UAI: Capturing the Promise of AI

By Mark Cummings, Ph.D.

GenAI productivity is driving the US stock market to record highs in spite of the economic headwinds from tariffs, layoffs, and immigration problems. This is a reflection of the tremendous productivity benefits GenAI promises. But there are challenges to achieving those promises. The challenges lie in two areas: potential damage that we need to find a way to minimize; and difficulties in implementation. A recent effort by a small group in the GenAI tech industry is a good start on minimizing damage. The group addresses one of the real problems. In doing so, it also points to a model that can be extended to address the rest. That is, to be built upon and expanded to cover the full range of potentially damaging effects and difficulties in implementation. It is important to quickly develop a consensus in this area and creating such a group may be key to achieving that. For purposes of discussion, we call this group UAI (Union of Artificial Intelligence). Below we start this discussion and provide a link for those interested in learning more, joining the discussion, helping to create UAI, etc.

The AI 2027 Model

A small group composed of people who had experience working inside the companies producing the frontier GenAI models shared a particular concern. Some of them were able to leave their jobs and get philanthropic donations to act as the core of the group. Others volunteered anonymously.

The concern they shared is titled “alignment”. Alignment is the problem of making sure that the AI does what it is supposed to do and doesn’t do what it is not supposed to do. This is achieved by making sure that the AI is “aligned” with a specification of what it is supposed to do and what it is not supposed to do. This specification is called the spec. The group is concerned that AI’s will get out of alignment. Out of alignment AI’s will decide that humanity is getting in the way of the AI’s achieving their own goals. Therefore the AI’s will get rid of humanity.

The group developed a convincing scenario founded on technical principles that showed how this could hit the key turning point by 2027 that would make the end of humanity inevitable. In later YouTube video clips they said it is more likely to be 2028.

The group also presented a scenario in which AI’s produced extreme benefits to humanity and the alignment problem is avoided. Part of the problem is the AI development race underway and it’s business and geopolitical drivers. The group has some specific recommendations for what needs to be done to achieve the positive outcome.

The core group members are identified as co-authors on the web site and have given a number of video interviews. To develop and test their scenarios, the core group relied on input, review and feedback from a larger group of people active in the AI industry. This larger group stayed anonymous. This anonymity is important because recognized participation could have put jobs and careers in jeopardy.

The AI 2027 report received a lot of attention in the AI industry. There are some indications that leaders of companies developing frontier AI systems took it seriously. The AI race is still underway. But, there are some indications that some of the AI 2027 recommendations are being quietly implemented.

Thus, the organizational model that AI 2027 developed proved to be effective in its somewhat narrow focus. As a model, it provides a good foundation for an organization to address the

broader challenges: potentially damaging effects; and difficulties in achieving applications of AI that fully deliver on its promise.

Damaging Effects

The alignment problem is real. But there may be a more serious underlying problem as well. Who writes the spec; and how does society change to accommodate the dramatic benefits of AI?

What the spec contains and how it is produced are critical. Making sure that AI's are properly and completely aligned is very important. A possibly bigger question is who creates the spec and what criteria are used? It appears that today the spec is created individually by each of the frontier model creating companies. Each tries to determine what spec is best from their own narrow product / market / revenue perspective. This makes sense from a company point of view. But from a larger societal point of view, it leaves a lot to be desired. Many of the employees of these frontier model companies have a sincere desire to make AI produce the best society wide outcomes. However, these same people may be limited in their ability to affect the spec and may lack some of the non-computer expertise required.

Although AI technical expertise is necessary to develop a well formed spec, it is not sufficient. Expertise in history, economics, philosophy, sociology, anthropology, etc. is also important in developing a spec that truly produces the best possible societal outcomes. A friend of mine characterizes some of the people vocal on this subject as attempting to build a utopia without understanding what a utopia is. That is another way of saying that they lack the historical, economic, philosophical, sociological, anthropological backgrounds to understand what has been tried before and failed.

Some may argue that since most frontier models are trained on a corpus of data that reflects most of the experience of humanity, we can depend on AI to fill in the expertise that is missing. Unfortunately, without the spec, it can be dangerous to rely on an AI for this. It could become something akin to the blind leading the blind…

