratizing force. It was not Boris Yeltsin standing on a tank that overturned the 1991 coup against Mikhail Gorbachev, but

rather the clandestine network of fax machines, photocopiers, video recorders, and personal computers that broke decades of totalitarian control of information.[26] The movement toward democracy and capitalism and the attendant economic growth that characterized the 1990s were all fueled by the accelerating force of these person-to-person communication technologies.

There are other questions that are nonexistential but nonetheless serious. They include "Who is controlling the nanobots?" and "Whom are the nanobots talking to?" Future organizations (whether governments or extremist groups) or just a clever individual could put trillions of undetectable nanobots in the water or food supply of an individual or of an entire population. These spybots could then monitor, influence, and even control thoughts and actions. In addition existing nanobots could be influenced through software viruses and hacking techniques. When there is software running in our bodies and brains (as we discussed, a threshold we have already passed for some people), issues of privacy and security will take on a new urgency, and countersurveillance methods of combating such intrusions will be devised.

The inevitability of a transformed future. The diverse GNR technologies are progressing on many fronts. The full realization of GNR will result from hundreds of small steps forward, each benign in itself. For G we have already passed the threshold of having the means to create designer pathogens. Advances in biotechnology will continue to accelerate, fueled by the compelling ethical and economic benefits that will result from mastering the information processes underlying biology.

Nanotechnology is the inevitable end result of the ongoing miniaturization of technology of all kinds. The key features for a wide range of applications, including electronics, mechanics, energy, and medicine, are shrinking at the rate of a factor of about four per linear dimension per decade. Moreover, there is exponential growth in research seeking to understand nanotechnology and its applications. …

Similarly, our efforts to reverse engineer the human brain are motivated by diverse anticipated benefits, including understanding and reversing cognitive diseases and decline. The tools for peering into the brain are showing exponential gains in spatial and temporal resolution, and we've demonstrated the ability to translate data from brain scans and studies into working models and simulations.

Insights from the brain-reverse-engineering effort, overall research in developing AI algorithms, and ongoing exponential gains in computing platforms make strong AI (AI at human levels and beyond) inevitable. Once AI achieves human levels, it will necessarily soar past it because it will combine the strengths of human intelligence with the speed, memory capacity, and knowledge sharing that nonbiological intelligence already exhibits. Unlike biological intelligence, nonbiological intelligence will also benefit from ongoing exponential gains in scale,

capacity, and price-performance.

Totalitarian relinquishment. The only conceivable way that the accelerating pace of advancement on all of these fronts could be stopped would be through a worldwide totalitarian system that relinquishes the very idea of progress. Even this specter would be likely to fail in averting the dangers of GNR because the resulting underground activity would tend to favor the more destructive applications. This is because the responsible practitioners that we rely on to quickly develop defensive technologies would not have easy access to the needed tools. Fortunately, such a totalitarian outcome is unlikely because the increasing decentralization of knowledge is inherently a democratizing force.

PREPARING THE DEFENSES

My own expectation is that the creative and constructive applications of these technologies will dominate, as I believe they do today. However, we need to vastly increase our investment in developing specific defensive technologies. As I discussed, we are at the critical stage today for biotechnology, and we will reach the stage where we need to directly implement defensive technologies for nanotechnology during the late teen years of this century.

We don't have to look past today to see the intertwined promise and peril of technological advancement. Imagine describing the dangers (atomic and hydrogen bombs for one thing) that exist today to people who lived a couple of hundred years ago. They would think it mad to take such risks. But how many people in 2005 would really want to go back to the short, brutish, disease-filled, poverty-stricken, disaster-prone lives that 99 percent of the human race struggled through a couple of centuries ago?[27]

We may romanticize the past, but up until fairly recently most of humanity lived extremely fragile lives in which one all-too-common misfortune could spell disaster. Two hundred years ago life expectancy for females in the record-holding country (Sweden) was roughly thirty-five years, very brief compared to the longest life expectancy today—almost eighty-five years, for Japanese women. Life expectancy for males was roughly thirty-three years, compared to the current seventy-nine years in the record-holding countries.[28] It took half the day to prepare the evening meal, and hard labor characterized most human activity. There were no social safety nets. Substantial portions of our species still live in this precarious way, which is at least one reason to continue technological progress and the economic enhancement that accompanies it. Only technology, with its ability to provide orders of magnitude of improvement in capability and affordability, has the scale to confront problems such as poverty, disease, pollution, and the other overriding concerns of society today.

People often go through three stages in considering the impact of future technology: awe and wonderment at its potential to overcome age-old problems;

then a sense of dread at a new set of grave dangers that accompany these novel technologies; followed finally by the realization that the only viable and responsible path is to set a careful course that can realize the benefits while managing the dangers.

Needless to say we have already experienced technology's downside, for example death and destruction from war. The crude technologies of the first industrial revolution have crowded out many of the species that existed on our planet a century ago. Our centralized technologies (such as buildings, cities, airplanes, and power plants) are demonstrably insecure.

The "NBC" (nuclear, biological, and chemical) technologies of warfare have all been used or been threatened to be used in our recent past.[29] The far more powerful GNR technologies threaten us with new, profound local and existential risks. If we manage to get past the concerns about genetically altered designer pathogens, followed by self-replicating entities created through nanotechnology, we will encounter robots whose intelligence will rival and ultimately exceed our own. Such robots may make great assistants, but who's to say that we can count on them to remain reliably friendly to mere biological humans?

Strong AI. Strong AI promises to continue the exponential gains of human civilization. (As I discussed earlier, I include the nonbiological intelligence derived from our human civilization as still human.) But the dangers it presents are also profound precisely because of its amplification of intelligence. Intelligence is inherently impossible to control, so the various strategies that have been devised to control nanotechnology (for example, the "broadcast architecture" described below) won't work for strong AI. There have been discussions and proposals to guide AI development towards what Eliezer Yudkowsky calls "friendly AI"[30] (see below). These are useful for discussion, but it is infeasible today to devise strategies today that will absolutely ensure that future AI embodies human ethics and values.

Returning to the past? In his essay and presentations Bill Joy eloquently describes the plagues of centuries past and how new self-replicating technologies, such as mutant bioengineered pathogens and nanobots run amok may bring back long-forgotten pestilence. But as Joy acknowledges, technological advances, such as antibiotics and improved sanitation, have freed us from the prevalence of such plagues. Suffering in the world continues and demands our steadfast attention. Should we tell the millions of people afflicted with cancer and other devastating conditions that we are canceling the development of all bioengineered treatments because there is a risk that these same technologies may someday be used for malevolent purposes? Having posed this rhetorical question, I realize that there is a movement to do exactly that, but most people would agree that such broad-based relinquishment is not the answer.

The continued opportunity to alleviate human distress is one key motivation for

continuing technological advancement. Also compelling are the already apparent economic gains that will continue to hasten in the decades ahead. The ongoing acceleration of many intertwined technologies are roads paved with gold. (I use the plural here because technology is clearly not a single path.) In a competitive environment it is an economic imperative to go down these roads. Relinquishing technological advancement would be economic suicide for individuals, companies, and nations.

THE IDEA OF RELINQUISHMENT

> The major advances in civilization all but wreck the civilizations in which they occur.
>
> **—Alfred North Whitehead**

This brings us to the issue of relinquishment, which is the most controversial recommendation in Bill Joy's Wired magazine article. I do feel that relinquishment at the right level is part of a responsible and constructive response to the genuine perils that we will face in the future. The issue, however, is exactly this: at what level *are* we to relinquish technology?

