Managing AI Risk with VUCA and Knowledge Engineering

Joe Glick, Chief Knowledge Scientist, RYAILITI LLC

Part 1 – Why?

The VUCA framework – Volatility, Uncertainty, Complexity and Ambiguity – precisely defines the drivers of all risk. It was defined over half a century ago at West Point, and in recent decades adopted by global management consultants. The more significant and challenging the risks, the more relevant is VUCA, and the risks associated with AI are very significant and challenging. Analytics using the VUCA framework do not eliminate risk but improve the precision of the problem identification and the viability of the mitigation strategy and supports an evidence-based approach to deal with what we cannot know or control. How does it apply to AI?

VOLATILITY – usually associated with external forces that we cannot control, such as geopolitics, markets, and competitors' strategies. With AI we have to add the internal volatility of black box algorithms at massive scale. A senior scientist I knew from Los Alamos Lab talked about "neurotic networks".

UNCERTAINTY – assumptions, premises, and theories for which we have insufficient evidence or cannot test experimentally. With AI we have to add the inherent uncertainty of statistical conclusions, which are core to AI/ML/NLP algorithms. Einstein said, "So far as the laws of mathematics refer to reality, they are not certain. And so far as they are certain, they do not refer to reality".

COMPLEXITY – the real world is highly complex and interconnected. In 2005, as part of its "Get Real" project, DARPA calculated the potential interaction of one million agents as 10^300,000 – a number that could give you a headache if you try and grasp it. Complexity is invisible to an AI algorithm because:

1. It can only see what it has been engineered to find.
2. It works within the narrow limits of the training data.

Neuroscientist Henry Markram, director of the Blue Brain Project, describes the mathematical challenge: "Now it would take a very long time to even try to calculate how many combinations there are (in the brain) when one can chose 10,000 possible genes from a set of 20,000, but what is sure is that it is more than the number of sub-atomic particles in this, and probably every other universe we can imagine."

AMBIGUITY – this risk is particularly significant in Large Language Models. Linguistic meaning is context based and evolving, especially online, which is the primary source of input. This issue is well understood and driving global concern, which is driving emerging regulation, which is inevitably ambiguous.

So, the VUCA framework is ideally suited to guide the architecture of solutions aimed at mitigating and managing AI risk. But how can it be done?

Part 2 – How?

As a reminder, the VUCA framework helps us to prepare for what we cannot predict, prevent, or control even when some elements of those risks can be mitigated. It does not address the standard software related risks that can be mitigated by good management and engineering practices.

Let's begin with an overview of the biomimetic principles and methods that apply to all four risk drivers in AI. I will address the methodology required for each one individually.

Principle 1 – Analysis of complex options and tradeoffs using real-world evidence (RWE) ecosystems. AI risks result from interactions that are invisible to us, part of what Gartner labelled dark data, because RWE is excluded/obscured by efforts to integrate databases and scale systems.

The RWE continuum: 1. Data - 2. Information - 3. Context - 4. Interpretation - 5. Learning - 6. Adaptability

AI developers focus on the first two because they are addressed by IT methods, and that's the easy part. The remaining four require scientific expertise, collaboration, experimentation, objectivity, and adaptability. That's the hard part, especially the last two. How is this addressed by RWE ecosystems?

Principle 2 – Decision analysis leveraging real-world reasoning (RWR) agents.

The RWR continuum: 1. Problem Definition - 2. Identification of Premises and Theories (including algorithms) - 3. Integration of Domain Expertise - 4. Relevance Computation Rules - 5. Dynamic Adaptability Process

The calculations from DARPA and Dr. Markram are RWE that we cannot achieve RWR by means of exhaustive computation of the possible interactions. A biomimetic approach is required that emulates human thinking.

Brain processes are SYSTEMIC and leverage what neuroscientists label PLASTISITY and SPARSITY.

- Plasticity is the ability to engage diverse combinations of neurons and synapses by relevance to the purpose of the analysis, and to dynamically adapt internal functional architectures.
- Sparsity is the ability to identify the minimum data required. The brain can respond to situations that are simultaneously new on multiple dimensions and can even categorize one data point.
- The neuronal and synaptic architecture of the brain is an ecosystem.

Systemic architecture, plasticity and sparsity are core to biological learning, but are NOT similar to ML algorithms. The biomimetic technologies that enable elements of RWR are:

- EXPERTESE GRAPHS and
- NEURAL SYSTEM DYNAMICS DIGITAL TWINS

We can imitate principles of plasticity and sparsity by implementing qualitative expertise graphs and leveraging them for contextual selection of data and methods from the in-memory model library.

