Differential Intellectual Progress as a Positive-Sum Project

Brian Tomasik
Foundational Research Institute
brian.tomasik@foundational-research.org

Abstract

Fast technological development carries a risk of creating extremely powerful tools, especially AI, before society has a chance to figure out how best to use those tools in positive ways for many value systems. Suffering reducers may want to help mitigate the arms race for AI so that AI developers take fewer risks and have more time to plan for how to avert suffering that may result from the AI's computations. The AI-focused work of the Machine Intelligence Research Institute (MIRI) seems to be one important way to tackle this issue. I suggest some other, broader approaches, like advancing philosophical sophistication, cosmopolitan perspective, and social institutions for cooperation.

As a general heuristic, it seems like advancing technology may be net negative, though there are plenty of exceptions depending on the specific technology in question. Probably advancing social science is generally net positive. Humanities and pure natural sciences can also be positive but probably less per unit of effort than social sciences, which come logically prior to everything else. We need a more peaceful, democratic, and enlightened world before we play with fire that could cause potentially permanent harm to the rest of humanity's future.

Contents

1 Introduction

> The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom. – Isaac Asimov

> The unleashed power of the atom has changed everything save our modes of thinking [...]. – Albert Einstein

Technology is an inherently double-edged sword: With great power comes great responsibility, and discoveries that we hope can help sentient creatures also have the potential to result in massive suffering. João Pedro de Magalhaes calls this "Alice's dilemma" (Magalhaes, 2004) and notes that "in the same way technology can save lives and enrich our dreams, it can destroy lives and generate nightmares."

In "Intelligence Explosion: Evidence and Import" (Muehlhauser and Salomon, 2012), Luke Muehlhauser and Anna Salamon propose "differential intellectual progress" as a way to reduce risks associated with development of artificial intelligence. From *Facing the Intelligence Explosion*:

> Differential intellectual progress consists in prioritizing risk-*reducing* intellectual progress over risk-*increasing* intellectual progress. As applied to AI risks in particular, a plan of differential intellectual progress would recommend that our progress on the scientific, philosophical, and technological problems of AI *safety* outpace our progress on the problems of AI *capability* [...].

I personally would replace "risk" with "suffering" in that quote, but the general idea is clear.

2 Encouraging more reflection

Differential intellectual progress is important beyond AI, although because AI is likely to control the future of Earth's light cone absent a catastrophe before then, ultimately all other applications matter through their influence on AI.

At a very general level, I think it's important to inspire deeper philosophical circumspection. The world is extremely complex, and making a positive impact requires a lot of knowledge and thought. We need more minds exploring big-picture questions like

- What kinds of futures do we want to see and want to avoid? What are their probabilities?
- How much control do we have over different aspects of the future? Which are mostly inevitable and which are more path-dependent?
- How can we avoid overconfidence and optimism bias in our expectations? Are there interventions that can be helpful across a broad range of possible scenarios?
- What political, social, and cultural institutions can we build to more reliably promote mutually beneficial cooperation?

As these questions suggest, greater reflectiveness by humanity can be a positive-sum (i.e., Pareto-improving) enterprise, because a more slow, deliberative, and clear-headed world is one in which all values have better prospects for being realized. In an AI arms race, there's pressure to produce *something* that can win, even if it's much less good than what your team would ideally want and gives no consideration to what the other teams want. If the arms race can be constrained, then there's more time to engage in positive-sum compromise on how AI should be shaped. This benefits all parties in expectation, including suffering reducers, because AIs built in a hurry are less likely to include safety measures against sentient science simulations, suffering subroutines (Tomasik, 2014), and so on.

3 Ideas for improving reflectiveness

MIRI does important work on philosophical and strategic issues related to AI and has written much on this topic. Below I discuss some other, broader approaches to differential intellectual progress, but in general, it's plausible that MIRI's direct focus on AI is among the most effective.