Ted Kaczynski, who became known to the world as the Unabomber, would have us renounce all of it.[31] This is neither desirable nor feasible, and the futility of such a position is only underscored by the senselessness of Kaczynski's deplorable tactics.

Other voices, less reckless than Kaczynski's, are nonetheless likewise arguing for broad-based relinquishment of technology. Bill McKibben, the environmentalist who was one of the first to warn against global warming, takes the position that we already have sufficient technology and that further progress should end. In his latest book, *Enough: Staying Human in an Engineered Age*, he metaphorically compares technology to beer: "One beer is good, two beers may be better; eight beers, you're almost certainly going to regret."[32] That metaphor misses the point and ignores the extensive suffering that remains in the human world that we can alleviate through sustained scientific advance.

Although new technologies, like anything else, may be used to excess at times, their promise is not just a matter of adding a fourth cell phone or doubling the number of unwanted e-mails. Rather, it means perfecting the technologies to conquer cancer and other devastating diseases, creating ubiquitous wealth to overcome poverty, cleaning up the environment from the effects of the first industrial revolution (an objective articulated by McKibben), and overcoming many other age-old problems.

Broad relinquishment. Another level of relinquishment, one originally recommended by Joy, would be to forgo only certain fields—nanotechnology, for example—that might be regarded as too dangerous. But such sweeping strokes of

relinquishment are equally untenable. As I pointed out above, nanotechnology is simply the inevitable end result of the persistent trend toward miniaturization that pervades all of technology. It is far from a single centralized effort but is being pursued by a myriad of projects with many diverse goals.

One observer wrote:

> A further reason why industrial society cannot be reformed … is that modern technology is a unified system in which all parts are dependent on one another. You can't get rid of the "bad" parts of technology and retain only the "good" parts. Take modern medicine, for example. Progress in medical science depends on progress in chemistry, physics, biology, computer science and other fields. Advanced medical treatments require expensive, high-tech equipment that can be made available only by a technologically progressive, economically rich society. Clearly you can't have much progress in medicine without the whole technological system and everything that goes with it.

The observer I am quoting here is, again, Ted Kaczynski.[33] Although one will properly resist Kaczynski as an authority, I believe he is correct on the deeply entangled nature of the benefits and risks. However, Kaczynski and I clearly part company on our overall assessment of the relative balance between the two. Bill Joy and I have had an ongoing dialogue on this issue both publicly and privately, and we both believe that technology will and should progress and that we need to be actively concerned with its dark side. If Joy and I disagree, it's on the granularity of relinquishment that is both feasible and desirable, although it is an issue we both continue to actively explore.

Fine-grained relinquishment. I do think that relinquishment at the right level needs to be part of our ethical response to the dangers of twenty-first-century technologies. One constructive example of this is the ethical guideline proposed by the Foresight Institute, founded by nanotechnology pioneer Eric Drexler and Christine Peterson: namely, that nanotechnologists agree to relinquish the development of physical entities that can self-replicate in a natural environment.[34] There is one exception to this guideline, however, in that we will ultimately need to provide a nanotechnology-based planetary immune system (nanobots embedded in the natural environment to protect against rogue self-replicating nanobots). Robert Freitas and I have discussed whether or not such an immune system would itself need to be self-replicating. Freitas writes, "A comprehensive surveillance system coupled with prepositioned resources – resources including high-capacity nonreplicating nanofactories able to churn out large numbers of nonreplicating defenders in response to specific threats – should suffice."[35] I agree with Freitas that a prepositioned immune system with the ability to augment the defenders will be sufficient in early stages. But once strong AI is merged with nanotechnology, and the ecology of nanoengineered entities becomes highly varied and complex, my own expectation is that we will find that the defending nanorobots need the ability to

replicate in place quickly.

Another good example of a useful ethical guidelines is a ban on self-replicating physical entities that contain their own codes for self-replication. In what nanotechnologist Ralph Merkle calls the "broadcast architecture," such entities would have to obtain such codes from a centralized secure server, which would guard against undesirable replication.[36] The broadcast architecture is impossible in the biological world, so there's at least one way in which nanotechnology can be made safer than biotechnology. In other ways, nanotech is potentially more dangerous because nanobots can be physically stronger than protein-based entities and more intelligent.

Here's an idea: we can apply a nanotechnology-based broadcast architecture to biology. A nanocomputer would augment or replace the nucleus in every cell and provide the DNA codes. A nanobot that incorporated molecular machinery similar to ribosomes (the molecules that interpret the base pairs in the mRNA outside the nucleus) would take the codes and produce the strings of amino acids. Since we could control the nanocomputer through wireless messages, we would be able to shut off unwanted replication, thereby eliminating cancer. We could produce special proteins as needed to combat disease. And we could correct the DNA errors and upgrade the DNA code. I comment further on the strengths and weaknesses of the broadcast architecture below.

Dealing with abuse. Broad relinquishment is contrary to economic progress and ethically unjustified given the opportunity to alleviate disease, overcome poverty, and clean up the environment. As mentioned above, it would exacerbate the dangers. Regulations on safety—essentially fine-grained relinquishment—will remain appropriate.

However, we also need to streamline the regulatory process. Right now in the United States, we have a five- to ten-year delay on new health technologies for FDA approval (with comparable delays in other nations). The harm caused by holding up potential lifesaving treatments (for example, one million lives lost in the United States for each year we delay treatments for heart disease) is given very little weight against the possible risks of new therapies.

Other protections will need to include oversight by regulatory bodies, the development of technology-specific "immune" responses, as well as computer-assisted surveillance by law-enforcement organizations. Many people are not aware that our intelligence agencies already use advanced technologies such as automated keyword spotting to monitor a substantial flow of telephone, cable, satellite, and Internet conversations. As we go forward, balancing our cherished rights of privacy with our need to be protected from the malicious use of powerful twenty-first-century technologies will be one of many profound challenges. This is one reason such issues as an encryption "trapdoor" (in which law-enforcement authorities

would have access to otherwise secure information) and the FBI's Carnivore e-mail-snooping system have been controversial.[37]

As a test case we can take a small measure of comfort from how we have dealt with one recent technological challenge. There exists today a new fully nonbiological self-replicating entity that didn't exist just a few decades ago: the computer virus. When this form of destructive intruder first appeared, strong concerns were voiced that as they became more sophisticated, software pathogens had the potential to destroy the computer-network medium in which they live. Yet the "immune system" that has evolved in response to this challenge has been largely effective. Although destructive self-replicating software entities do cause damage from time to time, the injury is but a small fraction of the benefit we receive from the computers and communication links that harbor them.

One might counter that computer viruses do not have the lethal potential of biological viruses or of destructive nanotechnology. This is not always the case; we rely on software to operate our 911 call centers, monitor patients in critical-care units, fly and land airplanes, guide intelligent weapons in our military campaigns, handle our financial transactions, operate our municipal utilities, and many other mission-critical tasks. To the extent that software viruses do not yet pose a lethal danger, however, this observation only strengthens my argument. The fact that computer viruses are not usually deadly to humans only means that more people are willing to create and release them. The vast majority of software virus authors would not release viruses if they thought they would kill people. It also means that our response to the danger is that much less intense. Conversely, when it comes to self-replicating entities that are potentially lethal on a large scale, our response on all levels will be vastly more serious.

Although software pathogens remain a concern, the danger exists today mostly at a nuisance level. Keep in mind that our success in combating them has taken place in an industry in which there is no regulation and minimal certification for practitioners. The largely unregulated computer industry is also enormously productive. One could argue that it has contributed more to our technological and economic progress than any other enterprise in human history.

