Chapter 2A
Robin Hanson on Muehlhauser and Salamon's "Intelligence Explosion: Evidence and Import"

Muehlhauser and Salamon [M&S] talk as if their concerns are particular to an unprecedented new situation: the imminent prospect of "artificial intelligence" (AI). But in fact their concerns depend little on how artificial will be our descendants, nor on how intelligence they will be. Rather, Muehlhauser and Salamon's concerns follow from the general fact that accelerating rates of change increase intergenerational conflicts. Let me explain.

Here are three very long term historical trends:

1. Our total power and capacity has consistently increased. Long ago this enabled increasing population, and lately it also enables increasing individual income.
2. The rate of change in this capacity increase has also increased. This acceleration has been lumpy, concentrated in big transitions: from primates to humans to farmers to industry.
3. Our values, as expressed in words and deeds, have changed, and changed faster when capacity changed faster. Genes embodied many earlier changes, while culture embodies most today.

Increasing rates of change, together with constant or increasing lifespans, generically imply that individual lifetimes now see more change in capacity and in values. This creates more scope for conflict, wherein older generations dislike the values of younger more-powerful generations with whom their lives overlap.

As rates of change increase, these differences in capacity and values between overlapping generations increase. For example, Muehlhauser and Salamon fear that their lives might overlap with

> [descendants] superior to us in manufacturing, harvesting resources, scientific discovery, social charisma, and strategic action, among other capacities. We would not be in a position to negotiate with them, for [we] could not offer anything of value [they] could not produce more effectively themselves. … This brings us to the central feature of [descendant] risk: Unless a [descendant] is specifically programmed to preserve what [we] value, it may destroy those valued structures (including [us]) incidentally.

The quote actually used the words "humans", "machines" and "AI", and Muehlhauser and Salamon spend much of their chapter discussing the timing and likelihood of future AI. But those details are mostly irrelevant to the concerns expressed above. It doesn't matter much if our descendants are machines or biological meat, or if their increased capacities come from intelligence or raw physical power. What matters is that descendants could have more capacity and differing values.

Such intergenerational concerns are ancient, and in response parents have long sought to imprint their values onto their children, with modest success.

Muehlhauser and Salamon find this approach completely unsatisfactory. They even seem wary of descendants who are cell-by-cell emulations of prior human

brains, "brain-inspired AIs running on human-derived "spaghetti code", or 'opaque' AI designs …produced by evolutionary algorithms." Why? Because such descendants "may not have a clear 'slot' in which to specify desirable goals."

Instead Muehlhauser and Salamon prefer descendants that have "a transparent design with a clearly definable utility function," and they want the world to slow down its progress in making more capable descendants, so that they can first "solve the problem of how to build [descendants] with a stable, desirable utility function."

If "political totalitarians" are central powers trying to prevent unwanted political change using thorough and detailed control of social institutions, then "value totalitarians" are central powers trying to prevent unwanted value change using thorough and detailed control of everything value-related. And like political totalitarians willing to sacrifice economic growth to maintain political control, value totalitarians want us to sacrifice capacity growth until they can be assured of total value control.

While the basic problem of faster change increasing intergenerational conflict depends little on change being caused by AI, the feasibility of this value totalitarian solution does seem to require AI. In addition, it requires transparent-design AI to be an early and efficient form of AI. Furthermore, either all the teams designing AIs must agree to use good values, or the first successful team must use good values and then stop the progress of all other teams.

Personally, I'm skeptical that this approach is even feasible, and if feasible, I'm wary of the concentration of power required to even attempt it. Yes we teach values to kids, but we are also often revolted by extreme brainwashing scenarios, of kids so committed to certain teachings that they can no longer question them. And we are rightly wary of the global control required to prevent any team from creating descendants who lack officially approved values.

Even so, I must admit that value totalitarianism deserves to be among the range of responses considered to future intergenerational conflicts.