James Blodgett
11 Fullerton Street
Albany, NY 12209
bjames1@nycap.rr.com
518-434-8860

20,144 characters

Steering Humanity
By
James Blodgett

Humanity is on the cusp of tremendous prospects—but also of disasters. Steering towards the prospects while avoiding the disasters has never been more important. We have no choice but to steer as best we can, since walking away from the steering wheel is also a form of steering.   Steering is a collective act of all of humanity, but a thesis of this essay is that steering starts with us: the author, the readers of this essay, and our colleagues. This is a point of view that puts a steering wheel directly in front of each one of us.

We cannot predict the future in detail, but more-or-less-plausible prospects are truly tremendous. Technology appears to be expanding exponentially in many areas. [1] Kurweil and others look forward to computers that are not only smarter than a single human, but smarter than the entire human species. [2] Also, Lewis and others find enough material in the asteroid belt to build O'Neill habitats for trillions of people. [3]   A paper by Metzger and others shows a more-or-less-plausible way to build them. [4] A more-or-less-plausible paper by Armstrong and Sandberg talks of settling not only the asteroid belt or the galaxy, but thousands of galaxies. [5]

Disaster is also plausible. There have been several mass extinctions in Earth's history caused by natural events. Fortunately, Muller and others see this as happening only about once every sixty-two million years, and none of the disasters that we know of have killed everything. [6] However, we do not have the comfort of being able to compute a low rate for mass extinctions caused by human technology because our technology has approached the capacity to cause them for only a short time. [7]   The history of lesser forms of human-caused disasters (for example, wars) suggests that we have plenty of capacity to make stupid mistakes and to do nasty things. I cite tremendous prospects above, but with great power comes great responsibility. Examples of potential dangers involving human technology that might kill us all include nuclear war, run-away global warming, and artificial intelligence or nanotech that becomes a Darwinian competitor. There are many more examples. On a more optimistic note, we have strong motivation to avoid disasters, so hopefully we will find a way to avoid them. If we can conceptualize existential threats (threats to human existence) we can try to develop ways to guard against them. For example, if our future really does include trillions of people dispersed across space, that will make most plausible disasters less than total because it is likely that some people will survive.

Our current management of these prospects is a problematic mixed bag. Science, the market, government, and the various arenas of public discourse all contribute to this management, and all are mixed bags.

Scientists and futurists have a plethora of enchanting concepts that enable plausible futures; enchanting but sometimes potentially dangerous. Most are speculative. Many probably won't ever happen, but some probably will. There is enormous value in checking them out and implementing the concepts that are good, workable, and safe. Unfortunately, scientific procedures for checking them out are one of the mixed bags. The good part is that science has already demonstrated its ability to develop marvelous technology, and the methodologies of science are a major intellectual achievement. The bad part is that science sometimes has trouble evaluating the prospects of speculative futures because of a conflict of interest, a conflict that is inherent in our human nature rather than malicious. The conflict is the fact that scientists want their theories to prove out, for both intellectual and career reasons. The desire that one's own theories and career should be fruitful conflicts even with activities that one would think would be objective, for example, evaluation of experimental results. This difficulty is why standard scientific methodology includes statistical tests, double-blind experiments, and peer review, so that the hopes of wistful scientists do not lead them to see results that are not really there. If this is a problem in the objective process of evaluating experimental results, it is even more of a problem in the subjective process of evaluating the prospects of speculative theories that enable speculative futures. The result is that scientists working in areas with both positive and negative potential sometimes overestimate the positive and downplay the negative. This sends just the wrong signals, signals that confuse those trying to manage positive and negative potentials. This is especially troublesome when those potentials have existential impact. We could use better methodology for vetting speculative potentials in both directions.

Another problem is that while market-driven entrepreneurship works well at the scale of current enterprise, it does not work well when very large investments are required to test very large but speculative potentials. Governments sometimes finance large projects in these areas, but their taste for such enterprise varies and is often less than optimal. For example, China financed massive naval explorations in the fifteenth century. Had they continued, China might have discovered Europe, but exploration ended shortly after Emperor Zhu Di died. NASA begins to sound like a similar case. It might seem that governments would take on the task of addressing the security aspects of speculative futures, since governments devote substantial resources to security against lesser threats with which we have long been familiar, for example military threats and threats to public order. However, governments are not often geared up to address speculative existential threats, especially threats that the public does not understand, threats for which the public is not ready to support expenses for amelioration. Raising understanding to the level where the public does understand both good and bad potentials is a job that is sometimes addressed in arenas of public discourse, but more could be done in this area.