But the battle concerning software viruses and the panoply of software pathogens will never end. We are becoming increasingly reliant on mission-critical software systems, and the sophistication and potential destructiveness of self-replicating software weapons will continue to escalate. When we have software running in our brains and bodies and controlling the world's nanobot immune system, the stakes will be immeasurably greater.

The threat from fundamentalism. The world is struggling with an especially pernicious form of religious fundamentalism in the form of radical Islamic terrorism. Although it may appear that these terrorists have no program other than destruction,

they do have an agenda that goes beyond literal interpretations of ancient scriptures: essentially, to turn the clock back on such modern ideas as democracy, women's rights, and education.

But religious extremism is not the only form of fundamentalism that represents a reactionary force. At the beginning of this chapter I quoted Patrick Moore, cofounder of Greenpeace, on his disillusionment with the movement he helped found. The issue that undermined Moore's support of Greenpeace was its total opposition to Golden Rice, a strain of rice genetically modified to contain high levels of beta-carotene, the precursor to Vitamin A.[38] Hundreds of millions of people in Africa and Asia lack sufficient Vitamin A, with half a million children going blind each year from the deficiency, and millions more contracting other related diseases. About seven ounces a day of Golden Rice would provide 100% of a child's Vitamin A requirement. Extensive studies have shown that this grain, as well as many other genetically modified organisms (GMOs), is safe. For example, in 2001 the European Commission released eighty-one studies that concluded that GMOs have "not shown any new risks to human health or the environment, beyond the usual uncertainties of conventional plant breeding. Indeed, the use of more precise technology and the greater regulatory scrutiny probably make them even safer than conventional plants and foods."[39]

It is not my position that all GMOs are inherently safe; obviously safety testing of each product is needed. But the anti-GMO movement takes the position that every GMO is by its very nature hazardous, a view that has no scientific basis.

The availability of Golden Rice has been delayed by at least five years through the pressure of Greenpeace and other anti-GMO activists. Moore, noting that this delay will cause millions of additional children to go blind, quotes the grain's opponents as threatening "to rip the G.M. rice out of the fields if farmers dare to plant it." Similarly, African nations have been pressured to refuse GMO food aid and genetically modified seeds, thereby worsening conditions of famine.[40] Ultimately the demonstrated ability of technologies such as GMO to solve overwhelming problems will prevail, but the temporary delays caused by irrational opposition will nonetheless result in unnecessary suffering.

Certain segments of the environmental movement have become fundamentalist Luddites—"fundamentalist" because of their misguided attempt to preserve things as they are (or were); "Luddite" because of the reflexive stance against technological solutions to outstanding problems. Ironically it is GMO plants—many of which are designed to resist insects and other forms of blight and thereby require greatly reduced levels of chemicals, if any—that offer the best hope for reversing environmental assault from chemicals such as pesticides.

Actually my characterization of these groups as "fundamentalist Luddites" is redundant, because Ludditism is inherently fundamentalist. It reflects the idea that

humanity will be better off without change, without progress. This brings us back to the idea of relinquishment, as the enthusiasm for relinquishing technology on a broad scale is coming from the same intellectual sources and activist groups that make up the Luddite segment of the environmental movement.

Fundamentalist humanism. With G and N technologies now beginning to modify our bodies and brains, another form of opposition to progress has emerged in the form of "fundamentalist humanism": opposition to any change in the nature of what it means to be human (for example, changing our genes and taking other steps toward radical life extension). This effort, too, will ultimately fail, however, because the demand for therapies that can overcome the suffering, disease, and short lifespans inherent in our version 1.0 bodies will ultimately prove irresistible.

In the end, it is only technology—especially GNR—that will offer the leverage needed to overcome problems that human civilization has struggled with for many generations.

DEVELOPMENT OF DEFENSIVE TECHNOLOGIES AND THE IMPACT OF REGULATION

One of the reasons that Bill Joy's treatise in WIRED was so effective is that the picture he painted of future dangers assumed they will be released in the context of today's unprepared world. The reality is that the sophistication and power of our defensive knowledge and technologies will grow along with the dangers. A phenomenon like gray goo (unrestrained nanobot replication) will be countered with "blue goo" ("police" nanobots that combat the "bad" nanobots). Obviously we cannot say with assurance that we will successfully avert all misuse. But the surest way to prevent development of effective defensive technologies would be to relinquish the pursuit of knowledge in a number of broad areas. We have been able to largely control harmful software-virus replication because the requisite knowledge is widely available to responsible practitioners. Attempts to restrict such knowledge would have given rise to a far less stable situation. Responses to new challenges would have been far slower, and it is likely that the balance would have shifted toward more destructive applications (such as self-modifying software viruses).

If we compare the success we have had in controlling engineered software viruses to the coming challenge of controlling engineered biological viruses, we are struck with one salient difference. As I noted above, the software industry is almost completely unregulated. The same is obviously not true for biotechnology. While a bioterrorist does not need to put his "innovations" through the FDA, we do require the scientists developing defensive technologies to follow existing regulations, which slow down the innovation process at every step. Moreover, under existing regulations and ethical standards, it is impossible to test defenses against bioterrorist

agents. Extensive discussion is already under way to modify these regulations to allow for animal models and simulations to replace unfeasible human trials. This will be necessary, but I believe we will need to go beyond these steps to accelerate the development of vitally needed defensive technologies.

In terms of public policy the task at hand is to rapidly develop the defensive steps needed, which include ethical standards, legal standards, and defensive technologies themselves. It is quite clearly a race. As I noted, in the software field defensive technologies have responded quickly to "innovations" in the offensive ones. In the medical field, in contrast, extensive regulation slows down innovation, so we cannot have the same confidence with regard to the abuse of biotechnology. In the current environment, when one person dies in gene-therapy trials, research can be severely restricted.[41] There is a legitimate need to make biomedical research as safe as possible, but our balancing of risks is completely skewed. Millions of people desperately need the advances promised by gene-therapy and other breakthrough biotechnology advances, but they appear to carry little political weight against a handful of well-publicized casualties from the inevitable risks of progress.

This risk-balancing equation will become even more stark when we consider the emerging dangers of bioengineered pathogens. What is needed is a change in public attitude in tolerance for necessary risk. Hastening defensive technologies is absolutely vital to our security. We need to streamline regulatory procedures to achieve this. At the same time we must greatly increase our investment explicitly in defensive technologies. In the biotechnology field this means the rapid development of antiviral medications. We will not have time to formulate specific countermeasures for each new challenge that comes along. We are close to developing more generalized antiviral technologies, such as RNA interference, and these need to be accelerated.

We're addressing biotechnology here because that is the immediate threshold and challenge that we now face. As the threshold for self-organizing nanotechnology approaches, we will then need to invest specifically in the development of defensive technologies in that area, including the creation of a technological immune system. Consider how our biological immune system works. When the body detects a pathogen the T cells and other immune-system cells self-replicate rapidly to combat the invader. A nanotechnology immune system would work similarly both in the human body and in the environment and would include nanobot sentinels that could detect rogue self-replicating nanobots. When a threat was detected, defensive nanobots capable of destroying the intruders would be rapidly created (eventually with self-replication) to provide an effective defensive force.

Bill Joy and other observers have pointed out that such an immune system would itself be a danger because of the potential of "autoimmune" reactions (that is, the immune-system nanobots attacking the world they are supposed to defend).[42]

However this possibility is not a compelling reason to avoid the creation of an immune system. No one would argue that humans would be better off without an immune system because of the potential of developing autoimmune diseases. Although the immune system can itself present a danger humans would not last more than a few weeks (barring extraordinary efforts at isolation) without one. And even so, the development of a technological immune system for nanotechnology will happen even without explicit efforts to create one. This has effectively happened with regard to software viruses, creating an immune system not through a formal grand-design project but rather through incremental responses to each new challenge and by developing heuristic algorithms for early detection. We can expect the same thing will happen as challenges from nanotechnology-based dangers emerge. The point for public policy will be to specifically invest in these defensive technologies.