3.1 Liberal-arts education

The social sciences and humanities contain a wealth of important insights into human values, strategies for pro-social behavior, and generally what philosopher Nick Bostrom calls "crucial considerations" for understanding how the universe works and how to make a positive impact on it. It's good to encourage people to explore this material, such as through liberal-arts education.
Ralph Nader:

> The liberal arts are really the core of higher education. Vocational education is an instrument, but the liberal arts represent the best of our values and they develop of critical thinking[. ...T]he liberal arts and the humanities and social sciences are so critical when higher education is often viewed primarily as vocational.

Of course, a pure focus on humanities or social sciences is not a good idea either, because the hard sciences teach a clarity of thinking that can dissolve some of the confusions that afflict standard philosophy. Moreover, since one of the ultimate goals is to shape technological progress in more positive and cooperative directions, reflective thinkers need a deep understanding of science and technology, not just of David Hume and Peter Singer.

3.2 Big-picture, cosmopolitan thinking

Beyond what students learn in school, there's opportunity to expand people's minds more generally. When scientists, policy makers, voters, and other decision-makers are aware of more ways of looking at the world, they're more likely to be open-minded and consider how their actions affect all parties involved, even those who may feel differently from themselves. Tolerance and cosmopolitan understanding seem important for reducing zero-sum "us vs. them" struggles and realizing that we can learn from each other's differences – both intellectually and morally.

TED talks, Edge, and thousands of other forums like these are important ways to expand minds, advance social discourse on big-picture issues, and hopefully, knock down boundaries between people.

While science popularization helps inform non-experts of what's coming and thereby advance insight into crucial considerations for how to proceed, it also carries the risk of simultaneously encouraging more people to go into scientific fields and produce discoveries faster than what society can handle. The net balance is not obvious, though I would guess that for many "pure" sciences (math, physics, ecology, paleontology, etc.), the net balance is positive; for those with more technological application (computer science, neuroscience, and of course, AI itself), the question is murkier.

3.3 Effective altruism

Expanding the effective-altruist (EA) movement is another positive-sum activity, in the sense that EAs aim to help answer important questions about how best to shape the future in ways that can benefit many different groups. Of course, the movement is obviously just one of many within the more global picture of efforts to improve the world, and it's important to avoid insular "EA vs. non-EA" dichotomies.

3.4 Improved public-policy epistemology??

Carl Shulman suggests the following ideas:

- Enhance decision-making and forecasting capabilities with things like the IARPA forecasting tournaments, science courts, etc, to improve reactions to developments including AI and others (recalling that most of the value of MIRI in [Eliezer Yudkowsky's] model comes from major institutions being collectively foolish or ignorant regarding AI going forward)
- Prediction markets, meta-research, and other institutional changes[.]

These and related proposals would indirectly speed technological development, which is a counter-consideration. Also, if used by a single national government, could they not accelerate arms races? Even if positive, it's not clear these approaches have the same value for negative-leaning utilitarians specifically as the other, more philosophical interventions, which seem more likely to encourage compassion and tolerance.

4 Are these meta things cost-effective?

Is encouraging philosophical reflection in general plausibly competitive with more direct work to explore the philosophical consequences of AI? My guess is that direct work like MIRI's is more important per dollar. That said, I doubt the difference in cost-effectiveness is vast, because everything in society has flow-through effects on everything else, and as people become more philosophically sophisticated and well-rounded, they have a better chance of identifying the most important focus areas, of which AI philosophy is just one. Another important focus area could be, for exam-

ple, designing international political structures that can make cooperative work on AI possible, thereby reducing the deadweight loss of unconstrained arms race. There are probably many more such interventions yet to be explored, and generally encouraging more thought on these topics is one way to foster such exploration.

Part of my purpose in this discussion was not to propose a highly optimized charitable intervention but merely to suggest some tentative conclusions about how we should regard the side-effects of other things we do. For example, should I Like intellectually reflective material on Facebook and YouTube? Probably. Should I encourage my cousin to study physics + philosophy or electrical engineering? These considerations push slightly more for physics + philosophy than whatever your prior recommendation might have been. And so on.

