Ontologies, Neuro-Symbolic and Generative AI Technologies
Toward Trustworthy AI Systems

Kenneth Baclawski[1], Michael Bennett[2], Gary Berg-Cross[3],
Todd Schneider[4], Ravi Sharma[5], Mark Underwood[6], Andrea Westerinen[7]

[1]Northeastern University, [2]Hypercube Limited,
[3]ESIP Semantic Harmonization, [4]Engineering Semantics,
[5]Senior Enterprise Architect, [6]Krypton Brothers, [7]OntoInsights

Abstract

This article is the Communiqué of the Ontology Summit 2024 which dealt with how ontologies and knowledge graphs can both aid and benefit from neuro-symbolic and generative AI technologies. The article synthesizes and summarizes the main points presented and discussed during the summit. Neuro-symbolic systems integrate neural networks with systems based on forms of logic applied to human-readable symbolic representations. The advantages and disadvantages of current neural network technologies and symbolic technologies are listed and found to be mostly complementary, thereby motivating the development of neuro-symbolic systems. Representative examples of applications are presented that combine semantic technologies with neuro-symbolic and generative AI technologies, which could help improve trustworthiness of AI systems. The risks and ethics associated with these systems are also presented.

Introduction

ARTIFICIAL INTELLIGENCE (AI) SYSTEMS are not new but have been invigorated by the subset of AI called machine learning (ML) using neural net architectures, which afford many new applications for tasks like image classifications and text generation. Part of the current excitement in current AI is the possibility of effective neuro-symbolic (NeSy) systems which combine current ML systems with symbolic technologies based on human-readable representations of problems, logic and search. NeSy systems have emerged as a promising approach for dealing with the recognized deficiencies of non-symbolic ML techniques that are neural network based.

The Ontology Summit 2024 was a series of sessions held from October 2023 to May 2024 and featured over 20 prominent researchers in the field of NeSy systems (Ontology Summit, 2024). The summit surveyed current techniques that combine neural network machine learning with symbolic methods, especially methods based on ontologies and knowledge graphs. At the simplest level ontologies capture some degree of meaning and are representations of a knowledge domain. They define the concepts, relationships, properties, axioms and rules within that domain, providing a framework that enables a deep understanding of that subject area. A knowledge graph (KG) is a representation of a set of statements in the form of a node- and edge-labeled directed multigraph allowing multiple, heterogeneous edges for the same nodes (Ontology Summit, 2020). As discussed in detail as part of the Ontology 2020 Summit, a KG can be based in whole or in part on an ontology (Ontology Summit 2020). An ontology enables machine reasoning by allowing a system to draw inferences and to derive new information and relationships between entities both in the ontology and in any KGs that are based on the ontology. In contrast neural network and other ML models acquire some degree of shallow knowledge by training on large corpora, learning the patterns and connections between words and images (Bzdok, Altman, and Krzywinski, 2018). Hence, although their "knowledge base" is broad, it is also sometimes incorrect and/or biased, and doesn't explicitly understand the semantics or relationships in that content.

This article is the Communiqué that synthesizes and summarizes the major points of the

Ontology Summit 2024. We begin by surveying both kinds of AI methods: deductive methods, such as ontologies and KGs, and inductive methods, such as neural networks. Deductive methods are also known as symbolic or logic-based AI, and inductive methods are also known as non-symbolic or subsymbolic AI. It was found that the two kinds of methods have complementary strengths and weaknesses. This motivates finding ways to make use of both methods in applications. Such combined approaches range from ad hoc uses to fully integrated NeSy systems. We give some examples of combining the technologies in the Case Studies section. We then classify NeSy systems, after which we discuss some applications of such systems. As with other automation opportunities, the introduction of AI technologies comes with risks, and we discuss these as well as some of the ethical issues and regulations of such systems.

Symbolic and Subsymbolic Methods

Symbolic methods are techniques based on human-readable representations of problems that are processed using various forms of logic, including classical logic, non-monotonic logic, and probabilistic logic. Common tools for symbolic methods include logic programming, production rules, semantic nets, and frames. Symbolic methods have a long history going back to the development of modern computers in the 1950s, and significant research and development continues to the present day. Significant recent work on symbolic methods include ontologies and knowledge graphs.

Some of the many applications of symbolic methods include knowledge-based systems, symbolic mathematics, automated theorem provers, ontologies, the semantic web, automated planning and scheduling systems, symbolic programming languages, and multi-agent systems. One of the applications of symbolic methods is to AI, and they were the dominant paradigm of AI from the mid-1950s to the mid-1990s. However, AI is only one of many applications of symbolic methods.

In contrast with symbolic methods, subsymbolic methods, also known as connectionist methods, are techniques that do not rely on formally specified symbols and logic. Subsymbolic methods, both generative and discriminant (for finding patterns in data), are primarily based on "deep" neural networks and include large language models (LLMs) and many other methods. The currently popular generative AI systems are most commonly implemented using subsymbolic methods. One of these is the ML-based attention method which evaluates inputs to determine their importance (weight), simulating how human attention works. Attention methods are critical in generative pre-trained transformer (GPT) architectures such as LLMs that are currently very popular for natural language processing (NLP). Instead of a model specified with rules and formulas, subsymbolic methods learn their models from a dataset called the training set. A subsymbolic system that has been trained is validated by applying its model to a validation dataset. Once the subsymbolic system has been trained and validated it can be applied to new data to generate outputs. Contemporary subsymbolic systems are trained on very large training sets although the foundational details are not usually publicly released.

Unlike subsymbolic systems, symbolic methods are transparent making them easier to debug and control, and therefore, explainable, reliable and trustworthy. The advantages of symbolic methods are important for many applications, such as mission- and life-critical applications. However, symbolic methods can struggle with real-world complexities and the need for hand-crafting knowledge into a processable form, which limit scalability. Still, when an application has a relatively small domain, the lack of scalability of symbolic methods is not an issue. All the same, when scalability is important, subsymbolic methods have significant advantages.