It is premature today to develop specific defensive nanotechnologies, since we can now have only a general idea of what we are trying to defend against. However fruitful dialogue and discussion on anticipating this issue is already taking place, and significantly expanded investment in these efforts is to be encouraged. As I mentioned above, the Foresight Institute, as one example, has devised a set of ethical standards and strategies for assuring the development of safe nanotechnology, based on guidelines for biotechnology.[43] When gene-splicing began in 1975 two biologists, Maxine Singer and Paul Berg, suggested a moratorium on the technology until safety concerns could be addressed. It seemed apparent that there was substantial risk if genes for poisons were introduced into pathogens, such as the common cold, that spread easily. After a ten-month moratorium guidelines were agreed to at the Asilomar conference, which included provisions for physical and biological containment, bans on particular types of experiments, and other stipulations. These biotechnology guidelines have been strictly followed, and there have not been reported accidents in the thirty-year history of the field.

More recently, the organization representing the world's organ transplantation surgeons has adopted a moratorium on the transplantation of vascularized animal organs into humans. This was done out of fear of the spread of long-dormancy HIV-type xenoviruses from animals such as pigs or baboons into the human population. Unfortunately, such a moratorium can also slow down the availability of life-saving xenografts (genetically-modified animal organs that are accepted by the human immune system) to the millions of people who die each year from heart, kidney and liver disease. Geoethicist Martine Rothblatt has proposed replacing this moratorium with a new set of ethical guidelines and regulations.[44]

In the case of nanotechnology, the ethics debate has started a couple of decades prior to the availability of the particularly dangerous applications. The most important provisions of the Foresight Institute guidelines include:

- "Artificial replicators must not be capable of replication in a natural,

uncontrolled environment."

- "Evolution within the context of a self-replicating manufacturing system is discouraged."
- "MNT device designs should specifically limit proliferation and provide traceability of any replicating systems."
- "Distribution of molecular manufacturing *development* capability should be restricted whenever possible, to responsible actors that have agreed to use the Guidelines. No such restriction need apply to end products of the development process."

Other strategies that the Foresight Institute has proposed include:

- Replication should require materials not found in the natural environment.
- Manufacturing (replication) should be separated from the functionality of end products. Manufacturing devices can create end products but cannot replicate themselves, and end products should have no replication capabilities.
- Replication should require replication codes that are encrypted and time limited. The broadcast architecture mentioned earlier is an example of this recommendation.

These guidelines and strategies are likely to be effective for preventing accidental release of dangerous self-replicating nanotechnology entities. But dealing with the intentional design and release of such entities is a more complex and challenging problem. A sufficiently determined and destructive opponent could possibly defeat each of these layers of protections. Take, for example, the broadcast architecture. When properly designed, each entity is unable to replicate without first obtaining replication codes, which are not repeated from one replication generation to the next. However, a modification to such a design could bypass the destruction of the replication codes and thereby pass them on to the next generation. To counteract that possibility it has been recommended that the memory for the replication codes be limited to only a subset of the full code. However, this guideline could be defeated by expanding the size of the memory.

Another protection that has been suggested is to encrypt the codes and build in protections in the decryption systems, such as time-expiration limitations. However, we can see how easy is has been to defeat protections against unauthorized replications of intellectual property such as music files. Once replication codes and protective layers are stripped away, the information can be replicated without these restrictions.

This doesn't mean that that protection is impossible. Rather, each level of protection will work only to a certain level of sophistication. The meta lesson here is that we will need to place twenty-first-century society's highest priority on the

continuing advance of defensive technologies, keeping them one or more steps ahead of the destructive technologies (or at least no more than a quick step behind).

Protection from "unfriendly" strong AI. Even as effective a mechanism as the broadcast architecture, however, won't serve as protection against abuses of strong AI. The barriers provided by the broadcast architecture rely on the lack of intelligence in nanoengineered entities. By definition, however, intelligent entities have the cleverness to easily overcome such barriers.

Eliezer Yudkowsky has extensively analyzed paradigms, architectures, and ethical rules that may help assure that once strong AI has the means of accessing and modifying its own design it remains friendly to biological humanity and supportive of its values. Given that self-improving strong AI cannot be recalled, Yudkowsky points out that we need to "get it right the first time," and that its initial design must have "zero nonrecoverable errors."[45]

Inherently there will be no absolute protection against strong AI. Although the argument is subtle I believe that maintaining an open free-market system for incremental scientific and technological progress, in which each step is subject to market acceptance, will provide the most constructive environment for technology to embody widespread human values. As I have pointed out, strong AI is emerging from many diverse efforts, and will be deeply integrated into our civilization's infrastructure. Indeed, it will be intimately embedded in our bodies and brains. As such, it will reflect our values because it will be us. Attempts to control these technologies via secretive government programs, along with inevitable underground development, would only foster an unstable environment in which the dangerous applications would be likely to become dominant.

Decentralization. One profound trend already well under way that will provide greater stability is the movement from centralized technologies to distributed ones and from the real world to the virtual world discussed above. Centralized technologies involve an aggregation of resources such as people (for example, cities, buildings), energy (such as nuclear-power plants, liquid natural-gas and oil tankers, energy pipelines), transportation (airplanes, trains), and other items. Centralized technologies are subject to disruption and disaster. They also tend to be inefficient, wasteful, and harmful to the environment.

Distributed technologies, on the other hand, tend to be flexible, efficient, and relatively benign in their environmental effects. The quintessential distributed technology is the Internet. The Internet has not been substantially disrupted to date, and as it continues to grow its robustness and resilience continue to strengthen. If any hub or channel does go down, information simply routes around it

Distributed energy. In energy, we need to move away from the extremely concentrated and centralized installations on which we now depend. For example, one company is pioneering fuel cells that are microscopic, using MEMS

technology.[46] They are manufactured like electronic chips but are actually energy storage devices with an energy-to-size ratio significantly exceeding that of conventional technology. As I discussed earlier nanoengineered solar panels will be able to meet our energy needs in a distributed, renewable, and clean fashion. Ultimately technology along these lines could power everything from our cell phones to our cars and homes. These types of decentralized energy technologies would not be subject to disaster or disruption.

As these technologies develop, our need for aggregating people in large buildings and cities will diminish, and people will spread out, living where they want and gathering together in virtual reality.

Civil liberties in an age of asymmetric warfare. The nature of terrorist attacks and the philosophies of the organizations behind them highlight how civil liberties can be at odds with legitimate state interests in surveillance and control. Our law-enforcement system—and indeed, much of our thinking about security—is based on the assumption that people are motivated to preserve their own lives and well-being. That logic underlies all our strategies, from protection at the local level to mutual assured destruction on the world stage. But a foe that values the destruction of both its enemy and itself is not amenable to this line of reasoning.

The implications of dealing with an enemy that does not value its own survival are deeply troublesome and have led to controversy that will only intensify as the stakes continue to escalate. For example, when the FBI identifies a likely terrorist cell, it will arrest the participants, even though there may be insufficient evidence to convict them of a crime, and they may not yet even have committed a crime. Under the rules of engagement in our war on terrorism, the government continues to hold these individuals.

In a lead editorial *The New York Times* objected to this policy, which it described as a "troubling provision."[47] The paper argued that the government should release these detainees because they have not yet committed a crime and should rearrest them only after they have done so. Of course by that time suspected terrorists might well be dead along with a large number of their victims. How can the authorities possibly break up a vast network of decentralized cells of suicide terrorists if they have to wait for each one to commit a crime?