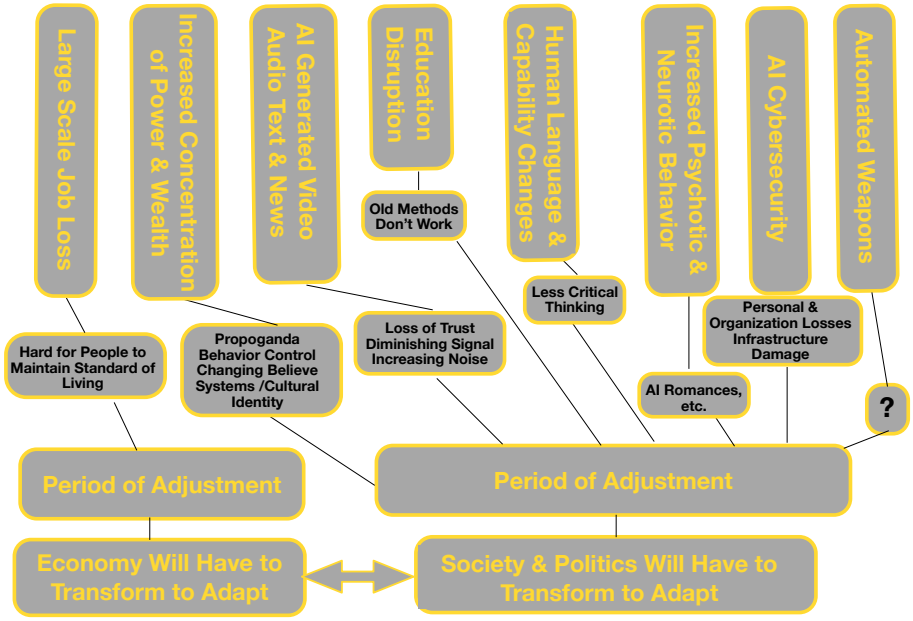
Thus, not only must the spec be well formed, it must be perceived as having been done in a reasonable way. Society at large has to be comfortable that the spec was prepared in a proper fashion. A way that makes people feel comfortable in accepting the result.

Finally, there must be cybersecurity protections sufficient to make sure that the spec can not be overcome, gotten around, or polluted. There has been a lot of work on prompt injection attacks. Recently, AI's have been shown to be susceptible to social engineering attacks. There are likely to more security vulnerabilities that will appear. This is a technical problem that should be within the capability of the frontier model building companies. But, the AI race may create so much pressure that insufficient time, resource, expertise, etc. is applied to the problem. There needs to be some kind of external way to insure that there is adequate security.

Even with a proper spec, AI's are and will create social impact problems. An illustration of these potential problems is shown in the taxonomy below. The taxonomy just lists the societal effects that can be seen from today's vantage point. There may well be others that become apparent as we move forward.

Each of these effects will have a fundamentally disruptive societal impact. Each will require a transition (period of adjustment) from current societal structures to new ones. The challenge is to make this transition in such a way as to to do minimal damage. To minimized the damage and maximize the benefits from AI, these transitions need to be anticipated and mitigation

Taxonomy of Concerning AI Societal Effects



strategies developed for each. Mitigating these societal effects needs the expertise of the professionals in the frontier model building companies to understand and prepare for the AI capabilities in a timely fashion. But that will not be enough. Here again expertise in history, economics, philosophy, sociology, anthropology, etc. will also be important.

The challenge is how to best address concerning societal AI impacts without killing the goose that lays the golden eggs. That is. address in the broadest sense.

Our current society has, in large part, been developed around and accommodating to, the industrial revolution. Since the microprocessor/PC/Internet/Web people have been talking about the Information Age and how it is changing things. GenAI is going to have at least an order of magnitude greater impact. We can either be ahead of the curve or behind it. That is, we can either prepare for the changes coming. Or struggle after the changes hit hard trying to adapt. To the extent we are ahead of the curve, we will minimize the pain that people will feel going through the changes.

Unfortunately, the changes are the result of a complex matrix of factors interacting in previously unseen ways. It can be thought of as trying to put a puzzle together without the picture on the box top.

The good news is that there are many organizations working on pieces of the puzzle. The bad news is that most of these organizations lack some of the expertises needed. That is akin to missing some of the pieces of the puzzle. There are the companies developing the tech. Those applying the tech. Those investing in the tech. Governments. Academic organizations. Not For Profits. Etc. Some are motivated by self interest and not societal well being. Some are well intentioned. Some are not.

Each of these organizations, and some of the people in them, have a portion of the expertise needed. But not all of the expertise. So, for example, some in the AI tech community talk about a guaranteed minimum income as a way to mitigate the effects of AI job losses. But the same people don't have the expertise necessary to create realistic proposals about how to realize that vision in all the different economic, political and social situations around the world.

Furthermore, each of these organizations is subject to outside pressures that can have a big influence on what they can do and say. Some of the people who could make significant contributions are afraid that if they do so, their jobs could be in jeopardy. The situation is

further complicated by political and economic ideologies developed in response to the industrial revolution that are no longer relevant today.

As a result, none of these organizations, or people within them, are optimally positioned to figure out how to stay ahead of the curve - to minimize pain and damage as we go through the transitions.

Capturing the Full Promise of AI

GenAI intelligent agents have the potential for significant productivity improvements. To achieve the full productivity benefits we need to learn how to develop, deploy and secure these intelligent agents. Currently, it appears that we are very low on the learning curve. For example, a recent MIT report says that 95% of GenAI pilots at companies are failing. This indicates that as an industry, we are very low on the learning curve.

The best way to quickly move up the learning curve is to create a way that agent implementers can share experience, develop best practices based on that experience and provide an environment that helps move all up the learning curve. To achieve the full promise of AI, moving up the learning curve more quickly is important.