The AI Landscape

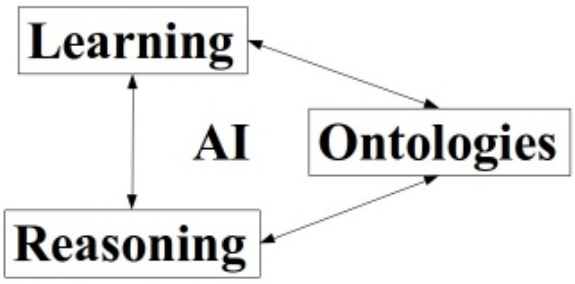

Figure 1: The three components of AI (Ontology Summit 2017)

Today we have a dynamic and noisy AI landscape with reports such as "large language models are changing AI and we need to understand them" and "The Oppenheimer moment in AI: How Generative AI (Gen-AI) and LLMs are changing the world" (McGuinness, 2023). This landscape supports new approaches to a fuller, integrated artificial intelligence (AI) that combine the strengths of past established symbolic approaches to cognition and new approaches. Broadly, this combines learning with reasoning and knowledge in symbolic forms like ontologies. Figure 1, from the Ontology Summit 2017, is a high level model depicting the three AI components.

Cognitive theory and research helps us understand differences in components like learning and reasoning and how the new field of integrated AI fits into the well known dual processing modes of evolved human cognition systems (Kelly & Barron, 2022). These two modes are called System 1 Thinking, which is fast, automatic, frequent, emotional, stereotypic, unconscious, opaque; and System 2 Thinking, which is slow, effortful, infrequent, logical, calculating, conscious, explainable (Kahneman, 2011). An important consequence of this theory of human cognition is that the two Systems can arrive at different results although given the same inputs. The theory has provided possible explanations for many cognitive biases, and investigations into the basis of cognitive biases is an ongoing and active area of research.

The distinction between subsymbolic and symbolic is analogous to the distinction between System 1 thinking and System 2 thinking, respectively. In some respects this analogy is helpful, but the analogy can break down. For example, using heuristics, Kahneman (2021) proposed that System 1 thinking involves associating new information with existing stereotypical patterns, or thoughts, rather than constructing new patterns for each new experience. This is similar to subsymbolic systems, which are based on large databases of patterns derived from the training data. However, while System 1 thinking is fast and System 2 thinking is slow, the reverse is the case for subsymbolic and symbolic systems; computers perform symbolic operations, especially deductive logic, extremely fast, but subsymbolic systems have become so large, using billions or even trillions of parameters, that they require large computing resources and as a result are relatively slow. The other features that distinguish System 1 and System 2 are mainly human characteristics and so attributing them to computer systems is problematic. Nevertheless, if one is careful to avoid cognitive biases, such as jumping to conclusions, the analogy can be useful. One distinction that does appear to be accurate is that System 1 thinking is opaque while System 2 thinking is explainable. This is largely also a distinction between subsymbolic and symbolic systems.

Another aspect of some symbolic methods is the generation of a consensus among the individuals in a community, generally due to the way that they are created. This is especially true for most ontologies, but it is also part of other symbolic methods. For example, in software engineering the creation of requirements requires a consensus among the stakeholders (often expressed using an ontology), but the later phases of software development do not have as much emphasis on

consensus. Subsymbolic systems such as LLMs are not designed for generating consensus. When the training set of a model includes the result of a prior consensus-building process, then it might appear that the subsymbolic system is generating a consensus, but the system is only retrieving an existing one. Furthermore, even if a subsymbolic system could propose a consensus when trained on a set of varying opinions, the system would still need to explain the proposal if the system would be able to convince the community to agree to it.

There was considerable research and development on symbolic systems in the early days of AI. Now, due in part to advances in computer chips and big data as well as improved neural algorithms, there is rapid progress with producing subsymbolic systems. For the rest of this section we begin with an overview of some of the many subsymbolic architectures and systems that have been developed. We then discuss the advantages and disadvantages of subsymbolic and symbolic techniques. For the most part the disadvantages of one class of techniques are advantages of the other class of techniques. We first survey the advantages of subsymbolic techniques and then discuss their limitations.

The Many Kinds of Subsymbolic Systems

The current AI era advances machine learning (ML) based on massive data repositories and fast chips for processing. Important capabilities include image recognition and language translation. These machine learning advances make use of neuro-inspired architectures. Text, images, and other data fuel ML that purportedly mimics the brain's layered neural networks. Current subsymbolic systems can have hundreds of connected layers. For instance, recognizing objects in images involves pre-training on millions of labeled images (the "P" in GPT) – cats, dogs, houses, cars – under varying conditions, which allow for general image recognition, avoiding the brittleness of early systems. Using a neural network to learn to recognize images, all neatly labeled, requires powerful computing with examples under different lighting and angle conditions as well as with different backgrounds and realistic partial views of the entities due to occluding objects. In a complex scene the system is trained using segments of the image, so it can identify a baseball player, a bat and a ball all in the same scene. Taken together, powerful computers with massive amounts of data allow some generality of image recognition.

Convolutional Neural Networks (CNNs) are an important aspect of some neural-based machine learning. These are networks that are inspired by the human visual cortex, extracting features and patterns from images in a manner that bears some similarity to how the brain identifies complex shapes and objects. Initial layers extract low-level image features like edges, corners, and textures, while layers deeper in the network take these features as outputs and combine them to identify more complex shapes and objects. Non-linear activation functions between layers allows networks to learn complex relationships between features.

For text processing, a large language model (LLM) is trained on a massive dataset of text, typically scraped from information on the Internet. This process helps achieve general-purpose language generation and avoids much of the brittleness of earlier systems for language generation.

The "T" in GPT stands for "Transformer." Neural network architectures have been designed for natural language processing tasks like translation and text generation. Transformer-based AI systems are a type of architecture that can process sequential data, such as text or speech. They can learn probabilistic relationships between different parts of a symbolically represented sequence to predict the next word or phrase (often referred to as a token). Multi-dimensional arrays use a tensor representation to capture the probability of how words are connected sequentially in text. The LLM and Chatbot systems learn to process these tensors in a way that allows it to process

questions and generate responses that are understandable to humans. Transformer-based chatbots have captured popular attention by engaging in dialog in response to questions. Transformers power many machine translation systems, allowing for more accurate and natural-sounding translations than earlier symbolic attempts. An essential idea of this architecture is that unlike traditional neural networks that process data sequentially, transformers use a technique called self-attention. This allows them to analyze many parts of the sequence simultaneously, identifying the likelihood of different elements being part of a sequence.