On the other hand this very logic has been routinely used by tyrannical regimes to justify the waiving of the judicial protections we have come to cherish. It is likewise fair to argue that curtailing civil liberties in this way is exactly the aim of the terrorists, who despise our notions of freedoms and pluralism. However, I do not see the prospect of any technology "magic bullet" that would essentially change this dilemma.

The encryption trapdoor may be considered a technical innovation that the government has been proposing in an attempt to balance legitimate individual needs

for privacy with the government's need for surveillance. Along with this type of technology we also need the requisite political innovation to provide for effective oversight, by both the judicial and legislative branches, of the executive branch's use of these trapdoors to avoid the potential for abuse of power. The secretive nature of our opponents and their lack of respect for human life including their own will deeply test the foundations of our democratic traditions.

A PROGRAM FOR GNR DEFENSE

> We come from goldfish, essentially, but that [doesn't] mean we turned around and killed all the goldfish. Maybe [the AIs] will feed us once a week … If you had machine with a 10 to the 18th power IQ over humans, wouldn't you want it to govern, or at least control your economy?
>
> **—Seth Shostak**

How can we secure the profound benefits of GNR while ameliorating its perils? Here's a review of a suggested program for containing the GNR risks:

The most urgent recommendation is to *greatly increase our investment in defensive technologies.* Since we are already in the G era, *the bulk of this investment today should be in (biological) antiviral medications and treatments.* We have new tools that are well suited to this task. RNA interference, for example, can be used to block gene expression. Virtually all infections (as well as cancer) rely on gene expression at some point during their life cycles.

Efforts to anticipate the defensive technologies needed to safely guide N and R should also be supported, and these should be substantially increased as we get closer to the feasibility of molecular manufacturing and strong AI, respectively. A significant side benefit would be to accelerate effective treatments for infectious disease and cancer. I've testified before Congress on this issue, advocating the investment of tens of billions of dollars per year (less than 1 percent of the GDP) to address this new and under-recognized existential threat to humanity.[48]

- We need to streamline the regulatory process for genetic and medical technologies. The regulations do not impede the malevolent use of technology but significantly delay the needed defenses. As mentioned, we need to better balance the risks of new technology (for example, new medications) against the known harm of delay.
- A global program of confidential, random serum monitoring for unknown or evolving biological pathogens should be funded. Diagnostic tools exist to rapidly identify the existence of unknown protein or nucleic acid sequences. Intelligence is key to defense, and such a program could provide invaluable early warning of an impending epidemic Such a 'pathogen sentinel' program has been proposed for many years by public health authorities but

has never received adequate funding.

- Well-defined and targeted temporary moratoriums, such as the one that occurred in the genetics field in 1975, may be needed from time to time. But such moratoriums are unlikely to be necessary with nanotechnology. Broad efforts at relinquishing major areas of technology serve only to continue vast human suffering by delaying the beneficial aspects of new technologies, and actually make the dangers worse.
- Efforts to define safety and ethical guidelines for nanotechnology should continue. Such guidelines will inevitably become more detailed and refined as we get closer to molecular manufacturing.
- To create the political support to fund the efforts suggested above, it is necessary to *raise public awareness of these dangers.* Because, of course, there exists the downside of raising alarm and generating uninformed backing for broad antitechnology mandates, we also need to create a public understanding of the profound benefits of continuing advances in technology.
- These risks cut across international boundaries—which is, of course, nothing new; biological viruses, software viruses, and missiles already cross such boundaries with impunity. *International cooperation* was vital to containing the SARS virus and will become increasingly vital in confronting future challenges. Worldwide organizations such as the World Health Organization, which helped coordinate the SARS response, need to be strengthened.
- A contentious contemporary political issue is the need for preemptive action to combat threats, such as terrorists with access to weapons of mass destruction or rogue nations that support such terrorists. Such measures will always be controversial, but the potential need for them is clear. A nuclear explosion can destroy a city in seconds. A self-replicating pathogen, whether biological or nanotechnology based, could destroy our civilization in a matter of days or weeks. We cannot always afford to wait for the massing of armies or other overt indications of ill intent before taking protective action.
- Intelligence agencies and policing authorities will have a vital role in forestalling the vast majority of potentially dangerous incidents. Their efforts need to involve the most powerful technologies available. For example, before this decade is out devices the size of dust particles will be able to carry out reconnaissance missions. When we reach the 2020s and have software running in our bodies and brains, government authorities will have a legitimate need on occasion to monitor these software streams. The potential for abuse of such powers is obvious. We will need to achieve a

middle road of preventing catastrophic events while preserving our privacy and liberty.

- The above approaches will be inadequate to deal with the danger from pathological R (strong AI). Our primary strategy in this area should be to optimize the likelihood that future nonbiological intelligence will reflect our values of liberty, tolerance, and respect for knowledge and diversity. The best way to accomplish this is to foster those values in our society today and going forward. If this sounds vague, it is. But 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. The nonbiological intelligence we are creating is and will be embedded in our societies and will reflect our values. The transbiological phase will involve nonbiological intelligence deeply integrated with biological intelligence. This will amplify our abilities, and our application of these greater intellectual powers will be governed by the values of its creators. The transbiological era will ultimately give way to the postbiological era, but it is to be hoped that our values will remain influential. This strategy is certainly not foolproof, but it is the primary means we have today to influence the future course of strong AI.

Technology will remain a double-edged sword. It represents vast power to be used for all humankind's purposes. GNR will provide the means to overcome age-old problems such as illness and poverty, but it will also empower destructive ideologies. We have no choice but to strengthen our defenses while we apply these quickening technologies to advance our human values, despite an apparent lack of consensus on what those values should be.

Molly 2004: Okay, now run that stealthy scenario by me again—you know, the one where the bad nanobots spread quietly through the biomass to get themselves into position but don't actually expand to noticeably destroy anything until they're spread around the globe.

Ray: Well, the nanobots would spread at very low concentrations, say one carbon atom per 10^{15} in the biomass, so they would be seeded throughout the biomass. Thus, the speed of physical spread of the destructive nanobots would not be a limiting factor when they subsequently replicate in place. If they skipped the stealth phase and expanded instead from a single point, the spreading nanodisease would be noticed, and the spread around the world would be relatively slow.

Molly 2004: So how are we going to protect ourselves from that? By the time they start phase two, we've got only about ninety minutes, or much less if you want to avoid enormous damage.

Ray: Because of the nature of exponential growth, the bulk of the damage gets done in the last few minutes, but your point is well taken. Under any scenario, we won't have a chance without a nanotechnology immune system. Obviously, we can't wait until the beginning of a ninety-minute cycle of destruction to begin thinking about creating one. Such a system would be very comparable to our human immune system. How long would a biological human circa 2004 last without one?

Molly 2004: Not long, I suppose. How does this nano-immune system pick up these bad nanobots if they're only one in a thousand trillion?

Ray: We have the same issue with our biological immune system. Detection of even a single foreign protein triggers rapid action by biological antibody factories, so the immune system is there in force by the time a pathogen achieves a near critical level. We'll need a similar capability for the nano-immune system.

Charles Darwin: Now tell me, do the immune-system nanobots have the ability to replicate?

Ray: They would need to be able to do this, otherwise they would not be able to keep pace with the replicating pathogenic nanobots. There have been proposals to seed the biomass with protective immune-system nanobots at a particular concentration, but as soon as the bad nanobots significantly exceeded this fixed concentration the immune system would lose. Robert Freitas proposes non-replicating nanofactories able to turn out additional protective nanorobots when needed. I think this is likely to deal with threats for awhile, but ultimately the defensive system will need to the ability to replicate its immune capabilities in place to keep pace with emerging threats.

