How should humanity steer the future? Humanity must benefit the thriving of its constituents!

Randal A. Koene RANDAL.A.KOENE@CARBONCOPIES.ORG
1087 Mission Street
San Francisco, CA 94103, USA

Abstract

Using historic developments that are perceived as progress we derived reliable requirements for a positive future. Humanity needs to survive and continues to exist, and it succeeds by helping its constituent members thrive in terms of fairness in the availability of opportunity for development, achievement and expression. With those goals, we analyzed humanity in terms of risks and proposed applicable solutions. The solutions have a set of common requirements that need to be emphasized: Human adaptability, Access to technological developments and trust through Transparency.

1. Introduction

The premise that humanity can strive toward a better future implies that humanity is facing serious problems, some beyond the scope of typical political and social discourse. Here, we consider the big picture and where appropriate a cosmic scope. Occasionally, such concerns do manage to enter public discourse. For example, the real and serious danger that near-Earth objects (e.g. asteroids) can pose received such attention and a tracking program was mandated. This shows that highlighting concerns for humanity anchored in real physical or cosmological properties and processes can lead to action and improve our odds.

A better future for humanity can involve: 1.) counter-acting or avoiding existing or predictable damaging influences, and 2.) stimulating developments that steer towards positive and desirable conditions, events and experiences. Positive and desirable are a higher bar than just survival. Just as we want children not only to survive, but to thrive, humanity should also thrive.

We begin our analysis with survival and thriving on the grand scale. What are predictable problems at that scale? We then propose solutions and proceed to relate them to immediate needs, implementations rooted in science and technology today, and our socio-political reality. We take this approach, because good proposals must insure that long-term requirements are met and are not sacrificed to near-term focus. After all, long-term consequences ultimately determine success or failure.

We interpret the question "How should humanity steer the future?" as "What is the future of humanity that we should steer towards". Therefore, we need to understand what humanity is in its present state, as well as to form substantiated ideas about what humanity should become. The information supporting those two insights was gathered simultaneously, but we begin in sections 2 and 3 with arguments for what humanity should become. Current conditions and their difference compared with desired conditions indicate a number of risks, and by describing those in section 4 we can use goal conditions already clearly derived for the selection of acceptable solutions. It will

become apparent that this is important, because many proposals for dealing with specific problems without that context can steer humanity toward dystopia. Besides, this is a logical way to present the goal-oriented top-down analysis, then how it connects with bottom-up analysis from present conditions, resulting in viable paths to the desired future.

2. What should humanity be?

On which points could we agree that would constitute a thriving humanity in the future? In prior work (Koene, 2014), I pointed out the importance of being on a continuing path of creating, exploring, experiencing, learning and understanding. The significance of those activities originates in the desire to satisfy individual interests and the drives that are associated with that. Notice that those activities and their results are among the accomplishments humanity has celebrated and memorialized: art, invention, exploration, leadership.

We speculate that those appreciations are a product of successful selection, not just in human evolution, but also favored by ongoing universal Darwinism (Dennet, 2005). It is reasonable to assume that drives to create and understand were selected for, because those activities impart survival advantages. A form of selection pressure exists whenever something can manifest as a comparison or competition. If that process is universal then some features of our definition of thriving may be mandatory in the sense that a species that does not exhibit them would not, on a cosmic scale, be able to persist to a significant degree.

A popular claim is that the purpose of life, and in particular intelligent life, is to counteract the entropic tendency toward disorder. At extremes, such claims go on to suppose that we might somehow incorporate the entirety of the universe into complex organization. Without further specification, that is a nonsensical aim and physically impossible.

2.1 Development of rights and opportunities

Looking at human history, we applaud those developments that improved the well-being of persons in terms of their rights and opportunities. We call it progress. Sometimes, those developments involved the removal of unfair disadvantages, thereby leveling the playing field for a greater proportion of humanity. Sometimes, progressive developments involved expanded empathy, acknowledging similarity with subsets of humanity and extending to them a similar status.