A GPT learns in a trial and error fashion using reinforcement learning. In the process it adjusts internal parameters (i.e., the weights of the connections between nodes of the network) to optimize future choices. The network iteratively learns from mistakes, refining its abilities with each training batch. This aspect is one that relies on powerful computing to provide many iterations. As a result, training a GPT model can be very expensive, e.g., for computer hardware and operational costs such as electrical power. While companies do not publicly release details about their models, some information has "leaked out." For example, there is some evidence that GPT-4 has about 1.8 trillion parameters across 120 layers and cost about $63 million to train.

There are several additional building blocks to a transformer system, including:

- Tokens: The input data (text, speech) is broken down into smaller units called tokens. These could be individual words, phrases, or even characters.
- Word Embeddings: Each token, say a word, is converted into a numerical representation as a vector in a multi-dimensional space, establishing a context with other tokens.
- Encoders and Decoders (for generative tasks): Transformers often have encoder and decoder parts. Encoders process the input sequence, analyzing the relationships between tokens. Decoders use the encoder's output to generate an output sequence, like translating a sentence or writing a summary.
- Self-Attention Mechanism: This is a key to an effective transformer. An attention mechanism allows the model to focus on specific parts of an input sequence and determine how they relate to other parts. It is strategically like highlighting relevant parts in a sequence to capture the overall relations.

Advantages of Subsymbolic Systems

We now examine the advantages of subsymbolic systems:

- Accuracy: By considering relationships between parts of massive input data, generative AI based on transformers architectures can achieve high human-level accuracy when tested on standard tasks like machine translation, text summarization, and question answering.
- Parallelization: Transformers show an ability to analyze many parts of a sequence in parallel, making them more efficient in processing large amounts of training data.
- Flexibility: A transformer architecture can be adapted to various tasks by changing the nature of the inputs and outputs, e.g., images rather than text. This makes them a more general tool useful for several AI applications.
- Statistical Learning (of a rich conceptual model): Subsymbolic systems excel at finding patterns and relationships within massive amounts of data. This ability enables them to model the world in a way that appears to mimic human general intelligence.
- Output Capability: Transformers can translate between languages and their notations including formal/logical ones. They can also effectively summarize and provide alternative outputs.

- Emergent Complexity: As subsymbolic systems become increasingly complex, with more parameters, more neural net layers, and more data processing capabilities, new and unforeseen abilities might emerge. To some this theoretical "emergent complexity" could lead to a system that exhibits human-like general intelligence, even if it wasn't explicitly programmed for it. Evidence for emergence is suggested by Rapid Performance Improvements. That is as subsymbolic systems grow in size and complexity, they can exhibit sudden and significant improvements on tasks such as simple reasoning, answering open ended questions comprehensively, and even code generation.

Research suggests that these emergent abilities might exhibit "phase transitions," critical points in the size of the system and/or the amount of training data beyond which new capabilities emerge out of nowhere. Finally there is the phenomena of few-shot learning. Some subsymbolic systems show the ability to learn new tasks with minimal training data, suggesting an underlying generalizability by creatively combining the knowledge going beyond rote memorization (Wei et al., 2022).

On the other hand, there are alternative explanations for emergent subsymbolic capabilities. One such alternative explanation is that for a particular task and model family, when fixed model outputs are analyzed, emergent abilities appear due the researcher's choice of metric rather than due to fundamental changes in model behavior with scale (Schaeffer et al., 2024). Specifically, "nonlinear or discontinuous metrics produce apparent emergent abilities, whereas linear or continuous metrics produce smooth, continuous, predictable changes in model performance.... alleged emergent abilities evaporate with different metrics or with better statistics, and may not be a fundamental property of scaling AI models" (Schaeffer et al., 2024).

Limitations of Subsymbolic Approaches

We now discuss some of the disadvantages of subsymbolic techniques that were identified during the Ontology Summit 2024. Gary Marcus (2020, 2024) thinks that subsymbolic systems do not provide a proper foundations for trustworthy artificial general AI for several reasons. Marcus argues that what is needed is a more "hybrid, knowledge-driven, reasoning-based approach, centered around cognitive models, that could provide the substrate for a richer, more robust AI than is currently possible" (Marcus, 2020).

Some of the key disadvantages of current subsymbolic systems include:

1. Rich cognitive models are needed that describe mental processes in detail to keep track of dynamic environments.
2. Extensive real world knowledge/experience is needed rather than just text-based knowledge.
3. Intelligent systems need to represent complex relationships between entities, such as causal relationships. These require a deep understanding of the world, and not just a recognition of patterns.
4. Representation of wholes and parts, i.e. composability, is necessary.
5. Common sense knowledge is developed over time through embodied experience and the ability to interact with the environment.
6. Sophisticated reasoning explicitly uses symbols, logic, and rules and require a symbolic foundation.
7. Some useful knowledge of human sentiment and preferences are important for tasks such as medical decision-making.
8. System behavior must be explainable and not a black box output.

A similar list of issues for subsymbolic systems was developed by John Sowa (2024). These include:

1. No fixed set of meanings can adequately describe a continuous, dynamically changing world (which we might note is also a limit on a formal ontology).
2. Written language, the source for training LLMs, is isolated from perception, feelings, actions, and reactions of people in a dynamically changing world.
3. Useful mental models are needed and are more fundamental than language or logic.
4. Much of human intelligence and underlying mental models are probably lost in a mapping to LLMs.
5. A linear language or notation is not ideal for thinking or communicating complex spatial patterns.