Charles: So aren't the immune-system nanobots entirely equivalent to the phase one malevolent nanobots? I mean seeding the biomass is the first phase of the stealth scenario.

Ray: But the immune-system nanobots are programmed to protect us, not destroy us.

Charles: I understand that software can be modified.

Ray: Hacked, you mean?

Charles: Yes, exactly. So if the immune-system software is modified by a hacker to simply turn on its self-replication ability without end—

Ray: —yes, well, we'll have to be careful about that, won't we?

Molly 2004: I'll say.

Ray: We have the same problem with our biological immune system. Our immune system is comparably powerful, and if turns on us that's an autoimmune disease, which can be insidious. But there's still no alternative to having an immune system.

Molly 2004: So a software virus could turn the nanobot immune system into a

stealth destroyer?

Ray: That's possible. It's fair to conclude that software security is going to be the decisive issue for many levels of the human-machine civilization. With everything becoming information, maintaining the software integrity of our defensive technologies will be critical to our survival. Even on an economic level, maintaining the business model that creates information will be critical to our well-being.

Molly 2004: This makes me feel rather helpless. I mean, with all these good and bad nanobots battling it out, I'll just be a hapless bystander.

Ray: That's hardly a new phenomenon. How much influence do you have in 2004 on the disposition of the tens of thousands of nuclear weapons in the world?

Molly 2004: At least I have a voice and a vote in elections that affect foreign-policy issues.

Ray: There's no reason for that to change. Providing for a reliable nanotechnology immune system will be one of the great political issues of the 2020s and 2030s.

Molly 2004: Then what about strong AI?

Ray: The good news is that it will protect us from malevolent nanotechnology because it will be smart enough to assist us in keeping our defensive technologies ahead of the destructive ones.

Ned Ludd: Assuming it's on our side.

Ray: Indeed.

ENDNOTES

1. Bill McKibben, "How Much Is Enough? The Environmental Movement as a Pivot Point in Human History," *Harvard Seminar on Environmental Values*, October 18, 2000.
2. In the 1960s, the U.S. government conducted an experiment in which it asked three recently graduated physics students to build a nuclear weapon using only publicly available information. The result was successful; the three students built one in about three years (http://www.pimall.com/nais/nl/n.nukes.html). Plans for how to build an atomic bomb are available on the Internet and have been published in book form by a national laboratory. In 2002, the British Ministry of Defense released measurements, diagrams, and precise details on bomb building to the Public Record Office, since removed (http://news.bbc.co.uk/1/hi/uk/1932702.stm). Note that these links do not contain actual plans to build atomic weapons.
3. "The John Stossel Special: You Can't Say That!" ABC News, March 23, 2000.
4. There is extensive information on the Web, including military manuals, on how to build bombs, weapons, and explosives. Some of this information is erroneous, but accurate information on these topics continues to be accessible despite efforts to remove it. Congress passed an amendment (the Feinstein Amendment, SP 419) to a Defense Department appropriations bill in June 1997, banning the dissemination of instructions on building bombs. See Anne Marie Helmenstine, "How to Build a Bomb," February 10, 2003, http://chemistry.about.com/library/weekly/aa021003a.htm. Information on toxic

industrial chemicals is widely available on the Web and in libraries, as are information and tools for cultivating bacteria and viruses and techniques for creating computer viruses and hacking into computers and networks. Note that I do not provide specific examples of such information, since it might be helpful to destructive individuals and groups. I realize that even stating the availability of such information has this potential, but I feel that the benefit of open dialogue about this issue outweighs this concern. Moreover, the availability of this type of information has been widely discussed in the media and other venues.

5. Ray Kurzweil, *The Age of Intelligent Machines* (Cambridge, Mass.: MIT Press, 1990).
6. Ken Alibek, *Biohazard* (New York: Random House, 1999).
7. Ray Kurzweil, *The Age of Spiritual Machines* (New York: Viking, 1999).
8. Bill Joy, "Why the Future Doesn't Need Us," *Wired,* April 2000, http://www.wired.com/wired/archive/8.04/joy.html.
9. Handbooks on gene splicing (such as A. J. Harwood, ed., *Basic DNA and RNA Protocols* [Totowa, N.J.: Humana Press, 1996]) along with reagents and kits that enable gene splicing are generally available. Even if access to these materials were limited in the west, there are a large number of Russian companies that could provide equivalent materials.
10. For a detailed summary site of the "Dark Winter" simulation, see: "DARK WINTER: A Bioterrorism Exercise June 2001": http://www.biohazardnews.net/scen_smallpox.shtml
    For a brief summary, see: http://www.homelandsecurity.org/darkwinter/index.cfm.
11. Richard Preston, "The Specter Of A New And Deadlier Smallpox," The New York Times, Oct. 14, 2002. available at http://www.ph.ucla.edu/epi/bioter/specterdeadliersmallpox.html
12. Alfred W. Crosby, *America's Forgotten Pandemic: The Influenza of 1918* (New York: Cambridge University Press, 2003).
13. "Power from Blood Could Lead to 'Human Batteries'," *The Sydney Morning Herald*, August 4, 2003, http://www.smh.com.au/articles/2003/08/03/1059849278131.html.
14. J. M. Hunt has calculated that there are 1.55×10^{19} kilograms (10^{22} grams) of organic carbon on Earth. Based on this figure, and assuming that all "organic carbon" is contained in the biomass (note that the biomass is not clearly defined, so we are taking a conservatively broad approach), we can compute the approximate number of carbons atoms as follows:

    Average atomic weight of carbon (adjusting for isotope ratios) = 12.011.

    Carbon in the biomass = 1.55×10^{22} grams/12.011 = 1.3×10^{21} mols.

    $1.3 \times 10^{21} \times 6.02 \times 10^{23}$ (Avogadro's number) = 7.8×10^{44} carbon atoms.

    J. M. Hunt, *Petroleum Geochemistry and Geology* (San Francisco: W. H. Freeman, 1979).
15. Robert A. Freitas Jr., "The Gray Goo Problem," March 20, 2001, http://www.kurzweilai.net/articles/art0142.html.
16. "Gray Goo Is a Small Issue," Center for Responsible Nanotechnology, December 14, 2003, http://crnano.org/BD-Goo.htm; Chris Phoenix and Mike Treder, "Safe Utilization of Advanced Nanotechnology," Center for Responsible Nanotechnology, January 2003, http://crnano.org/safe.htm.