Steering humanity involves two different scales. There is a steering wheel in front of each one of us, but the steering that individuals can do is generally accomplished by indirect means. Individuals rarely have the resources required to steer all of humanity, but there are sometimes larger groups that do have the requisite resources. Steering by individuals is accomplished by rhetoric, activism, and leadership that influence those larger groups. Often there are other individuals and other groups steering in different directions. Steering by multiple entities is aggregated (or ignored) by various decision and implementation mechanisms such as persuasion, voting, markets, war, etc.

The effectuality of steering varies depending on the steerability of the underlying situation. For example, cars are easy to steer, bicycles are unstable but have a dynamic stability that is easy to manage when one has the opportunity to learn, and unicycles can be steered but with difficulty. Mousetraps are at the most difficult end of this scale because they are deceptive. A mouse steering toward the bait is due for a surprise. Hopefully the universe is not malicious. Hopefully we rarely face a mousetrap-level steering problem. At any other level, an attempt to steer usually improves the odds by at least a little bit since steering tries for a good direction and might succeed. Hopefully the universe is not malicious, but the essence of the mousetrap is not maliciousness, it is the fact that the mouse does not understand the problem. Non-malicious natural forces can set a mousetrap for us by coming to a balance that we do not understand and might upset. The good news is that humans are smart, so we have a good chance of figuring out balances. An even worse mousetrap for those following a utilitarian philosophy (seeking the greatest good for the greatest number) is the unintended consequences along the many branches of possibility in the future, many more branches than we can confidently predict and evaluate, branches containing consequences that may overwhelm the good that we seek. However, unintended consequences are less of a problem for the events under consideration since steering around existential risk makes it easy to evaluate all of the branches that involve falling in, and steering towards a technological singularity [8] trivializes all of the branches that do not include a singularity. In many cases, if there is a positive or a negative singularity in the picture, nothing else matters much. That simplifies evaluation, and that simplifies steering.

It is difficult to steer something as humongous as all of humanity with little opportunity to learn from experience and with many hands on the wheel, albeit those many hands usually vote better than just one. Despite this difficulty, the effort is worthwhile because the odds are generally improved by the attempt. The odds improve because steering sometimes works. We can learn from similar efforts in history. First, humanity can be steered. History is a record of successful attempts to steer large sections of humanity to build things like the Roman Empire and the Great Wall. Second, singularities can be steered. The Industrial Revolution could be considered a mini version of a singularity. Thousands of innovators like Newton and Edison helped to build, or one might say steer, the Industrial Revolution. Third, ordinary people can help to steer. Minor reductions in the existential risks (i.e. negative singularities) of nuclear weapons and of asteroid impact have had a crowdsourced component. In the case of nuclear weapons, anti-bomb protestors and essayists countered in the public mind a certain military enthusiasm for nuclear weapons that might otherwise have torpedoed the Gorbachev-Reagan negotiations. In fact, the Russian military tried a coup. It might have been successful had public opinion in Russia been different. In the case of asteroid impact, the odds of disaster have been reduced somewhat by asteroid surveys that so far have not found a threat, but that might motivate deflection projects if necessary. Amateur astronomers have helped with these surveys. None of these contributions definitively saved the world in the style of Superman, but they did contribute.

Despite the lack of superhero effectuality, and despite a sometimes low probability of success that can sometimes seem daunting, an attempt to steer humanity is often worthwhile because of expected value (probability times value), a standard metric of decision theory. If we win a dollar each time a coin flip comes up heads, then that flip has an expected value of fifty cents, and in a sense the opportunity to flip is worth that much. Expected value and utilitarianism are good compasses to steer by, despite philosophical issues at the extremes. Even if we have only a one-in-a-thousand chance of steering humanity away from an existential risk, then that steering has an expected value of seven million lives (i.e. seven billion lives, the current population of Earth, multiplied by the probability of one in a thousand.) Few heroes of legend have saved more. (We should also count future lives, but those are harder to estimate.) If we are trying to steer humanity toward a future that enables trillions rather than billions of lives, expected values can be even larger. The utilitarian "greatest good for the greatest number" is much larger when there is a much larger number to experience the good of existence. Part of an individual's low odds in this form of activism is due to the difficulty of influencing an organization, often necessary because an organization's resources are required to enable effective steering. Once that influence is attained, then the odds improve because a difficult part of the process is out of the way. The odds are also better for organizations because they are more able than individuals to undertake portfolios of projects. If organizations have portfolios of scores of projects for improving the prospects of humanity, there is a fairly good chance that at least one of them will succeed. Even if despite all these considerations our probability of success is discouraging, there is human satisfaction and honor in the quest of doing what one can do for a transcendent cause, even against overwhelming odds. [9] The future of humanity is a transcendent cause. Hopefully the odds are not overwhelming.