5 Idealism meets competitive constraints

Many of the ideas suggested in this piece are cliché – observations made at graduation ceremonies or moralizing TV programs, about expanding people's minds so that they can better work together in harmony. Isn't this naïve? The future is driven by economic competition, power politics, caveman emotions, and other large-scale evolutionary pressures, so can we really make a difference just by changing hearts and minds?

It's true that much of the future is probably out of our control. Indeed, much of the present is out of our control. Even political leaders are often constrained by lobbyists, donors, and popularity ratings. But a politician's personal decisions can have some influence on outcomes, and of course, the opinions and wealth distribution of the electorate and donors are themselves influenced by ideas in society.

Many social norms arise from convention or expediency, due to the fact that beliefs often follow action rather than precede it. Still, there is certainly leeway in the space of memes toward with society gravitates, and we can tug on them, either directly or indirectly. The founders of the world's major religions had an immense and non-inevitable impact on the course of history. The same is true for other writers and thinkers from the past and present.

Another consideration is that we don't want selective reflectiveness. For example, suppose those currently pursuing fast technological breakthroughs kept going at the same pace, while the rest of society slowed down to think more carefully about how to proceed. This would potentially make things worse because then circumspection would have less chance of winning the race. Rather, what we'd like to see is an across-the-board recognition of the need for exploring the social and philosophical side of how we want to use future technology – one that can hopefully influence all parties in all countries.

As a specific example, say the US slowed down its technological growth while China did not. China currently cares less about animal welfare and generally has more authoritarian governance, so even from a non-ethnocentric viewpoint, it could be slightly worse for China to control the future. But my guess is that this consideration is very small compared with the direct, potentially adverse effect of faster technology on the whole planet, especially since most non-military technological progress isn't confined within national boundaries. China could catch up to America's level of humane concern in a few decades anyway, and the bigger issue seems to be how fast the world as a whole moves. Also, in the case of military technology, the US tends to set the pace of innovation, and probably slower US military-tech growth would reduce the pressure for military-tech development by other countries.

6 Areas where the sign is unclear

6.1 Faster technology

It's not always the case that accelerated technology is more dangerous. For example, faster technology in certain domains (e.g., the Internet that made Wikipedia possible) accelerates the spread of wisdom. Discoveries in science can help us reduce suffering faster in the short term and improve our assessment for which long-term trajectories humanity should pursue. And so on. Technology is almost always a mixed bag in what it offers, and faster growth in some areas is probably very beneficial. However, from a macro perspective, the sign is less clear.

6.2 Education

Promoting education wholesale is another double-edged sword because it speeds up technology as well as wisdom. However, differentially advancing cross-disciplinary and philosophically minded education seems generally like a win for many value systems at once, including suffering reduction.

6.3 Cognitive enhancement

In "Intelligence Amplification and Friendly AI", Luke Muehlhauser enumerates arguments why improving cognitive abilities might help and hurt chances for controlled AI. Nick Bostrom (Bostrom, 2014) reviews similar considerations in Ch. 14 of *Superintelligence:*

Paths, Dangers, Strategies.

6.4 Transhumanism

Benefits:

- Transhumanists recognize the importance of thinking about the future ahead of time.

Drawbacks:

- Transhumanists often want to accelerate the future, perhaps due to starry-eyed optimism.
- Transhumanists typically support colonizing space and spreading sentience far and wide, even though this likely will mean a massive increase in expected suffering.

6.5 Economic growth

A similar double-edged sword is economic growth, though perhaps less dramatically. One primary effect of economic growth is technological growth, and insofar as we need more time for reflection, this seems to be a risk. On the other hand, economic growth has several consequences that are more likely positive, such as

- Increasing international trade, with the side effect of making people more sympathetic to those of other nationalities and reducing odds of inter-country warfare
- Promoting democracy, which is a powerful way to resolve disputes among conflicting factions
- Enhancing stability and therefore concern for longer-term outcomes, with reduced unilateral risk-taking
- Allowing for more intellectual awareness and reflection on important questions generally.