Sheth (2024) argues that subsymbolic systems, like GPT-4, have impressive pattern recognition capabilities, including language processing; but generating coherent text based on input has serious reasoning limitations. Subsymbolic systems can perform certain types of reasoning tasks, such as simple logical deductions or basic arithmetic, and can produce responses that appear rational on the surface. However, they lack genuine comprehension or logical consistency since their reasoning capabilities are limited without real, cognitive understanding or awareness of concepts, contexts, or causal relationships. These systems cannot go beyond the statistical patterns in the data on which they were trained. Thus, as also argued by Marcus and Sowa, subsymbolic systems often need help with more complex forms of reasoning that require deeper understanding, context awareness, or commonsense knowledge. Furthermore, their reasoning does not adapt well to the dynamicity of the environment, i.e., the changing environment in which the AI model is operating (e.g., changing data and knowledge). More of the implications of these limitations is discussed in the Risk section and consideration of future developments are discussed in the Summary section.

Case Studies

In this section we give some examples of systems that make use of both symbolic and subsymbolic methods, including a system that predates current ML systems. We first discuss some examples of systems that use subsymbolic techniques for creating and maintaining KGs and ontologies. Then we discuss examples of systems that use KGs and ontologies to improve subsymbolic systems.

From Subsymbolic to Symbolic

OntoLearn is an example of a knowledge extraction and ontology development tool which predated current ML systems and aids in the creation and population of domain ontologies (Missikoff et al., 2002). The tool utilizes a three-phase approach:

1. Extraction and filtering of terms from domain documents (using natural language processing and statistical techniques),
2. Determination of the terms' underlying semantics and concepts, along with the assignment of concept identifiers (using knowledge bases such as WordNet (Fellbaum, 1998)), and
3. Generation of taxonomic, similarity and other relationships linking the terms.

These steps result in the construction of a domain concept forest, ultimately resulting in an ontology.

The recent release of LLMs makes many of OntoLearn's natural language and semantic concept matching tasks easier. LLMs have captured the imagination of the public and researchers

alike. In contrast to previous generations of machine learning and statistical models, LLMs are more general-purpose tools, which can communicate with humans. They excel at many natural language processing (NLP) tasks such as information extraction and summarization, translation, classification and more (Bennett & Westerinen, 2023).

LLMs can define terms, find relevant, published resources, and answer factual questions based on their internally represented knowledge. So a question is, can they automatically extract and structure something useful given what they can process from texts? Can they elevate their output into richer forms such as are found in an ontology? Certainly, at a minimum, LLMs can aid and enhance several aspects of ontology design and KG population. Several presentations from the Fall Pre-Summit series demonstrated these capabilities, as well as the value of using ontologies and KGs to validate and support the creation of trustworthy output by LLMs.

The following bullets summarize the value of coupling ontologies and LLMs (Bennett & Westerinen, 2023):

- Use of LLMs to aid in ontology creation and population:
  - Extraction of information from text and mapping to an ontology
  - Creation of lists of initial concepts to begin or extend an ontology, or to create exemplary KG instances
  - Generation of SPARQL queries from natural language
  - Summarization of an ontology or KG
  - Assistance in alignment of ontologies and KGs
  - Generation of competency questions and use cases for an ontology
- Use of ontologies to aid in the output of trustworthy information by LLMs:
  - Validation of the responses of an LLMs
  - Creation of training data or prompt inputs for an LLM
  - Support logical reasoning and inference

As an empirical test of the value of LLMs, Giglou et al. (2023) performed a comprehensive evaluation of nine different LLM model families (Bert, Bloom, Llama, GPT, etc.) using a zero-shot prompting method. Three learning tasks were evaluated:

1. Taxonomy discovery (studying whether modular structure can be determined)
2. Term typing
3. Extraction of non-taxonomic relations

It was found empirically that LLMs are not yet suitable for ontology construction that entails a high degree of reasoning skills and domain expertise, such as found in biomedical and food ontologies. On the other hand, it seems reasonable that: "... when effectively fine-tuned they [LLMs] just might work as suitable assistants, alleviating the knowledge acquisition bottleneck, for ontology construction" (Giglou et al., 2023).

Another example where generative AI using LLMs have been used to improve ontologies is the vast area of the Ontology of Chemical Entities of Biological Interest (ChEBI). A challenge for such large ontologies is how they can be maintained as the number of elements (chemical entities in this case) expands rapidly and as novel structures are developed. The complexity of such ontologies makes their development and maintenance especially challenging. Entities can have many labels and many parent entities, and the parent hierarchies can be very unbalanced with some parts being sparse while others are very dense. Even a rule-based extension of a chemical ontology based on structural features of the molecules using a tool such as ClassyFire (Djoumbou

et al., 2016) is hard to maintain or integrate with ontologies. Thus ChemOnt's 4,825 classes and rules are not integrated with ChEBI's definitions of ontology classes (Mossakowski, 2024). For such ontologies, the classical ML approaches have generally not been able to learn the ontology hierarchy, and automated text mining has failed to replace labor intensive hand curation.

Hastings et al. (2021) have attempted to use machine learning to extend the ChEBI ontology to represent a novel molecule by starting with the knowledge regarding the classes to which the molecule belongs. This knowledge helps to predict the molecule's chemical behavior and uses, as well as to enrich data and to drive discovery approaches. Subsymbolic techniques for ontology extension such as recurrent neural networks, scale better than previous techniques as the size the molecule and the ontology increases, but they struggle with complex structures like aromatic ring structures.

ChEBI classification techniques can be compared with other techniques in terms of speed, reliability, and adaptability to noisy data. Results show that RoBERTa and ELECTRA transformers with self-attention, pre-training, fine-tuning and utilization of sub-sampling for generating learning-ready datasets out-compete other approaches (Mossakowski, 2024).

On the whole, early empirical results show that LLMs are useful for simple ontology construction. However, they are not sufficient (without retraining or supplemented knowledge) for tasks requiring domain expertise or when advanced reasoning is needed.

Continued research might lead to some changes to the knowledge engineering process a well as address a family of research questions such as:

- What are useful neuro-symbolic architectures for (grounded) knowledge representation?
- Can an LLM act like willing but mediocre domain experts to get some initial views of a domain?