- The process has involved the removal of previously existing systems of biased treatment intrinsic to a power structure that inhibited mobility. Examples are career opportunities for women, opportunities for minorities, opportunities independent of 'noble' birth and independent of financial status or parentage.

- The process has involved developing technological solutions that work around limitations that affect specific groups or individuals. Examples are electronic delivery of learning tools to children in poor countries, access to news via cell phones in many countries, treatments for hereditary diseases, technology that enables the handicapped to go about their lives, medical treatments and adapted living for the elderly.

- We value *Fairness*, after all, we might in some circumstances ourselves experience the downsides of unfairness.

- We value *Opportunity for Development and Achievement*.
- We value *Protected Expression*, because each of us is in the minority sometimes.

2.2 Thriving of humanity is a function of the thriving of its constituents

From the above, we derive this big-picture understanding: Part of what we consider thriving humanity involves continuing improvement according to such shared values as they relate to the conditions and opportunities of *members* of humanity. In other words, thriving humanity cannot be abstracted and taken out of the context of the well-being of the constituents of humanity.

Note that we consider it acceptable and just to ask or expect members of humanity to make sacrifices voluntarily or in solidarity when that is necessary for humanity as a whole. We do not generally consider it justified if members of humanity are instead sacrificed to an abstract or ideological 'good' of humanity.

Our responses show that we treat humanity as something that exists and should thrive for the security, support and benefit of the constituent members of humanity. Sometimes, human society is compared with a multicellular organism, but our value judgments about the parts and the whole are clearly not in line with that example. After all, selection pressures resulted in organisms where the constituent components (cells) serve the whole with no prioritization of individual thriving.

Importantly, this means that when we consider solutions to problems humanity may face, options that involve disregard for or obsolescence of constituent members are undesirable. We should favor solutions that empower humanity's members to overcome those problems themselves. We are measuring the success of humanity according to the perceptions of the constituents of humanity, and especially the perceptions of human beings.

3. Sensible goals

1. We want to ensure that humanity exists in the future.
2. We also want future humanity – all of human society and what it contains, including artificial intelligence (AI) – to exist for the benefit of its members, for their thriving and opportunities in accordance with principles that we have been elevating for many centuries.

There are some consequences to consider: As analysis in section 4 will demonstrate, risks that threaten those two goals may emerge due to the limitations of human beings. Proposed solutions that remove those limitations imply changes, not just to human society, but in human beings. It is a self-directed, engineered process that is therefore distinct from genetic speciation and natural selection.

The members of humanity may therefore experience changes, yet favorably, so that they are not subject to out-selection in competition with successors. The change may therefore be more reminiscent of the in-person changes we experience as we develop from child to adult. In order for humanity to succeed and benefit the thriving of its members, those members need to grow and adapt to novel challenges.

Goals for humanity and its members in mind, what is needed in order to achieve them? There is much about ourselves that we do not yet understand. In fact, we do not yet have access to the data that would allow us to come to such understanding. Clearly, a crucial component is to gain access to information about ourselves, and to learn to understand that information.

Maintaining fairness of opportunities to the benefit of all members of humanity will require that we avoid scenarios that lead to domination by a fractured subset of humanity, as in the oft-discussed scenario where superhuman AI dominates (Yudkowski, 2008). Foreseeing undesirable development of that kind means that insight and planning often needs to precede application, but actual understanding emerges from iterative study and application. A carefully considered step-wise process alternating insight an application is possible.

Now, we take our goals for a desirable future and switch to the bottom-up perspective accomplished by a present-day analysis of humanity. We identify long-term problems and propose solutions that take into account the essential requirements that a.) humanity continues to exist, and that b.) it serves to benefit the thriving of its members, including further improvements along the directions that the preceding analysis found we generally approve of.

4. What is humanity now and how do we address problems so that humanity can thrive?

Humanity is a collection of interacting parts that has progressed to a point where, with technology, our connectivity is extensive. Nevertheless, humanity is ultimately composed of many individuals, primarily, but not exclusively, human beings. Humans, of course, are one of many animal species that emerged from a process of natural selection.