System dynamics has been taught at MIT for some time, and digital twinning is not new. Combining the above methods delivers the evidence to enable VUCA analytics.

I will address each driver separately.

Part 3 – VOLATILITY

Volatility refers to rapid and unpredictable change. Since the risks are unpredictable, how can they be managed? There are three areas where human-led governance that is enabled by biomimetic knowledge engineering methodology and software (digital twins) can optimize our risk management process.

1. Before AI/ML models are designed
2. Before model outputs are integrated with downstream processes
3. Monitoring the results of utilizing the outputs

Before Design – Twin the Problem Definition

Premises driving the approach, and theories guiding the architecture need to be identified, modeled, and validated. One twin is required for premises and another for theories, and the modeling process helps to distinguish between expected and proven knowledge, supporting human-led governance. Although the usage of the terms varies, from a scientific viewpoint:

- Theories are preliminary conclusions based on observed evidence, but not yet tested
- Premises are assumptions in search of evidence, and frequently vague, for example: "In a lot of neuroscience, the premises remain unexamined, but everything else is impeccable. "- John Krakauer, neuroscientist at Johns Hopkins University

Similarly, impeccable algorithms that implement unexamined premises fuel unpredictable risks and are a key source of bias. Likewise, theories that are not clearly defined and mapped to existing and expected evidence are difficult to validate. This initial phase of human-led governance is most critical to mitigate significant AI risk, and it enables the next two phases.

Before Integration – Twin the Solution Methods

Usually, the final specification of the software is delivered once the coding is done. Each implemented solution needs to be modeled as an independent agent in a Solution Methods twin that can interact stochastically with the Premises and Theories twins. The three twins create an ecosystem to validate the new software by outputting various scenarios that are reviewed by relevant experts.

Monitoring Results – Test Using Problem and Solution Twins

Once the software is in Beta testing, all the outputs need to be delivered to a data lake accessible to the digital twin ecosystem. A fourth twin, Beta Discovery, explores the outputs for dark data (categories not specified in the solution schema and unexpected correlations) and interacts with the Problem and Solution twins to produce evidence of quality, correctness, and bias, enabling experts to manage volatility risk.

The biomimetic digital twin ecosystem is then leveraged and adapted to address the remaining VUCA drivers.

Part 4 – UNCERTAINTY

This year, the MIT EmTech session about knowledge and learning concluded that in the past we measured success as learning to predict the future; now we need to measure success as learning to respond to the present. Very relevant to managing AI risk.

Uncertainty risk refers to inherent limitations of the knowledge we use to address problems and goals, including assumptions, premises, and theories for which we have insufficient evidence or cannot test experimentally. As discussed in Part 1 of this series of posts, with AI we have to add the inherent uncertainty of statistical conclusions which are core to AI/ML/NLP algorithms. Einstein said, "So far as the laws of mathematics refer to reality, they are not certain. And so far as they are certain, they do not refer to reality". AI trained on rapidly proliferating junk web sites scales the problem.

The conclusion of Part 3 stated that a biomimetic digital twins ecosystem of independent agents that incorporate expert knowledge graphs can be leveraged to address VUCA risk drivers, including our ability to respond to uncertainty. How?

1. Exposing dark data in the problem domain and including it in the analytics
2. Stochastic discovery of scenarios
3. More objective exploration of response options and tradeof
4. More precise and comprehensive modeling of the problem domain

The value of the first three benefits is straightforward and clear. The fourth item is intuitively correct but may be more challenging to visualize. Consider an example: what makes food taste good?

If you are a chef, the problem domain includes ingredients, preparation, and some knowledge about your guests, including preferences and dietary constraints.

If you are a healthcare researcher working on conditions that prevent people from enjoying food, the domain could include biological, neuronal, psychological systems that produce and govern the senses of taste and smell, the communication of the senses with the brain, the interpretation of those communications by the brain, the genomics responsible for the enjoyment of food, and other domain elements that an expert in the field would define, which I am not.

The point is that when dealing with uncertainty risk, in addition to trying to minimize what we do not know and cannot measure, we need to model our problem domain as precisely and as comprehensively and we can. A biomimetic digital twins ecosystem where each agent is designed independently and incorporates multidisciplinary expert knowledge graphs can significantly improve problem analysis and dynamic adaptability to uncertainty.