Neuhaus (2024) argues that subsymbolic techniques are not likely to replace ontologies or to fully automate the knowledge engineering process. However, in partnership with knowledge engineers and domain experts and as an assistant, they are likely to provide very valuable tools for knowledge building.

From Symbolic to Subsymbolic

We now give a brief discussion of the use of symbolic techniques to improve the performance of subsymbolic systems. Ontologies can provide high-quality symbolic knowledge that captures a consensus among members of a community as discussed in the AI Landscape section above, where it was noted that subsymbolic systems are not well suited for generating such a consensus. This suggests that ontologies could help subsymbolic systems overcome this limitation. This is illustrated by the approach to integrate knowledge from ontologies into the structure of a neural network called ontology pre-training (Glauer et al., 2022, 2023). Pre-training allows the network to predict membership in ontology classes so that the structure of the ontology becomes embedded into the network. A subsequent training process prepares the system for a particular prediction task. Glauer et al. (2023) used this approach to predict potential toxicity for small molecules based on molecular structure. This, along with many other tasks in life sciences chemistry, is normally a challenging task for ML. Their approach improves on the state of the art, and shows that the model learns to focus attention on more meaningful chemical groups when making predictions with ontology pre-training than without. This may provide a path towards greater robustness and interpretability of such tasks. In addition, the training time was reduced after ontology pre-training, indicating that the toxicity space model was better structured to learn what

matters for toxicity prediction with ontology pre-training than without. This is one example of a neuro-symbolic approach to embed meaningful semantics into neural networks.

Retrieval-Augmented Generation (RAG) is a NeSy method for dealing with the disadvantages of transformer-based systems. A RAG system combines both symbolic and subsymbolic AI techniques to make information more reliable by building on a base of trusted information. A RAG eliminates or at least mitigates mistakes by using a set of documents as the only oracle for truth. Techniques like filtering retrieved documents based on specific criteria or adjusting the number of documents retrieved (k most similar) can help ensure that retrieved information directly addresses a question in a more focused way. For KGs, the filtering can be specified using SPARQL queries to retrieve relevant documents and other objects dependent on the domain. If a RAG system cannot determine an answer from available documents, it can report this – basically stating that "it doesn't know". In addition, RAG can avoid the issue of black box reasoning because it can be used to return the specific documents/texts that support an answer. For example, AllegroGraph integrates with ChatGPT and utilizes RAG, processing natural language queries based only on the statements in a knowledge graph (AllegroGraph, n.d.). RAG is valuable to aid in calculating confidence in results or to state that information is not available. This benefits users to understand limitations and avoid misunderstandings. For a more in-depth discussion of RAGs, see Kurt Cagle's talk on complementary thinking (Cagle, 2023) and the article by DeBellis and Underwood in this same issue (DeBellis & Underwood, 2024).

Neuro-Symbolic Systems

NeSy themes have a long history, but only recently have substantial projects been developed. The integration of symbolic and subsymbolic methods remains a challenge, but integration appears to be useful for addressing complex AI problems that cannot be solved by purely symbolic or subsymbolic means alone. So there are benefits to integrating the techniques of both paradigms. Coupling may be through different methods, including the calling of deep learning systems within a symbolic algorithm, or the acquisition of symbolic rules during training. Ideally an AI system should include a sound, symbolic reasoning layer in combination with deep learning. In this section we classify the existing and proposed techniques for integrating symbolic and subsymbolic systems.

NeSy architectures were classified by Kautz at the Ontology Summit 2021 and summarized in the Communiqué (Kautz, 2021, 2022; Baclawski et al., 2022, §4). Kautz's taxonomy consists of the following six types of systems:

- A Type 1 neural-symbolic integration is standard deep learning. This is included by Kautz to note that the input and output of a neural network can be made of symbols, e.g. text in the case of language translation or question answering applications. Other types use different inputs.
- Type 2 are systems such as DeepMind's AlphaGo and other systems where the core neural network is loosely-coupled with a symbolic problem solver such as Monte Carlo tree search.
- Type 3 involves a neural network focusing on one task (e.g., object detection) and interacts via its input and output with a symbolic system which may specialize in a complementary task (e.g., query answering). Examples include the neuro-symbolic concept learner (Mao et al., 2019) and deepProbLog (Manhaeve et al., 2018).
- Type 4 neuro-symbolic systems compile symbolic knowledge into the training set of a neural network as was discussed for the toxicity case study.
- Type 5 systems are tightly-coupled but are distributed systems in which symbolic logic rules

are mapped onto embeddings which act as soft-constraints on the network's loss function used for learning. An example of this is the Logic Tensor Network (Schlag & Schmidhuber, 2018).

- Finally, Kautz's Type 6 system is a fully integrated system capable of true symbolic reasoning inside a neural engine.

Some of the main benefits of NeSy systems include:

- Transparency: Combining symbolic representations with neural networks can make AI decisions more understandable and explainable.
- Flexibility: Symbolic reasoning allows for adapting to new situations and reasoning beyond data patterns, offering greater generalizability.
- Efficiency: Efficiently leveraging the strengths of both approaches can lead to faster learning and improved performance.

Applications

There are several domains where the disadvantages of subsymbolic systems are especially problematic, including healthcare, military applications and control systems (e.g., nuclear reactors). In this section we discuss healthcare applications and then discuss how one can integrate domain knowledge by combining ontologies, KGs and subsymbolic methods.

Healthcare

Ophthalmology is one area where healthcare specialists have utilized ChatGPT to create summaries and notes for treatments. An additional use of subsymbolic systems is to make complex medical statistics more comprehensible through condensation (Thirunavukarasu, et al., 2023). When researchers need to explain complex medical data or statistical results, subsymbolic systems can provide clear and concise explanations, making them easier for readers to understand (Meng et al., 2024).