That said, these seem like properties that result from the *absolute amount* of economic output rather than the *growth rate* of the economy. It's not controversial that a richer world will be more reflective, but the question is whether the world would be more reflective *per unit of GDP and technology* if it grew faster or slower.

As a suggestive analogy, slower-growing crystals have fewer defects. More slowly dropping the temperature in a simulated-annealing algorithm allows for finding better solutions. In the case of economic growth, one might say that if people have more time to adapt to a given level of technological power, they can make conditions better before advancing to the next level. So, for example, if the current trends (Pinker, 2011) toward lower levels of global violence continue, we'd rather wait longer for growth, so that the world can be more peaceful when it happens. Of course, some of that trend toward peace may itself be due to economic growth.

Imagine if people in the Middle Ages developed technology very rapidly, to the verge of building general AI. Sure, they would have improved their beliefs and institutions rapidly too, but those improvements wouldn't have been able to compete with the centuries of additional wisdom that our actual world got by waiting. The Middle-Age AI builders would have made worse decisions due to less understanding, less philosophical sophistication, worse political structures, worse social norms, etc. The arc of history is almost monotonic toward improvements along these important dimensions.

A counterargument is that conditions are pretty good right now, and if we wait too long, they might go in worse directions in the meantime, such as because of another Cold War between the US and China. Or, maybe faster economic growth means more trade sooner, which helps prevent wars in the short run. (For example, would there not have been a Cold War if the US and Soviet Union had been important trading partners?) A friend tells me that Peter Thiel believes growth is important for cooperation because in a growth scenario, incentives are positive-sum, while in stagnation, they're more zero-sum. Carl Shulman notes that "Per capita prosperity and growth in per capita incomes are associated (Friedman, 2005) with more liberal postmaterialist values, stable democracy, and peace." Faster growth by means other than higher birth rates might increase GDP per capita because growth would happen more rapidly than population could keep up.

Suppose AI would arrive when Earth reached some specific level of GDP. Then even if we saw that faster growth correlated with faster increases in tolerance, cooperation, and wisdom, this wouldn't necessarily mean we should push for faster growth. The question is whether some percent increase in GDP gives more increase in wisdom when the growth is faster or slower.

Alternatively, in a model where AI arrives after some amount of cumulative GDP history for Earth, regardless of whether there has been growth, then if zero GDP growth meant zero moral growth (which is obviously unrealistic), then we'd prefer to have more GDP growth so that we'd have more wisdom when AI arrived.

Another relevant consideration Carl Shulman pointed out is that growth in AI technology specifically may only be loosely coupled with economic growth overall. Indeed, if slower growth caused wars that triggered AI arms races, then slower economic growth would mean faster AI. Of course, some take the opposite view: Environmentalists might claim that faster growth would mean more future catastrophes like cli-