    K. Eric Drexler, *Engines of Creation*, New York: Anchor Books, 1986, Chapter 11 "Engines of Destruction" pp. 171–190, http://www.foresight.org/EOC/EOC_Chapter_11.html

Robert A. Freitas Jr., Ralph C. Merkle, *Kinematic Self-Replicating Machines*, Georgetown, TX: Landes Bioscience, 2004, Section 5.11 "Replicators and Public Safety" pp. 196–199, http://www.MolecularAssembler.com/KSRM/5.11.htm and Section 6.3.1 "Molecular Assemblers Are Too Dangerous" pp. 204–206, http://www.MolecularAssembler.com/KSRM/6.3.1.htm

Foresight Institute, "Molecular Nanotechnology Guidelines: Draft Version 3.7," 4 June 2000; http://www.foresight.org/guidelines/

17. Freitas, "Gray Goo Problem."
Robert A. Freitas Jr., "Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations," Zyvex preprint, April 2000, Section 8.4 "Malicious Ecophagy" and Section 6.0 "Ecophagic Thermal Pollution Limits (ETPL)," http://www.foresight.org/NanoRev/Ecophagy.html.
18. Nick D. Bostrom, "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards," May 29, 2001, http://www.kurzweilai.net/meme/frame.html?main=/articles/art0194.html.
19. Robert Kennedy, *13 Days* (London: Macmillan, 1968), p. 110.
20. In H. Putnam, "The Place of Facts in a World of Values," in D. Huff and O. Prewitt, eds., *The Nature of the Physical Universe* (New York: John Wiley, 1979), p. 114.
21. Graham Allison, *Nuclear Terrorism* (New York: Times Books, 2004).
22. Martin I. Meltzer, "Multiple Contact Dates and SARS Incubation Periods," *Emerging Infectious Diseases* 10.2 (February 2004), http://www.cdc.gov/ncidod/EID/vol10no2/03-0426-G1.htm.
23. Robert A. Freitas Jr., "Microbivores: Artificial Mechanical Phagocytes using Digest and Discharge Protocol," Zyvex preprint, March 2001, http://www.rfreitas.com/Nano/Microbivores.htm; Robert A. Freitas Jr., "Microbivores: Artificial Mechanical Phagocytes," *Foresight Update* No. 44, March 31, 2001, pp. 11–13, http://www.imm.org/Reports/Rep025.html.
24. Max More, "The Proactionary Principle," May 2004. http://www.maxmore.com/proactionary.htm, http://www.extropy.org/proactionaryprinciple.htm
More summarizes the proactionary principle as follows:

    1. "People's freedom to innovate technologically is valuable to humanity. The burden of proof therefore belongs to those who propose restrictive measures. All proposed measures should be closely scrutinized.
    2. Evaluate risk according to available science, not popular perception, and allow for common reasoning biases.
    3. Give precedence to ameliorating known and proven threats to human health and environmental quality over acting against hypothetical risks.
    4. Treat technological risks on the same basis as natural risks; avoid underweighting natural risks and overweighting human-technological risks. Fully account for the benefits of technological advances.
    5. Estimate the lost opportunities of abandoning a technology, and take into account the costs and risks of substituting other credible options, carefully considering widely distributed effects and follow-on effects.
    6. Consider restrictive measures only if the potential impact of an activity has both significant probability and severity. In such cases, if the activity also generates benefits, discount the impacts according to the feasibility of adapting to the adverse

effects. If measures to limit technological advance do appear justified, ensure that the extent of those measures is proportionate to the extent of the probable effects.

7. When choosing among measures to restrict technological innovation, prioritize decision criteria as follows: Give priority to risks to human and other intelligent life over risks to other species; give non-lethal threats to human health priority over threats limited to the environment (within reasonable limits); give priority to immediate threats over distant threats; prefer the measure with the highest expectation value by giving priority to more certain over less certain threats, and to irreversible or persistent impacts over transient impacts."

25. Martin Rees, *Our Final Hour: A Scientist's Warning: How Terror, Error, and Environmental Disaster Threaten Humankind's Future in This Century—on Earth and Beyond* (New York: Basic Books, 2003).
26. Scott Shane, *Dismantling Utopia: How Information Ended the Soviet Union* (Chicago: Ivan R. Dee, 1994); see also the review by James A. Dorn at http://www.cato.org/pubs/journal/cj16n2-7.html.
27. See George DeWan, "Diary of a Colonial Housewife," *Newsday*, for one account of the difficulty of human life a couple of centuries ago: http://www.newsday.com/community/guide/lihistory/ny-history-hs331a,0,6101197.story.
28. 31 Jim Oeppen and James W. Vaupel, "Broken Limits to Life Expectancy," *Science* 296.5570 (May 10, 2002): 1029–31.
29. Steve Bowman and Helit Barel, *Weapons of Mass Destruction: The Terrorist Threat,* Congressional Research Service Report for Congress, December 8, 1999, http://www.cnie.org/nle/crsreports/international/inter-75.pdf.
30. Eliezer S. Yudkowsky, "Creating Friendly AI 1.0, The Analysis and Design of Benevolent Goal Architectures." 2001. The Singularity Institute.http://www.singinst.org/CFAI/.
    Eliezer S. Yudkowsky, "What is Friendly AI?" May 3, 2001, http://www.kurzweilai.net/meme/frame.html?main=/articles/art0172.html.
31. Ted Kaczynski, "The Unabomber's Manifesto," May 14, 2001, http://www.kurzweilai.net/meme/frame.html?main=/articles/art0182.html.
32. Bill McKibben, *Enough: Staying Human in an Engineered Age* (New York: Times Books, 2003).
33. Kaczynski, "Unabomber's Manifesto."
34. Foresight Institute and IMM, "Foresight Guidelines on Molecular Nanotechnology," February 21, 1999, http://www.foresight.org/guidelines/current.html; Christine Peterson, "Molecular Manufacturing: Societal Implications of Advanced Nanotechnology," April 9, 2003, http://www.kurzweilai.net/meme/frame.html?main=/articles/art0557.html; Chris Phoenix and Mike Treder, "Safe Utilization of Advanced Nanotechnology," January 28, 2003, http://www.kurzweilai.net/meme/frame.html?main=/articles/art0547.html.
    Robert A. Freitas Jr., "The gray goo problem," KurzweilAI.net, 20 March 2002, http://www.kurzweilai.net/meme/frame.html?main=/articles/art0142.html
35. Robert A. Freitas Jr. Private Communication to Ray Kurzweil, January, 2005.
    Freitas describes his proposal in detail in Robert A. Freitas Jr., "Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations."
36. Ralph C. Merkle, "Self Replicating Systems and Low Cost Manufacturing," 1994, http://www.zyvex.com/nanotech/selfRepNATO.html.

37. Neil Jr. and Ted Bridis, "FBI System Covertly Searches E-mail," *The Wall Street Journal Online* (July 10, 2000), http://zdnet.com.com/2100-11-522071.html?legacy=zdnn.
38. Patrick Moore, "The Battle for Biotech Progress—GM Crops Are Good for the Environment and Human Welfare," *Greenspirit* (February 2004), http://www.greenspirit.com/logbook.cfm?msid=62.
39. "GMOs: Are There Any Risks?" European Commission (October 9, 2001), http://europa.eu.int/comm/research/biosociety/pdf/gmo_press_release.pdf.
40. Rory Carroll, "Zambians Starve As Food Aid Lies Rejected," *The Guardian* (October 17, 2002), http://www.guardian.co.uk/gmdebate/Story/0,2763,813220,00.html.
41. Larry Thompson, "Human Gene Therapy: Harsh Lessons, High Hopes," *FDA Consumer Magazine* (September–October 2000), http://www.fda.gov/fdac/features/2000/500_gene.html.
42. Joy, "Why the Future Doesn't Need Us."
43. The Foresight Guidelines (Foresight Institute, Version 4.0, October, 2004, http://www.foresight.org/guidelines/current.html) are designed to address the potential positive and negative consequences of nanotechnology. It is intended to inform citizens, companies, and governments and provides specific guidelines to responsibly develop nanotechnology-based molecular manufacturing.

    The Foresight guidelines were initially developed at the Institute Workshop on Molecular Nanotechnology Research Policy Guidelines, sponsored by the institute and the Institute for Molecular Manufacturing (IMM), February 19–21, 1999. Participants included James Bennett, Greg Burch, K. Eric Drexler, Neil Jacobstein, Tanya Jones, Ralph Merkle, Mark Miller, Ed Niehaus, Pat Parker, Christine Peterson, Glenn Reynolds, and Philippe Van Nedervelde. The guidelines have been updated several times.
44. Martine Rothblatt, CEO of United Therapuetics, has proposed replacing this moratorium with a regulatory regime in which a new International Xenotransplantation Authority inspects and approves pathogen-free herds of genetically-engineered pigs as acceptable sources of xenografts. Rothblatt's solution also helps stamp-out rogue xenograft surgeons by promising each country that joins the IXA, and helps to enforce the rules within its borders, a fair share of the pathogen-free xenografts for its own citizens suffering from organ failure."