Recruiting people to help is one objective of this essay. Recruiting people is a form of crowdsourcing. We could also use the help of organizations, governments, and philanthropists. I expect that most readers will not respond, but I also expect that those who do respond will do so because they get the point, and just the fact that they get the point makes them better qualified than most. This is an area that requires getting the point, i.e. understanding the technical possibilities, both positive and negative. It also requires understanding of the rhetorical and political contingencies of getting people and organizations to take action. It also requires that one focus on the public interest. Some individuals focus on their private interests. The invisible hand of the market converts this to public interest in areas where a market functions, but the market generally does not function to reward improvements in the future of humanity. However, some large organizations are supposed to focus on the public interest, so policies benefiting the public have a selling point that may help in recruiting these organizations. Officials of these organizations have at least the role of advancing the public interest, and so should welcome projects that can be seen to advance that goal. Working on the public interest can also be a noble pursuit of which we can be proud, a worthy use for our skills and time (and money, if we have some to spare.)

Once one signs on, what does one do? First, do no harm. However, following that dictum too closely is a prescription for inaction. There are few actions that involve no possibility of harm. There is more expected value in evaluating both positive effects and negative side effects, and taking care to assure that the former dominate. If our current situation is dangerous, as some authors suggest [10], then a solution that involves some danger may be better than doing nothing and accepting the background level of danger. For example, artificial intelligence may have an element of risk, but it is also a component of several solutions to other risks. Study these areas carefully, understand them well, and develop or recognize a good solution. Then persuade people and organizations to implement that solution. It is important to think well, but even if you come up with an idea that is bad for a reason you do not recognize, it is likely that bad ideas will be weeded out because of the need to convince others, and convincing others is easier with good ideas. The process is a cooperative one, usually best achieved with standards of courtesy and respect for fellow contributors, even when their ideas for steering compete with one's own. The real win is a good solution, but not necessarily our own solution.

To study these areas, read about various versions of singularities and about various existential risks. Google these topics or ask reference librarians, and follow up on interesting references within initial sources. Read authors with different viewpoints. Also check out organizations that are active in these areas. A few examples are the Future of Humanity Institute at Oxford, the Lifeboat Foundation, the National Space Society, Singularity Institute, and Lesswrong.com. Some of these organizations have mixed bag aspects too. [11] The existence of mixed bag aspects demonstrates valuable diversity, room for improvement, and work to be done.

When you develop or recognize good ideas, and have checked to verify that they are good, the next step is to implement those that can be done personally, or to work to convince groups to implement those that require additional resources. Several types of what I call "effectuality groups" have the purpose of accomplishing large projects. Effectuality groups include governments, markets, limited-liability-joint-stock companies, and non-government organizations, of which foundations are an example. All were invented to facilitate large projects. If existing organizations do not work for specific projects that improve the prospects of humanity, consider that the process of inventing new organizations may need to be continued, so that we can invent and found a new type of organization that will work. For example, Metzger advocates starting an Industrial Revolution in space, starting with a few tons of equipment sent to the moon, equipment like 3D printers that use regolith or its components, machine tools, and robots (initially teleoperated). This equipment would be used to build more equipment that in turn would build more equipment, with each generation using fewer components from Earth and expanding exponentially in numbers and in capabilities. [4 op. cit.] This requires reinventing Earth machinery so that it can be constructed by the available fabrication methods and can work in the space environment. Existing organizations might implement developing these reinventions via large cost-plus contracts. Recently, important software like Linux and Android has been developed by open software groups. Perhaps some of the reinvention that Metzger's concept requires could be done by similar groups, perhaps encouraged by directed prizes. Perhaps teleoperation could be financed by renting time on equipment in space and establishing a market for products.

I just received an email from Keith Henson. He has been trying to get us into space for years, co-founding the L5 society almost forty years ago. His latest proposal is based partly on an idea by Steve Nixon that cuts the start-up cost by $80 billion. It involves rapidly developing space solar power to solve energy, carbon, climate and economic problems. He has extensive engineering studies showing this to be much cheaper than fossil fuel on a large scale given low cost launch capacity which he hopes to achieve by using a British space plane (Skylon) to carry a load up 16 miles, then taking it into orbit using banks of giant lasers in geosynchronous orbit that can heat reaction mass and achieve higher exhaust velocity than chemical rockets. I don't know whether he can make it work, but if we had a thousand Keith Hensons, some of them would come up with something that would work. Keith is a colorful character who cannot be duplicated, but with luck crowdsourcing might encourage thousands to be as creatively productive as Keith. Readers of this essay are invited to try to be one of those thousands.