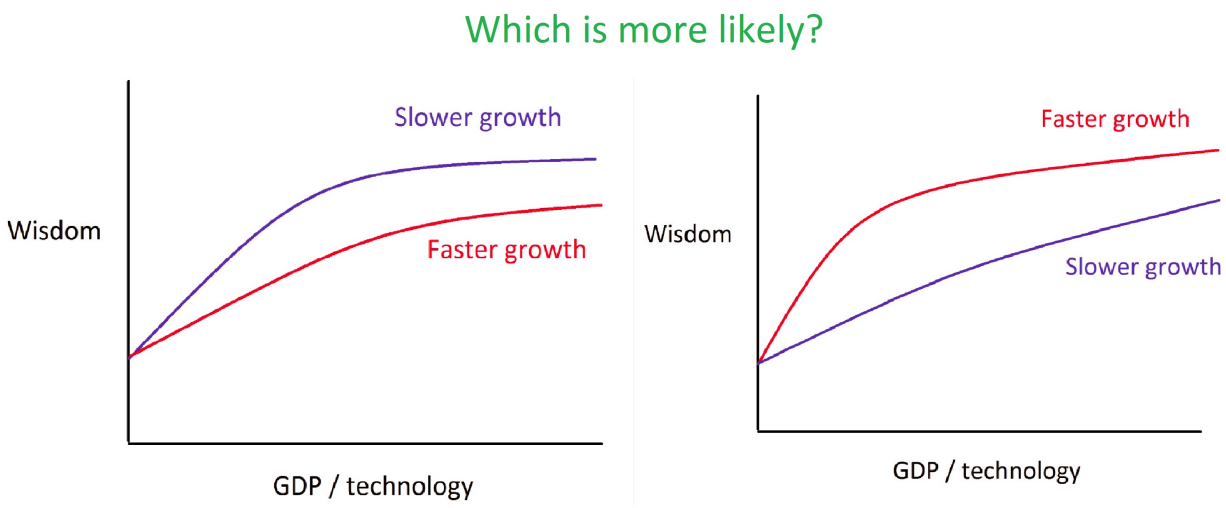

Figure 1: *How does wisdom per unit of GDP and technology depend on the growth rate of that GDP and technology?*

mate change and water shortages, and these would lead to more wars. The technologists then reply that faster growth means faster ways to mitigate environmental catastrophes. And so on.

Also, a certain level of economic prosperity is required before a country can even begin to amass dangerous weapons, and sometimes an economic downturn can push the balance toward "butter" rather than "guns." David E. Jeremiah predicted (Jeremiah, 1995) that "Conventional weapons proliferation will increase as more nations gain the wealth to utilize more advanced technology." In the talk "Next steps in nuclear arms control," Steven Pifer suggested that worsening economic circumstances might incentivize Russia to favor disarmament agreements to reduce costly weapons that it would struggle to pay for. Of course, an opposite situation might also happen: If the budget is tight, a country might, when developing new technologies, strip away the "luxuries" of risk analysis, making sure the technologies are socially beneficial, and so on.

More development by Third World countries could mean that more total nations are able to compete in technological arms races, making coordination harder. For instance, many African nations are probably too poor to pursue nuclear weapons, but slightly richer nations like India, Pakistan, and Iran can do so. On the other hand, development by poor nations could mean more democracy, peace, and inclination to join institutions for global governance.

The upshot is unclear. In any event, even if faster economic growth were positive, it seems unlikely that advancing economic growth would be the most cost-effective intervention in most cases, especially since there are strong competitive and political pressures pushing for it already. Of course, there are some cases where the political pressures are stronger the other way (e.g., in opposing open borders for immigrants), when there's a perceived conflict between national and global economic pie.

Also, while the effects of "economic growth" as an abstract concept may be rather diffuse and double-edged, any particular intervention to increase economic growth is likely to be targeted in a specific direction where the differential impact on technology vs. wisdom is more lopsided.

6.5.1 Wars and arms races may dominate

Quoting Kawoomba on LessWrong:

> R&D, especially foundational work, is such a small part of worldwide GDP that any old effect can dominate it. For example, a "cold war"-ish scenario between China and the US would slow economic growth – but strongly speedup research in high-tech dual-use technologies.
>
> While we often think "Google" when we think tech research, we should mostly think DoD in terms of resources spent – state actors traditionally dwarf even multinational corporations in research investments, and whether their [investments] are spurned or spurred by a slowdown in growth (depending on the non-specified cause of said slowdown) is anyone's guess.

Luke_A_Somers followed up:

> Yes - I think we'd be in much better shape with high growth and total peace than the other way

> around. Corporations seem rather more likely to be satisfied with tool AI (or at any rate AI with a fixed cognitive algorithm, even if it can learn facts) than, say, a nation at war.

The importance of avoiding conflict and arms races is elaborated in "How Would Catastrophic Risks Affect Prospects for Compromise?"

In general, warfare is a major source of "lost surplus" for many value systems, because costs are incurred by each side, resources are wasted, and the race may force parties to take short-sighted actions that have possibly long-term consequences for reducing surplus in the future. Of course, it seems like many consequences of war would be temporary; I'm not sure how dramatic the "permanently lost future surplus" concern is.