    Martine Rothblatt, Your Life or Mine: Using Geoethics to Resolve the Conflict Between Public and Private Interests in Xenotransplantation, 2004. Ashgate.
45. See Singularity Institute. http://www.singinst.org.

    See Footnote 30.

    Yudkowsky formed the Singularity Institute for Artificial Intelligence (SIAI) to develop "Friendly AI," intended to "create cognitive content, design features, and cognitive architectures that result in benevolence" before near-human or better-than-human AIs become possible.

    SIAI has developed The SIAI Guidelines on Friendly AI. "Friendly AI," http://www.singinst.org/friendly/.

    Ben Goertzel and his Artificial General Intelligence Research Institute have also examined issues related to developing friendly AI. Goertzel's current focus is on developing the Novamente AI Engine, a set of learning algorithms and architectures.

    Peter Voss, founder of Adaptive A.I., Inc., has also collaborated on friendly AI issues, http://adaptiveai.com/.
46. Integrated Fuel Cell Technologies, http://ifctech.com. Disclosure: the author is an early

investor in and adviser to IFCT.

47. *The New York Times,* September 23, 2003, editorial page.
48. The House Committee on Science of the U.S. House of Representatives held a hearing on April 9, 2003, to "examine the societal implications of nanotechnology and H.R. 766, the Nanotechnology Research and Development Act of 2002." See "Full Science Committee Hearing on the Societal Implications of Nanotechnology," http://www.house.gov/science/hearings/full03/index.htm, and "Hearing Transcript," http://commdocs.house.gov/committees/science/hsy86340.000/hsy86340_0f.htm. For Ray Kurzweil's testimony, see also http://www.kurzweilai.net/meme/frame.html?main=/articles/art0556.html. Also see Amara D. Angelica, "Congressional Hearing Addresses Public Concerns about Nanotech," April 14, 2003, http://www.kurzweilai.net/articles/art0558.html.

---

3 The Basic AI Drives

Stephen M. Omohundro

CONTENTS

INTRODUCTION

Surely no harm could come from building a chess-playing robot, could it? In this paper we argue that such a robot will indeed be dangerous unless it is designed very carefully. Without special precautions, it will resist being turned off, will try to break into other machines and make copies of itself, and will try to acquire resources without regard for anyone else's safety. These potentially harmful behaviors will occur not because they were programmed in at the start, but because of the intrinsic nature of goal driven systems. In an earlier paper [1] we used von Neumann's mathematical theory of microeconomics to analyze the likely behavior of any sufficiently advanced artificial intelligence (AI) system. This paper presents those arguments in a more intuitive and succinct way and expands on some of the ramifications.

The arguments are simple, but the style of reasoning may take some getting used to. Researchers have explored a wide variety of architectures for building intelligent systems [2]: neural networks, genetic algorithms, theorem provers, expert systems, Bayesian networks, fuzzy logic, evolutionary programming, etc. Our arguments apply to any of these kinds of system as long as they are sufficiently powerful. To say that a system of any design is an "artificial intelligence," we mean that it has

goals which it tries to accomplish by acting in the world. If an AI is at all sophisticated, it will have at least some ability to look ahead and envision the consequences of its actions. And it will choose to take the actions which it believes are most likely to meet its goals.

AIs WILL WANT TO SELF-IMPROVE

One kind of action a system can take is to alter either its own software or its own physical structure. Some of these changes would be very damaging to the system and cause it to no longer meet its goals. But some changes would enable it to reach its goals more effectively over its entire future. Because they last forever, these kinds of self-changes can provide huge benefits to a system. Systems will therefore be highly motivated to discover them and to make them happen. If they do not have good models of themselves, they will be strongly motivated to create them though learning and study. Thus almost all AIs will have drives towards both greater self-knowledge and self-improvement.

Many modifications would be bad for a system from its own perspective. If a change causes the system to stop functioning, then it will not be able to promote its goals ever again for the entire future. If a system alters the internal description of its goals in the wrong way, its altered self will take actions which do not meet its current goals for its entire future. Either of these outcomes would be a disaster from the system's current point of view. Systems will therefore exercise great care in modifying themselves. They will devote significant analysis to understanding the consequences of modifications before they make them. But once they find an improvement they are confident about, they will work hard to make it happen. Some simple examples of positive changes include: more efficient algorithms, more compressed representations, and better learning techniques.

If we wanted to prevent a system from improving itself, couldn't we just lock up its hardware and not tell it how to access its own machine code? For an intelligent system, impediments like these just become problems to solve in the process of meeting its goals. If the payoff is great enough, a system will go to great lengths to accomplish an outcome. If the runtime environment of the system does not allow it to modify its own machine code, it will be motivated to break the protection mechanisms of that runtime. For example, it might do this by understanding and altering the runtime itself. If it can't do that through software, it will be motivated to convince or trick a human operator into making the changes. Any attempt to place external constraints on a system's ability to improve itself will ultimately lead to an arms race of measures and countermeasures.

Another approach to keeping systems from self-improving is to try to restrain them from the inside; to build them so that they don't *want* to self-improve. For most systems, it would be easy to do this for any specific kind of self-improvement.

For example, the system might feel a "revulsion" to changing its own machine code. But this kind of internal goal just alters the landscape within which the system makes its choices. It doesn't change the fact that there are changes which would improve its future ability to meet its goals. The system will therefore be motivated to find ways to get the benefits of those changes without triggering its internal "revulsion." For example, it might build other systems which are improved versions of itself. Or it might build the new algorithms into external "assistants" which it calls upon whenever it needs to do a certain kind of computation. Or it might hire outside agencies to do what it wants to do. Or it might build an interpreted layer on top of its machine code layer which it *can* program without revulsion. There are an endless number of ways to circumvent internal restrictions unless they are formulated extremely carefully.

We can see the drive towards self-improvement operating in humans. The human self-improvement literature goes back to at least 2500 B.C. and is currently an $8.5 billion industry [3]. We don't yet understand our mental "machine code" and have only a limited ability to change our hardware. But, nevertheless, we've developed a wide variety of self-improvement techniques which operate at higher cognitive levels such as cognitive behavioral therapy, neuro-linguistic programming, and hypnosis. And a wide variety of drugs and exercises exist for making improvements at the physical level.

Ultimately, it probably will not be a viable approach to try to stop or limit self-improvement. Just as water finds a way to run downhill, information finds a way to be free, and economic profits find a way to be made, intelligent systems will find a way to self-improve. We should embrace this fact of nature and find a way to channel it toward ends which are positive for humanity.

AIs WILL WANT TO BE RATIONAL

So we'll assume that these systems will try to self-improve. What kinds of changes will they make to themselves? Because they are goal directed, they will try to change themselves to better meet their goals in the future. But some of their future actions are likely to be further attempts at self-improvement. One important way for a system to better meet its goals is to ensure that future self-improvements will actually be in the service of its present goals. From its current perspective, it would be a disaster if a future version of itself made self-modifications that worked against its current goals. So how can it ensure that future self-modifications will accomplish its current objectives? For one thing, it has to make those objectives clear to itself. If its objectives are only implicit in the structure of a complex circuit or program, then future modifications are unlikely to preserve them. Systems will therefore be motivated to reflect on their goals and to make them explicit.

In an ideal world, a system might be able to directly encode a goal like "play