To consider characteristics of humanity let us first realize that humanity is more than a group of humans. Humans are creators and tool-makers, and over millenia have adapted humanity to depend crucially on many of our creations. Humanity has depended on the inclusion of work animals, some of which had degrees of agency (e.g. hunting dogs). Today, humanity includes computer systems that have the agency to take action on the stock market, auto-pilots, and an increasing number of emerging AI. For our purposes, when we use the general terms 'constituents' or 'members' of humanity, we include all of that.

7 billion human beings: Human beings are multicellular organisms suited to life in the environment of Earth's biosphere. Humans are cognitive agents that have intrinsic drives, consciousness, individual characteristics and personal interests. Eventually, one problem is that human beings are not the result of engineered optimization and are not naturally upgradable the way software is, for example. We are products of natural selection. Natural selection does not transition a species from one set of fitness characteristics to another. Instead, it kills off any species that prove unfit for prevailing circumstances. Evolution by natural selection is therefore not a survival strategy for humanity, but a possible scenario for extinction.

New selection pressures may appear when facing an environmental challenge (e.g. changes in temperature, radiation, atmospheric chemistry), a challenging event (e.g. impending meteor strike, gamma ray burst, resistant disease pandemic), or a more elaborate, subtle but serious challenge (e.g. exponentially increasing information processing and decision making demands, contact with another form of intelligence). Fitness for the totality of humanity may then be attained in a number of ways. Here, consider two solution categories, namely those that 1.) work around human characteristics, or that 2.) modify human characteristics to achieve fitness. The first type of solution requires additional translation layers between evolved human characteristics and novel conditions in the form of technology and new constituent members required for a functional humanity, shifts in the combined demographic. The less well-suited humans do not benefit from opportunity as well

in that scenario as do more well-suited agents. This approach does steer towards the continued existence requirement, but not the fairness of opportunity requirement.

The second type of solution may meet both requirements, and involves emphasizing the development of therapies and procedures that can adapt human beings to address needed capabilities. Developments in bio-engineering, synthetic biology and gene therapy can do that when requirements are within the reach of biology. Beyond that, as when adapting to a lack of vital resources such as air, extremes of temperature or pressure, or information processing requirements, it is necessary to include or transition to adaptations implemented in non-biological substrates that are then integral to a person. This broad range of adaptability requires strong developments in cybernetic technology, brain and body prostheses, and brain-machine interfaces. Its logical conclusion is to develop complete access to the mechanisms of body and brain, as well as the potential for transition of any or all parts to more easily adaptable substrates (Koene, 2012b,a). Comparative risk along this path of technology development and its mitigation were recently described in a paper by Eckersley and Sandberg (2013).

Inhabitants of Earth: The gravity wells of Earth and Sun stake out a tiny portion of our universe. The resources within this little pocket are limited. Furthermore, existing only within a single location presents a problem, a single point-of-failure for humanity faced with cosmic risks (e.g. solar transition to the red giant phase of its life-cycle, cosmic gamma ray bursts, large asteroid impact). Exposure to a single environment also means that all of our experience, learning and skills are based on that environment, as is the totality of genetic evolution.

This problem is solved by extending humanity past the confines of our solar system, which can include both a presence in interstellar space and travel to other stars and planets. Space travel is a prime example where human biology is not well-suited. Our current life spans are not suited to the duration of travel and our biology is not suited to the environment. No wonder the vast majority of space probes are robots. So, the solution to this problem leads back to the previous problem of new selection pressures and the two different solution categories. In the first case, humanity beyond our solar system consists primarily of non-human robots. In the second, human beings are adapted to the challenge, which again levels the playing field in terms of opportunity.