It's not obvious that economic growth would reduce the risk of arms races. Among wealthy countries it might, since more trade and prosperity generally lead to greater inter-dependence and tolerance. On the other hand, more wealth also implies more disposable income to spend on technology. Economic growth among the poorest countries could exacerbate arms races, because as more countries develop, there would be more parties in competition. (For instance, there's no risk of arms races between the developed world and poor African nations in the near future.) But international development might also accelerate global coordination.

7 There are many exceptions

My assessments in the previous section are extremely broad generalizations. They're akin to the claim that "girls are better at language than boys" – true on average, but the distributions of individual measurements have huge overlap. Likewise with my statements about technology and social institutions: There are plenty of advances in each category that are very good and plenty that are very bad, and the specific impact of an activity may be very different from the average impact of the category of which it's a part. The main reason to generalize about categories as a whole is in order to make high-level assessments about policies, like "Should we support more funding of engineering programs in the US?" When evaluating a particular activity, like what you do for your career, a specific analysis of that activity will be far more helpful than just labeling it "technology" or "social science".

8 Technologies that are probably bad to accelerate

8.1 Computer hardware

In *Superintelligence* (Ch. 14), Bostrom outlines reasons why faster hardware is likely to make AI control harder:

- It may accelerate general AI, giving less time for reflection and cooperation.
- It may favor more brute-force and less transparent forms of AI, which seem harder to predict and align with our values. (I would add that this is debatable depending on how the brute force was applied. Brain emulations are a type of brute-force AI that may actually be easier to control. Even minds evolved via genetic algorithms might resemble humans in important ways, more so than strictly mathematical AIs.)
- It may create a "computing overhang", i.e., more hardware capacity than software know-how for developing AI. That means that when crucial insights for AI software are developed, the takeoff is likely to be more abrupt.
- It would lower the resource requirements for creating general AI, potentially allowing more parties to enter an AI arms race, including more extreme groups.
- While some computer technologies like the Internet may accelerate wisdom, it's unclear how much marginal hardware improvements would further contribute along such dimensions.

8.2 Artificial consciousness

Artificial consciousness seems net harmful to advance because:

- It helps accelerate AI in general.
- It's better to wait until society is wiser and more humane before conscious computer agents are developed. For instance, imagine violent video games that are marketed for their ability to generate conscious, lifelike enemies.

Steve Grand defended his work toward artificially conscious creatures on the following grounds:

> This is what I care about. I want to help us find out what it means to be conscious and I want to challenge people to ask difficult questions for themselves that they can't do with natural life because of their unquestioned assumptions and prejudices. But we really are talking about creatures that are incredibly simple by natural standards. What I'm trying to explore is what it means to have an imagination. Not a rich one like humans

> have, but at all. The only way to find that out is to try to build one and see why it is needed and what it requires. And in doing so I can help people to ask questions about who they are, who other creatures are, and what it means to be alive. That's not such a bad thing, is it?

This resembles an argument that Bostrom calls an instance of "second-guessing" in Ch. 14 of *Superintelligence*: basically, that in order to get people to take the risks of a technology seriously, you need to advance work on the technology, and it's better to do so while the technology has limited potential so as to bound risks. In other words, we should advance the technology before a "capability overhang" builds up that might yield more abrupt and dangerous progress in the technology. Bostrom and I are both skeptical. Armed with such a defense, one can justify any position on technological speed because either we (a) slow the technology to leave more time for reflection or (b) accelerate the technology so that others will take risks more seriously while the risks remain manageable.

In the case of artificial consciousness, we should advance the public discussion by focusing our energies on philosophy rather than on the technical details of building software minds. There's already enough technical work on artificial consciousness to fuel plenty of philosophical dialogue.

9 Caveats: When are changes actually positive-sum?

9.1 Positive-sum in resources does not mean positive-sum in utility

Improved social wisdom is positive-sum in terms of the resources it provides to different value systems: Because they know more, they can better accomplish each of their goals. They have more tools to extract value from their environment. However, it's not always the case that an action that improves the resources of many parties also improves the utility of each of those parties. Exceptions can happen when the goals of the parties conflict.