There is a range of healthcare assistance (along with challenges) that are driven by NeSy AI. One example is to use neural network-based processing for depression assessment, to highlight the symptoms and definitions of the disease along with contextualized recommendations, and to provide decision support information in cases of depression. In this application, text related to depressive symptoms, such as feelings of tiredness, changes in appetite, and emotional fluctuations are run through a neural network to predict and explain the individual's feelings and thoughts related to depression. It can provide predictive analysis for depression assessment, indicating the likelihood of depression based on input text with a high degree of confidence. It can also propose useful assessments such as the disease severity, which assists in recommending further assessment, appropriate interventions or patient treatments such as suggesting pharmacotherapy as a potential treatment option. Virtual healthcare assistance incorporates human-like communication and understanding with intuitive interfaces to help bridge between users and complex AI systems. It recognizes different types of depression and the specific circumstances of each (Roy, 2024).

ALLEVIATE is a second generation AI-enabled virtual assistant used for telehealth. It focuses on mental health cases. ALLEVIATE provides personalized, user-explainable views for both patient and clinician users during interactions, incorporating safety-constrained operations and user-level explanations for outcomes. It offers dynamic evaluator feedback-based refinements and a domain knowledge-guided safety envelope for medically safe patient interactions and responsible emergency response (Roy, 2024).

Pharmaceutical applications are another important class of healthcare applications since this is a big part of treatment. While a great amount of drug design data is available, an important goal is to acquire in-depth knowledge of the chemical properties that determine whether a candidate drug will be safe and effective. This knowledge would aid design, control, optimization and safety. Steps towards this goal include development of a series of related ontologies including:

- An equipment ontology (such as STEP, AP231, FIATECH),
- A general recipe ontology (such as OntoCAPE),
- A process safety ontology covering deviation, cause, consequence etc.,
- A material ontology for pharmaceutical product development,
- A reaction mechanism ontology about atoms, molecules, bonds, etc. (Steinbeck, 2006), and
- A model and guideline ontology (Venkatasubramanian, 2024).

Ontologies preserve drug domain semantics, offer efficient knowledge representation, and help organize information hierarchically in the form of class-subclass relationships. In addition, ontologies capture relationships and instances that can provide the basis for ontology-based NeSy systems. The Columbia Ontology for Pharmaceutical Engineering (COPE) includes a data driven module to support information extraction and semantic search coupled with a user interface employing a Chat system (Mann, 2023). COPE provides weak supervision to programmatically annotate important words in text documents, particularly in the context of custom-built drug development ontologies and health and biomedical ontologies. It can also generate labeled datasets using text overlap, with each word being labeled as important or not important. Additionally, COPE highlights the focus on developing therapeutic agents that elevate levels of high-density lipoprotein cholesterol (HDL-C) and the development of cholesteryl ester transfer protein (CETP) inhibitors.

Information extracted from unstructured pharmaceutical documents could identify important entities, along with context and relations between the entities, using custom-built drug development ontologies combined with standard domain ontologies. The identified entities and associated context and relations could help auto-generate KGs. The generated KGs could then be used in downstream applications such as visual representation, text summarization, and efficient search. COPE has used this approach to support current work on CETP inhibitors such as evacetrapib (Venkatasubramanian, 2024).

Integrating Domain Knowledge

Ontologies and KGs are useful tools to connect disparate, but potentially related, areas of knowledge. Recent advances in AI can aid in rigorously exploring relationships that cut across distinct areas, such as mechanics, biology, general science, or even art (Buehler, 2024). This exploration capability can deepen our understanding and accelerate innovation in these areas (Buehler, 2024). As an example, information, starting with a set of 1,000 scientific papers in the field of biological materials, was represented as a detailed ontological KG. An analysis of the graph structure resulted in the conclusion that there was no characteristic size to the node structure of the KG. In other words, the KG had "an inherently scale-free nature" (Buehler, 2024).

Integration of similar knowledge into the KG was accomplished by use of a large language embedding model to compute deep node representations. A path sampling strategy using combinatorial node similarity ranking allowed linking of what had been assumed dissimilar/unrelated concepts across the KG. Figure 2 shows an example of a distant relation between a flower and nacre-inspired cement via a node-path in the KG.

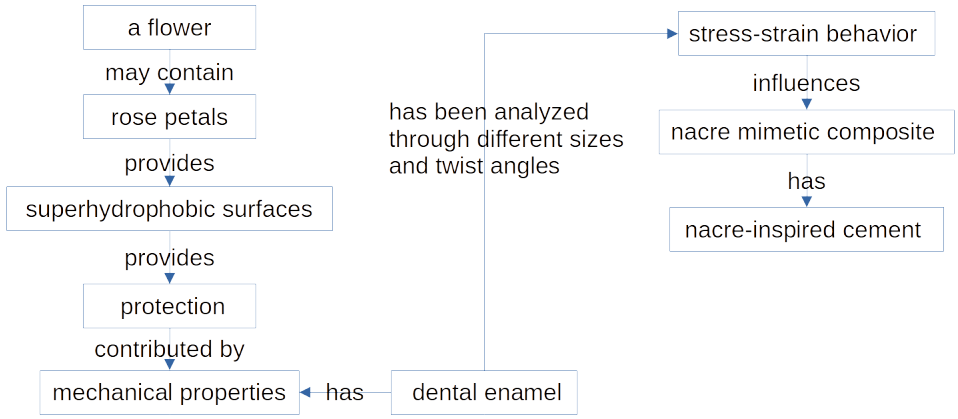

Figure 2: A path in a knowledge graph (Buehler, 2024)

The result of the integration of similar knowledge into the KG was to allow queries that reveal unprecedented interdisciplinary relationships and insights which identified gaps in knowledge about material designs, and to predict material behaviors. One comparison revealed detailed structural parallels between biological materials and Beethoven's 9th Symphony. This highlights the value of isomorphic mapping of complex, shared patterns.

In a second example, the approach "proposed" an innovative hierarchical mycelium-based composite based on a joint synthesis of path sampling with principles extracted from Kandinsky's "Composition VII" painting. The resulting composite was described as a "balance of chaos and order, adjustable porosity, mechanical strength, and complex patterned chemical functionalization" (Buehler, 2024). Other structural parallels were found across science, technology and art, that suggest dynamic, context-dependent heterarchical interplay of entities beyond traditional hierarchical paradigms.