Humanity is a collective with sentience: The constituents of humanity include intelligent agency, memory, language, empathy and collaboration, reasoning, associative learning, awareness and consequent tools and science. The level of activity of humanity as a whole has increased with networked interactivity and the scale of collaborative action. The sentience of individual constituents of humanity, human beings in particular, is constrained by biological limitations. There are problematic scenarios: The inability of biological humans to augment their experiential capabilities, even when other constituents (e.g. AI) may be able to experience new capabilities may result in a.) fracturing of humanity into subsets with different levels of sentience, and b.) incorrect assessment, goal-setting and action when those processes depend on humans that cannot fully appreciate the data available to humanity.

For a future in which human beings fully benefit from opportunities, limitations of human sentience may be overcome through modification with new processing algorithms (e.g. perfect memory, date-addressable memory, problem solving with parameter fitting by quantum annealing, microsecond reaction times). To be an integral part of the sentient experience it is not sufficient to strap on Google Glass. A modification needs to be situated within the activity flow of brain circuitry. For that to be possible, we need to encourage the development of brain machine interfaces that can

connect with large numbers of neurons in many brain regions. Developments in this direction are closely related to the development of new neuroscience tools for brain activity mapping (e.g. the 'Neural Dust' prototype developed at UC Berkeley (Seo et al., 2013), as well as neural prostheses that can interpret and respond in the language of patterns of brain activity (e.g. the artificial hippocampus developed at USC (Berger et al., 2012).

The dominant species: The dominant species has a great deal more agency and control than other species. Plans and interests of the dominant species supersede those of any other. Dominance improves the odds of species survival and maintaining dominance is therefore a survival strategy. Dominance depends on fitness to deal with challenges encountered. The dominance of humanity also depends on the absence of a stronger competitor, where competition can take place in the domains of intelligence, speed, size (i.e. resources) and dedication (perseverance). Maintaining dominance can rely on an "arms race" mentality and on explicit awareness of the strategic significance of dominance, as expressed through focus, goals, prediction or preemption, and keeping an eye out for potential competition. (As dictated by game theory, competition may, at least temporarily, be transformed into collaboration.)

We could seek to maintain dominance of humanity by building machines that have ever greater intelligence, speed, resource options and goal-directedness, without offering similar growth to human beings. That may satisfy the overall requirements for existence of the totality, but it is clear that in a competition where human beings are handicapped in that manner, opportunities could rapidly shift away from us and toward our machine creations. A solution that also meets opportunity and fairness requirements will need to extend the possibility for such improvements to us. Engineered adaptability beyond natural selection is again key, and ultimately may involve including or transitioning to non-biological substrates that better support some of the target adaptations.

A species with a track-record of a couple of million years: On a cosmic scale, humanity has existed for but a short moment. Overall, most species survive on Earth only for a limited duration, constrained by their environmental niche. The likelihood that time of existence is constrained is itself an obvious problem for humanity, and it brings to mind the probable risks suggested by the Fermi paradox (Wesson, 1990). Activities that are in line with the positive goals we expressed for the future of humanity may require a lot of time, for example, distant travel and exploration, ambitious creations, and deep understanding. Limited survival of a species is normally linked to competitive natural selection in the face of novel challenges. Consider these problematic scenarios: Significant environmental change, appearance of a competing species, disappearance of a vital resource, intrinsic decay of the species (e.g. accumulated genetic defects, uncontrollable disease, stagnation), or events that we are not equipped to notice or deal with.

As a totality, humanity may be able to persist as a species through its inclusion of constituent parts (e.g. AI, robotic agents) that can deal with local and cosmic changes. To benefit from the persistence, we, the human beings, need to be that adaptive as well.

Composed of many overlapping groups categorized or measured along several dimensions: Categories often also represent certain interests and therefore nuances in the assessment of what marks desirable or undesirable developments for the future of humanity. We list a few examples:

- biologically female / male / etc
- cultural or historic groups

- national groups
- language groups
- skill-specific groups
- health or disease groups
- age groups
- relationship identifiers (by DNA, contract, allegiance, etc)

The various subsets of humanity have somewhat different experiences of progress, because a variety of issues affect access and availability of opportunities. Some examples are bearing children, cultural values, necessary skills, or physical performance. Much of what we assess as progress in prior human history has involved the adoption of ways to level the playing field, to exclude fewer members of humanity from given opportunities. An important part of steering toward a good future is to continue that effort.