Take a toy example. Suppose Earth contained only Stone Age humans. One tribe of humans thought the Earth was beautiful in its untouched natural state. Another tribe felt that the Earth should be modified to better serve human economic interests. If these humans remained forever in the Stone Age, without greater wisdom, then the pro-preservation camp would have gotten its way by default. In contrast, if you increased the wisdom of both tribes – equally or even with more wisdom for the pro-preservation tribe – then it would now be at least possible for the pro-development tribe to succeed. Thus, despite a positive-sum increase in wisdom, the pro-preservation tribe is now worse off in expected utility.

However, this example is somewhat misleading. A main point of the present essay was to highlight the potential risks of greater technology, and one reason wisdom is beneficial is that it better allows both sides to cooperate and find solutions to reduce expected harms. For example, absent wisdom, the pro-development people might just start a war with the pro-preservation people, and if the pro-development side won, the pro-preservation side would have its values trashed. If instead both sides agreed to undertake modest development with safeguards for nature preservation, then each side could end up better off in expectation. This is an example of the positive-sum *utility* benefits that wisdom can bring.

Perhaps there are some examples where wisdom itself, not just technology, causes net harm to a certain ideology, but it seems like on the whole wisdom usually is positive-sum even in utility for many factions.

9.2 Are changes determined by fractions of people or by absolute numbers?

The main intuition why wisdom and related improvements should be positive-sum is that they hold constant the fraction of people with different values and instead distribute more "pie" to people with each set of values. This fractional view of power makes sense in certain contexts, such as in elections where the proportion of votes is relevant. However, in other contexts it seems that the absolute number of people with certain values is the more appropriate measure.

As an example, consider the cause of disaster shelters that serve to back up civilization following near-extinction-level catastrophes. Many altruists support disaster shelters because they want humanity to colonize space. Suffering reducers like me probably oppose disaster shelters because shelters increase the odds of space colonization without correspondingly increasing the odds of more humane values. If work towards disaster shelters is proportional to (# of people in favor) minus (# of people opposed), and if, say, 90% of people support them by default, then greater education might change

$$(10\ in favor) - (1\ opposed) = 9\ net$$

to

$$(1000\ in favor) - (100\ opposed) = 900\ net,$$

which is a 100-fold increase in resources for disaster shelters. This makes the suffering reducers worse off, so in this case, education was not positive-sum.

My intuitions that wisdom, education, cooperation, etc. are *in general* positive-sum presupposes that most of the work that people do as a result of those changes is intrinsically positive for both happiness increasers and suffering reducers. Disaster shelters seem to be a clear exception to this general trend, and I hope there aren't too many other exceptions. Suffering reducers should keep an eye out for other cases where seemingly positive-sum interventions can actually hurt their values.

See also

"Progress and Prosperity" by Paul Christiano

References

Bostrom, Nick. 2014. *Superintelligence. Paths, Dangers, Strategies.* Oxford: Oxford UP. Print.

De Magalhaes, João Pedro. "Alice's dilemma." *Futures* 36 (2004): 85-89. doi:10.1016/S0016-3287(03) 00141-1.

Friedman, Benjamin M. *The Moral Consequences of Economic Growth.* New York: Knopf, 2005. Print.

Jeremiah, David E. "Nanotechnology and Global Security." *Nanotechnology.* Technology Strategies and Alliances, 9 Nov. 1995. Web. 03 Mar. 2016. http://www.zyvex.com/nanotech/nano4/jeremiahPaper.html.

Muehlhauser, Luke, and Anna Salamon. "Intelligence Explosion: Evidence and Import." *The Frontiers Collection Singularity Hypotheses* (2012): 15-42. http://intelligence. org/files/IE-EI.pdf.

Pinker, Steven. *The Better Angels of Our Nature: Why Violence Has Declined.* New York: Viking, 2011. Print.

Tomasik, Brian. "Do Artificial Reinforcement-Learning Agents Matter Morally?" *ArXiv.org.* N.p., 30 Oct. 2014. Web. 03 Mar. 2016. http://arxiv.org/abs/ 1410.8233v1.