Humanity has tremendous future potential. Not achieving that future would be a tremendous tragedy. Somehow or other we should do a better job of getting there. We, individually and collectively, can help. Let's get going!

- - - - - - -

Endnotes: [In citation/discussion format, therefore part of exposition and part of character count]

[1] Moore's "law" says that the number of transistors on a chip doubles approximately every two years. This rate has been approximated for over forty years, and has already resulted in a very large increase in capacity. See [Wikipedia, "Moore's Law," accessed 4/13/14.] Ray Kurzwiel (and others) say that much of technology also grows exponentially. [Kurzwiel, The Singularity is Near, Penguin Group, 2005.] However, others think differently, for example: [Roland Wagner-Döbler, "Rescher's Principle of Decreasing Marginal Returns of Scientific Research," Scientometrics, 2001, 419-436.] I don't think we can predict the capacity of future technology precisely, but technology has been improving rapidly, it is plausible that this improvement will continue at least for a while, and it is plausible that it will improve enough to enable singularities.

[2] Kurzweil, op. cit. p. 136, states that in the mid 2040s, ". . . the intelligence created per year . . . will be about one billion times more powerful than all human intelligence today." Also see [Kenneth Baldouf and Ralph Star, Succeeding With Technology, Course Technology, 2010, p. 168.]

[3] I am using "O'Neill habitats" generically to signify any large space habitat. O'Neill himself did not expect his designs to be implemented precisely. [O'Neill, The High Frontier: Human Colonies in Space, William Morrow & Company, 1977.] John Lewis makes a rough estimate that there is enough iron in the asteroid belt to build a habitat for 10,000 trillion people. [Lewis, Mining the Sky: Untold Riches from the Asteroids, Comets, and Planets, Perseus Publishing, 1997, pg. 194.] This estimate is based on the amount of iron and does not consider loss of materials for reaction mass etc. Others estimate that the asteroid belt can support a population in the range of 10-100 trillion. **{** [Curreri & Detweiler, "A Contemporary Analysis of the O'Neill – Glaser Model for Space-based Solar Power and Habitat Construction," NSS Space Settlement Journal, Dec 2011.], cites this estimate as appearing in [Billingham, Gilbreath, & O'Leary, Editors, Space Resources and Space Settlements, SP-428, NASA, Washington, D.C., 1979.] However, in a personal communication, Curreri says that this citation is an error and the estimate actually appears in [Johnson & Holbrow, "Space Settlements: a Design Study," NASA SP-413, NASA, 1977.] I haven't had time to check. **}**

[4] [Philip Metzer et al, "Affordable, Rapid Bootstrapping of Space Industry and Solar System Civilization," Journal of Aerospace Engineering, April 2012.] Metzer's ideas are described briefly in the third paragraph from the end of this paper. (The paper you are reading.)

[5] [Stuart Armstrong & Anders Sandberg, "Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox," Acta Astronautica, Aug–Sept 2013.] The authors suggest firing what I call "seed ships" at many galaxies. These are low-mass pods stuffed with AI, nanotech, the DNA of many species including humans, and human memories; pods able to reconstruct a biosphere and civilization wherever they land. (My wife sees moral issues in this.)

[6] Evidence for a 62-million-year cycle of mass extinction is summarized in [Paul Glister, "Galactic Drift and Mass Extinction," July 2007 at http://www.centauri-dreams.org/?p=1378, accessed 4/15/14.]

[7] Willard Wells makes this point in [Wells, Apocalypse When?: Calculating How Long the Human Race Will Survive, Springer, 2010.]

[8] In mathematics, a singularity is a point at which a function goes to infinity. Exponential growth is metaphorically similar. A "technological singularity" is a postulated event in the near future when exponentially increasing technology "goes to infinity" and becomes transcendently amazing. See [Wikipedia, "Technological singularity," accessed 4/15/14.]

[9] Google "Impossible Dream lyrics," or listen to a YouTube version. This is a good one: [https://www.youtube.com/watch?v=XuH34mXJ5i4 , accessed 4/16/14.]

[10] Lord Martin Rees, Astronomer Royal of Great Britain, estimates in [Rees, Our Final Hour, Basic Books, 2003] that humanity's chance of surviving the next century is 50%. Willard Wells, op. cit., calculates a similar estimate.

[11] For example, see the following evaluation of Singularity Institute posted on Lesswrong by Holden Karnofsky, Co-Executive Director of GiveWel: [http://lesswrong.com/lw/cbs/thoughts_on_the_singularity_institute_si/#Arguments , accessed 4/15/14.]