Taken as a whole the use of graph similarity methods, illustrated by the examples mentioned above, represents a potentially useful framework for "innovation, drawing from diverse fields such as materials science, logic, art, and music, by revealing hidden connections that facilitate discovery" (Buehler, 2024). Of course, a suggested connection may be a spurious relationship due to a coincidence or a confounding factor (Burns, 1997). We discuss this issue and other risks in the next section.

Risks

Current AI technology seems to promise a wave of benefits ranging from task efficiency and automation, better data and information analysis with resulting insights and enhanced social well being and progress. Like previous automation opportunities it also comes with risks ranging from job dislocations, algorithmic discrimination, data bias, and lack of accountability. Our discussion of risks is broadly divided into several parts starting with general risks that may apply to software-based systems and perhaps would be enhanced by them having an artificially intelligent type of knowledge, reasoning and learning. We then consider more specific risks associated with current AI systems. There have been attempts to mitigate and manage AI risks by means of regulations and guidelines, and we mention a few of these. Finally, we consider how NeSy systems might help to mitigate and manage risks.

There are many notions of risk that depend on the domain as well as the particular context. A taxonomy of risk notions (or more precisely measures of risk) was presented at the Ontology

Summit 2022. While the risk taxonomy was discussed in the context of disasters, it was intended for measuring the level of risk for both adverse and beneficial events in contexts other than disasters such as information technology, project development, health and safety (Baclawski, 2022). The following are the main classes in the risk taxonomy:

1. Probability of the event. For this measure one wishes to minimize the probability of adverse consequences and to maximize the probability of beneficial consequences. This measure is often used in the context of an information technology project.
2. Downside Probability. This measure focuses on the probability of adverse consequences. Many contexts make use of this approach to risk, such as environmental risks, individual health risks, as well as occupational, safety and security risks.
3. Expected consequences. The expectation is the product of the probability of the event times the impact of the event, typically expressed in financial terms. The expectation is frequently used in business contexts as well as in other contexts such as the ones listed above when the impact can be quantified.
4. Variance. Unlike probabilities and expectations, which are first-order measures, the variance is a second-order effect. It is used in economic, financial and insurance contexts. For example, automobile insurance is a means by which an individual can reduce the possibility of a large cost due to an accident. Another example is modern portfolio management, which uses the variances of potential investments to determine the best way to allocate the total investment, subject to the amount of risk tolerance of the investor.

In the case of the risks associated with AI systems, any of the risk measures listed above might be appropriate. When developing new AI systems, the first measure in the list above would apply. On the other hand, when considering the effects of the introduction of AI systems on society, the second or third measure would be more appropriate, depending on whether one can quantify the potential impact. The fourth risk measure in the list could be used by an enterprise that is using products supplied by AI companies (Baclawski et al., 2024).

General Risks

The following are some of the general risks for software systems that also apply to AI systems:

1. The transparency of a system is determined by the extent of documentation of its scope, mission, policies, capabilities and status. There is a strong incentive for companies to maintain trade secrets to ensure advantages over their competitors. This makes it difficult for someone who is not part of a company to understand its systems, especially their limitations.
2. The next risk to consider is responsibility for the consequences of the use of a system. What are the standards for stewardship involved? There is a risk that developers of an AI system ignore issues of stewardship and responsibility because they do not regard these issues as being relevant to the capabilities of current AI systems and assume that the customer is responsible for any harmful use of their system.
3. Ease of use is a desirable feature of a system, but it also has risks. Chat systems support easy user dialogues and provide plausible responses. As a result chat systems can be seductively attractive for users, but a risk is that users may become overly reliant and may no longer bother to check that the responses are valid.
4. Sustainability of any system involves organizational resources and funding, digital object management, technical infrastructure, and security risk management.

5. The energy and physical infrastructure requirements for ML can be very large to the point of having an adverse impact on the environment.

This illustrates the fact that there are many interacting factors in assessing the risks of software systems.

Specific Risks

There are some risks that are more specific to AI systems because of the way that they are trained and architected, and many of these have been described in NIST (2024). The following risks are of note:

1. A major concern is that a large corpus naturally contains biases and misinformation. Training is one of the foundations of subsymbolic systems, and the resulting knowledge model can perpetuate biases and misinformation in its outputs. This can result in generating discriminatory content and/or factually incorrect information which sounds plausible but isn't grounded in reality. Users can easily mistake these fabrications as established facts.
2. Many current types of subsymbolic systems, such as those based on LLMs, are inherently unreliable and should never be used in mission- or life-critical applications without effective monitoring and supervision. A symbolic system might help in this monitoring and/or act as an executive by using meta-cognition to guide executive activities. A particular suggestion to manage misinformation risk is to mitigate these by using RAGs which don't use the LLM's knowledge base but use a core set of information for its decisions.
3. Current AI systems lack transparency ranging from how they are trained to their architecture. Indeed, even if the developers of an AI system were willing to be transparent, it might not be possible because such systems often lack explainability of their functioning and results. Despite the widespread deployment of foundation models based on standard deep learning and transfer learning we currently lack a clear understanding of how they work, when they fail, and what they are even capable of due to the potential for emergent properties as discussed in the Benefits of Subsymbolic Systems section above. This demands caution, because any defects in a foundation model will be inherited by all the adapted models downstream. Many believe that to tackle this risk, much of the critical research on foundation models will require deep interdisciplinary collaboration commensurate with their fundamentally sociotechnical nature. (Bommasani et al., 2021)

Because of the specific risks of AI systems mentioned above, new methods of monitoring and evaluating system function and impact may be needed. As it has been widely noted existing approaches to emerging AI system evaluation have tended to focus exclusively on model evaluation. Are they, for example, generating correct responses? However, such evaluations may potentially be "less sensitive to more general ways in which AI assistants may underperform when considered as part of a broader sociotechnical system. New methodologies and evaluation suites focusing in particular on human-AI interaction, multi-agent and societal effects are needed to support strong evaluation and foresight in this area" (Gabriel et al., 2024). More about these issues will be discussed in the Ethics section below.