Contains hierarchical structures within networks of human interconnection and roles: Such hierarchies structure the activities and direction of humanity and they afford it the ability to organize major accomplishments. There are dynamics at work within humanity that restructure the hierarchies, which have taken many forms throughout history. Existing hierarchies have played a role in, been accepted or discarded due to morals and ethics that prevailed at a given time. Hierarchies can put a small group in a position of power and feedback within the system that further exacerbates the distinctions has led to problems. Hierarchies can be more or less rigid. Within and between them, the concept of "balance of power" has had a significant influence (Sheehan, 2000).

Rigid power structures reduce the opportunities for achievement of many of the members of humanity. In the extreme case, we imagine a situation where a strong focus on the development of AI leads to overwhelming and unchangeable domination by a small collection of such agents. While the totality my continue to exist, may even continue to develop and create, and in that sense may arguably be successful, that does not attain the fairness and opportunity goals proposed. One method to counter such a scenario (AI dominance, absolute monarchy or other) is through de-facto or institutional balance and constraint of power. Established democracies attempt to guarantee that in an institutional manner. Historically, alliances of national powers applied to contain rogue dictators have relied on the de-facto balance offered by the existence of many independent nodes of power.

Here, we propose that vigilance about transparency and early rapid distribution of access to new developments may improve the odds of malleability in hierarchies. Broad access to technology that enables substrate-transition and improved adaptation in humans was recognized in Eckersley and Sandberg (2013) as a means of mitigating the risks of the technology itself, as well as a potential means to maintain a balance of power with regard to AI. Recognizing the value of balance for the whole community, it is in the interest of the community to pool resources where needed so that access to new developments (health, education, adaptations) can be provided the as many members of humanity as possible.

Has infrastructure: Infrastructure are the connections between us that make it possible to share burdens and specialize activities. Improved infrastructure is responsible for rapid growth and the ability to sustain a much larger and more capable whole of humanity. Infrastructure, when regarded from different perspectives, can be either robust or fragile. Robustness appears with increased size of the network. There is no single supplier or single destination. Fragility comes from shared underlying requirements. All of the many paths in the network have certain fundamental requirements in common, for example, an energy supply or common goods (light, water, etc).

Reliance on a small selection of underlying requirements can be reduced through diversification of needs. Ultimately, this leads back to adaptability to different fitness criteria and the solutions proposed above.

Has economic activity: Economic activity is the existence of communication and trade, which allows constituents of humanity to benefit from the existence of other members. A large economy (and also an extensive infrastructure) are essential in order for humanity to accomplish grand things, such as most major scientific or social programs. Well known examples are the Apollo program, but also accomplishments such as systems that share the burden of health care. Of course, the existence of an economy does not proscribe which accomplishments are valued, and therefore which goals benefit from significant effort by humanity. Economic activity relies on a great deal of trust and when trust breaks down sustained economic activity can break down as well.

Trust is related to transparency and balance of power. Human collaboration is aided by our similarities in that we can empathize, imagine, model the motivations and behaviors of other humans, as well as by our ability to communicate and thereby negotiate. Balance of power was addressed previously. Transparency in communications and negotiation with explicit safeguards that allow for trust can often be built into systems, such as is demonstrated by the ledger block-chain solution that is implemented in Bitcoin and similar cryptocurrencies. We propose that in addition to institutions and standards that seek to safeguard certain areas of transparency and trust, algorithmic solutions may be extended into many of the areas where the processing and exchange of information is involved. That applies to most aspects of economic activity, to personal information, ideas, contracts, secure updates, secure backups (of external data or even of DNA and brain data), secure 'undo' operations.