Part 5 – COMPLEXITY

Part 1 addressed why VUCA is appropriate for managing AI risk. On the issue of complexity DARPA's 2005 "Get Real" project calculated the potential interactions of one million agents as 10^300,000 to stress that exhaustive modeling of any significant element of the real world is not technically feasible. Their recommendation was research into real-world reasoning methods in imitation of the human mind, which filters out the vast majority of ingested data and uses what is needed from relevant domains. Computing relevance has two key challenges:

1. Identifying relevant information domains for the analysis purpose
2. Identifying relevant data required to perform the analysis

Steven Eppinger, Professor of Innovation at MIT, used to teach a course in complexity management at the graduate school of engineering. The example used for practical application was designing a new jet engine, but in the final test he asked us to apply the principles to a pizza recipe. It helped focus on what was relevant and necessary and we used that at Pfizer to explain complexity modeling during a business transformation. Part 4 discussed how differently a chef, or a healthcare researcher would answer the question: what makes food taste good? I asked ChatGPT and got a list of attributes for the interaction between food and the senses – flavor, aroma, texture, etc. As Part 1 explained, complexity is invisible to an AI algorithm because:

1. It can only see what it has been engineered to find.
2. It works within the narrow limits of the training data.

The LLM response is a classic example of both points above. If we implement the principles discussed in Part 2, independent real-world reasoning agents that apply small/wide data methods to real-world data,

we can build expert knowledge graphs to comprehensively define the relevant domains and data. Those models can be used to test AI outputs for correctness and completeness and identify gaps that need to be addressed. Based on the output, engineers can make needed adjustments to training or explore other options to correct the issues.

Sometimes, precise definition of relevant domains and data is insufficient to validate the AI outputs because we do not fully understand activities that we can observe and measure. A frustrating example is photosynthesis. It doesn't appear complex and is the most common process on the planet, but we can't reproduce it in the lab. Achieving that could eliminate food and energy shortages worldwide. So if we have to validate a model that is too complex to understand, what can we do? Model and validate the underlying premises. The biomimetic digital twin ecosystem described in Part 3 can enable the exploration of complex systems objectively and test the conclusions stochastically.

With LLMs, however, the fourth VUCA risk driver, Ambiguity, is a serious issue.

Part 6 – AMBIGUITY

Natural language processing (NLP) is a branch of AI that is being rapidly adopted, especially in biomedicine. The hope is to include the mushrooming volume of publications in data-driven R&D. This technology is subject to risk, especially driven by ambiguity, but there are mitigation strategies and tools.

NLP AMBIGUITY SOURCES: the nature of human language and algorithmic bias.

The nature of language – Chomsky's generative linguistics work in the 1950s was seminal to modern generative AI, but in 2011 he stood before a packed auditorium at a CogSci Society conference and asked: "What is human language?" He stated that no theory of the mind to date explains what human language is or how it came about and affirmed that he will not look at another paper on the subject unless it accounts for sign languages used by the deaf around the world, which have visual and spatial grammar, very different from spoken language. No one challenged his view that the gap between human language and lower primates is so huge and complex that it proves that we all descended from one individual. His reasoning: "That it happened in one individual is statistically inexplicable, that it happened in two is inconceivable." In 2022 clues began to appear about the genetic differences between humans and other species in what used to be viewed as junk DNA (search HAQER).

Since we cannot explain fundamentally what human language is, we cannot write software that understands it. This drives the risk of interpretation errors. In an interview with MIT Technology Review (7/5/23) Eric Schmidt, former Google CEO said, "we should be cognizant of the limitations—and even hallucinations—of current LLMs before we offload much of our paperwork, research, and analysis to them.

Algorithmic bias – LLMs are biased in three ways:

1. The designers' understanding of the problem space limits what the algorithm classifies as relevant to the search
2. The volume of poor-quality content on the internet that has trained LLMs, and accounts for many of the "hallucinations"
3. Differentiated meaning of terms across scientific and technical domains that impede precise classification

NLP AMBIGUITY STRATEGY: Human-led governance of AI risks and bias

The FDA has begun addressing this issue with emerging requirements and the introductory FDA article with links to the discussion papers is here:

https://www.fda.gov/news-events/fda-voices/fda-releases-two-discussion-papers-spur-conversation-about-artificial-intelligence-and-machine

NLP AMBIGUITY TOOLS: Biomimetic Digital Twins

To address these issues and to provide guidance to the biomedical community, the National Academies of Sciences, Engineering and Medicine (NAS) sponsored by the National Institutes of Health (NIH), the National Science Foundation (NSF) and the Department of Energy (DOE), began advocating research into the use of biomedical digital twins technology to more effectively model multidimensional and multi-scale biological complexity (https://www.nationalacademies.org/event/01-30-2023/opportunities-and-challenges-for-digital-twins-in-biomedical-sciences-a-workshop).

Conclusion

Implementing the methodology described here has two critical requirements:

1. Early adopter mindset
2. Network of innovative partners to evolve the ecosystem

We are committed to it. If you want to explore a potential role for your organization, please contact the author at jglick@ryailiti.com.