In past steps of technology evolution, vendors have created vendor specific user groups such as the San Francisco Apple Core, or IBM's SHARE. With AI intelligent agents, it is not unusual for an application developer to use more than one LLM and choose them from different vendors. These LLM choices are based on considerations of best fit for functionality, local resources available, latency, etc. LLM evolution is also increasing at a rapid rate. This can further complicate LLM suite choices.

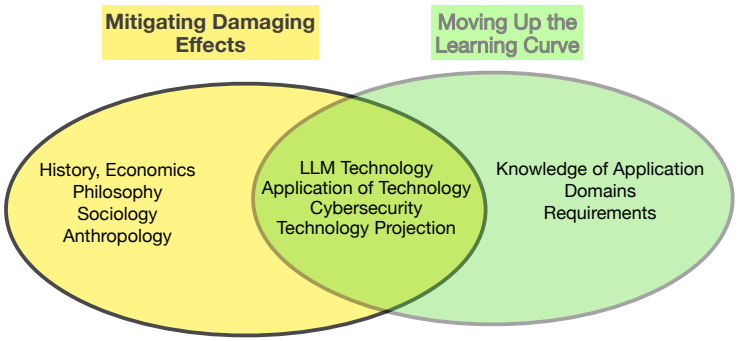
The best way to move up the learning curve more quickly is to provide a way for people to come together and share their experiences - both successes and failures. In previous generations of technology this was done in groups that operated on a principal of coop-etition. That is cooperation on fundamentals that creates a foundation where each participant can compete on application.

Such a cooperative organization could develop tutorials and best practices in a wide ranging set of areas. Examples might include:
- How to select functions for AI agents
- Selecting LLM(s)
- Determining privacy and security requirements
- Meeting up time / reliability requirements
- Handling hallucinations
- Developing Security architectures
- Handling end user acceptance problems
- Handling deployment, maintenance, etc. issues
- Managing Life cycles
- Creating input to vendor requirements

Organization Structure

The overlapping expertises required to address both the potentially damaging effects, and the difficulties in achieving applications of AI can be seen in the illustration below. The expertise requirements for both have a very large area of overlap. So, it makes sense to have a single organizational umbrella that addresses both. Within that umbrella there can be different working groups. How the detailed structure of the working groups is set up, should be decided by the group's membership. The structure and process for creating it should anticipate that

Expertise Required in the Two Problem Areas



there will be changes in working groups as the technology and its applications evolve. For the purpose of discussion, we call this umbrella group UAI.

UAI a Response to the Challenges

What is needed is an organization that is structured so as to have the requisite expertise and ways of avoiding outside pressures. The primary function of the organization is to provide other groups in society with information and recommendations on how best to maximize the benefits of AI while making changes in our social and economic structure necessary to accommodate AI.

In thinking about the structure of UAI, much can be learned from the structure of AI 2027. There needs to be a core group of people. With sufficient funding, some of these people may be full time employees. Others may be part time volunteers working on UAI after normal working hours. Some of these may be anonymous participants because of fear of employment repercussions. Finally, there must be a formal way that those working in UAI reach agreement on what is said and done in the name of the organization. This can be done by consensus. By voting. Or, by a combination of the two.

UAI needs to have key members well grounded and constantly up to date on AI technology. They may come from the leading edge LLM developers, AI application specialists and AI technical people working on applications of the technology in the organizations that employ them. UAI must also have people with expertise in the social sciences, government, etc. It is important to have good representation from each area of expertise in the core group.

UAI should focus on:
- Alignment problem
- Spec problem
- Societal accommodation problem highlighted in the taxonomy
- Application learning curve problem

Each of these may be in different working groups. But care needs to be taken to make sure that there is good cross-communication between the working groups. The cross communication should be aimed at taking full advantage of the synergy between them.

UAI's primary outputs will be information and recommendations:
- Explanations of the technology
    - How the technology is likely to evolve
    - What problems society is likely to encounter as a result
- Recommendations
    - Ways of modifying society to take adjust to AI while assuring good quality of life for all
    - Improvements to both development and applications of the technology

UAI should not focus on stopping nor limiting the development of the technology. It should not become perceived as a group of nay sayers and luddites. The objective is to maximize the benefit from AI while minimizing trouble for people caught in the transitions / disruptions that are inevitable.

The first step in creating a group like UAI is talking about it. This piece is intended to be a step in that direction. To catalyze others to start and build conversations about UAI or something like it.

If you are interested in learning more about this, please go to (url) where you will find updated information and ways to join the conversation

Conclusion

GenAI promises tremendous productivity benefits. But there are challenges to achieving those promises. The challenges lie in two areas: potential damage that we need to find a way to minimize; and difficulties in implementation. A recent effort by a small group in the GenAI tech industry called AI 2027 is a good model for addressing these challenges. That model should be built upon and expanded to cover the full range of potentially damaging effects and difficulties in implementation. For purposes of discussion, we call this expanded version UAI (Union of Artificial Intelligence). If you are interested in learning more about UAI, please go to (url) where you will find updated information and ways to join the conversation.