5. Conclusions

We first examined historic developments that have been considered progress. From that, we derived specific requirements for positive future developments that are agreeably considered progress. From that perspective, we then analyzed aspects of humanity in terms of risks and applicable solutions.

We determined that a good future has a couple of key requirements: Humanity continues to exist as it develops, and Humanity succeeds by helping its constituent members thrive. That thriving is expressed in terms of fairness in the availability of opportunity for development, achievement and expression.

We considered solutions to problems and implementations of those solutions that mitigate risks intrinsic to many aspects of humanity. Throughout the endeavor, we found several shared requirements and common fields in which development should be emphasized. Briefly, we might label the shared requirements: *Adaptability*, *Access* and *Transparency*.

Firstly, this means that we should emphasize technological developments that enable us to make humans adaptable, to overcome fundamental human limitations. Such technology involves being

able to acquire data about ourselves and to re-engineer ourselves (biologically and otherwise). Programs such as the European Human Brain Project and the BRAIN Initiative in the US are a promising start, although the magnitude of the task demands even greater attention and effort from government, academia and private industry. Secondly, we should emphasize the facilitation of universal access to new developments, especially those that improve human adaptability. Thirdly, we should press for transparency in crucial interactions. The length restrictions for this paper bar a closer examination of each point, though there are of course many more interesting details to address. Government, international standards organizations and industry groups, as well as informed interest from the general public can set expectations for broad access and transparency.

Finally, an important realization is that our understanding is constrained, that the unfolding of events is a complex process during which novel insights arise. It is therefore very likely that analyses of the nature of humanity and of desired conditions will lead to modified goals. Changes are especially likely as the constituents of humanity include an increasing number of members who experience motivations and drives that have (purposeful and sometimes exploratory) differences from those we typically experience today. Greater diversity of thought is a likely consequence of combining an increasing understanding of mind with the technology to access, modify or design mental functions. The analyses and proposals presented here are attempted with the acknowledged horizon of current understanding. Actual progress will involve many step-wise iterations of development, assessment and recalibration.

References

Berger, T. W.; Song, D.; Chan, R. H.; Marmarelis, V. Z.; LaCoss, J.; Wills, J.; Hampson, R. E.; Deadwyler, S. A.; and Granacki, J. J. 2012. A hippocampal cognitive prosthesis: multi-input, multi-output nonlinear modeling and VLSI implementation. *IEEE Trans Neural Syst Rehabil Eng* 20(2):198–211.

Dennet, D. 2005. *Darwin's Dangerous Idea*. New York, NY: Touchstone Press.

Eckersley, P., and Sandberg, A. 2013. Is Brain Emulation Dangerous? *Journal of Artificial General Intelligence* 4(3):170–194.

Koene, R. 2012a. Experimental Research in Whole Brain Emulation: The Need for Innovative *In-Vivo* Measurement Techniques. *Special Issue of the International Journal of Machine Consciousness* 4(1). doi: 10.1142/S1793843012500047.

Koene, R. 2012b. Fundamentals of whole brain emulation: State, transition and update representations. *International Journal of Machine Consciousness* 4(1).

Koene, R. 2014. Supporting the complex requirements of a long-term project for whole brain emulation. In *Presented at the 2014 Conference of the Mormon Transhumanist Association*. Salt Lake City, UT: Mormon Transhumanist Association.

Seo, D.; Carmena, J.; Rabaey, J.; Alon, E.; and Maharbiz, M. 2013. Neural Dust: An Ultrasonic, Low Power Solution for Chronic Brain-Machine Interfaces. *arXiv:1307.2196*.

Sheehan, M. 2000. *The Balance of Power: History & Theory*. Routledge.

Wesson, P. 1990. Cosmology, extraterrestrial intelligence, and the resolution of the Fermi-Hart paradox. *Quarterly Journal of the Royal Astronomical Society* 31:161–170.

Yudkowski, E. 2008. Artificial Intelligence as a Positive and Negative Factor in Global Risk. In *Global Catastrophic Risk*. Oxford University Press.