Regulations

The risks associated with AI systems has led to regulations in many countries and even for more extreme measures such as halting development entirely. It is beyond the scope of this article to survey this very large area. We only discuss some examples of how companies, governments and professional organizations are addressing AI risks.

Open AI has developed the following list of safety standards in order to mitigate some of risks mentioned above (Open AI, 2024):

1. Minimize harm by the promise to "build safety into our AI tools where possible, and work hard to aggressively reduce harms posed by the misuse or abuse of our AI tools."
2. Build trust in the user and developer community. The promise is to share the responsibility of supporting safe, beneficial applications of our technology.
3. Learn and iterate by observing and analyzing how our models behave and are used and seek input on our approach to safety in order to improve our systems over time.
4. A promise to "Be a pioneer in trust and safety". Open AI promised to support research into the unique trust and safety challenges posed by generative AI, to help improve safety beyond our ecosystem.

Some regulations, such as in the European Union's AI Act, consider reasonableness and other principles for deciding an adequate amount of generative AI risk management. It requires, for example, that algorithms be explainable for "high-risk AI systems" such as those deployed for remote biometric identification, law enforcement or access to education, employment or public services (Hutson, 2023). Current subsymbolic systems used as components of chat systems are not categorized as high-risk and might escape this legal need for explainability except in some specific use cases such as in healthcare.

The IEEE recommended practice for defining and evaluating AI risk in particular emphasizes safety, trustworthiness, and responsibility as well as considering the global context in adopting AI systems (Baclawski et al., 2024). Ideas for governance, and collaboration are also considered. It is worth noting that simple AI concepts have been defined by IEEE (ISO/IEC TR 24027).

Risks Specific to NeSy Systems

As has been noted throughout this article, hybrid systems combining symbolic and subsymbolic capabilities offer potential advantages over either approach by themselves. However, there are many possible NeSy architecture as discussed in the Neuro-Symbolic Systems section. Accordingly, we are still in the early phases of research, and thus experience with serious applications are a ways off. Nevertheless, it is prudent to anticipate and mitigate risks that have already been identified, as discussed above in this section. Aside from the risks already identified, NeSy systems may have risks specific to such systems.

- The combination of subsymbolic and symbolic components adds complexity to the system, which could make NeSy systems more difficult to debug and to control. This could have an impact on reliability and sustainability.
- There is much less experience with NeSy systems than with subsymbolic and symbolic systems by themselves. Moreover, each NeSy architecture will need to be tested and analyzed to build safety and trust as required by the regulations listed above. Since there are may architectures, there is a risk that there will not be enough resources to build safety and trust in every architecture.
- In principle, NeSy systems should be more transparent and explainable, but there is the risk that the combination of subsymbolic and symbolic may be even less transparent and explainable because it may be difficult to identify which component has responsibility for a result. This risk could be mitigated by carefully designing the system to be explainable.
- NeSy systems may not have any effect on biases or misinformation, or they could magnify these errors, unless steps are taken to monitor and mitigate biases and misinformation. For

example, if the symbolic component of the NeSy architecture is a front-end to the system, then it could attempt to reverse errors that were produced by the subsymbolic component, or it could at minimum detect and report them.

As systems become more capable with human-like capabilities, the impact of some risks over time, such as replacing people in typical tasks, will be harder to measure. Expensive random trial experiments (RTEs) may be needed (Yohsua et al, 2024). In clinical studies, such as diagnosis and outcome predictions, RTEs are widely recognized as the soundest way to determine the actual value of an automated system. The value of advanced NeSy systems is likely to be tested for clinical practices and actual outcomes over time. A first step for this and system design, evaluation and monitoring would be to leverage responsible, professional software development standards and give them enough strength to work in particular domains like healthcare with their own professional standards and practices. More discussion of these issues is presented in the Ethics section below.

Ethics

AI systems raise many ethical issues. While most of these issues are the same as for any software system, there are some issues that are unique. A complete exploration of AI ethics is beyond the scope of this article; however, the broad outline of several general approaches to support ethics for NeSy systems can be listed. These approaches are accompanied by a use case to illustrate the role of an ontology or knowledge graph. We then list some examples of recommendations for AI ethics. We end with some general ethical considerations. For more about AI ethics see the AI Ethics Panel at the Ontology Summit 2024 (Bennett & Krishnan, 2024).

1. Domain-specific ethical considerations are critical in computer security, health care and many other areas of specialization. As noted in the Blueprint for an AI Bill of Rights, "Systems should undergo pre-deployment testing, risk identification and mitigation, and ongoing monitoring that demonstrate they are safe and effective based on their intended use, mitigation of unsafe outcomes including those beyond the intended use, and adherence to domain-specific standards" (OSTP, n.d.). Domain-specific standards can be implemented as ontologies.
   Use case: Governance of a healthcare AI model can be implemented using the HL7 FHIR data model, in the manner attempted by Das and Hussey (2023) and including domain-specific privacy and security constraints suggested by a NIST Working Group – and implemented as ontologies (Chang et al., 2018).
2. Subsymbolic systems can generate datasets used for subsymbolic system training and other machine learning applications. Complete automation may not be possible in all cases, but artifacts generated by subsymbolic systems can be analyzed using rules in an ontology-based framework.
   Use case: Mousavi and Termehchy (2023) demonstrated the use of a simple ontology of people to implement declarative constraints on a subsymbolic system, and also survey other approaches to accomplish similar results.
3. By providing an affordable and ubiquitous NLP interface, a subsymbolic system can enable individuals and groups to interrogate AI implementations of interest to them or their organizations. This can include AI training sets, provenance, test engineering and interoperability. Ontologies provide explicit guidance over what information can be provided and to whom as well as guidance as to what questions to ask.
   Use case: A user can query a subsymbolic system to understand that a subsymbolic system voice model was trained only on English speakers, but a data protection ontology limits responses to anonymized information about individual speakers.

4. Subsymbolic systems can display bias even when guardrails have been erected to address them (Kotek et al., 2024). They note that "bias [occurs] across minoritized groups, but in particular in the domains of gender and sexuality, as well as Western bias, in model generation. The model not only reflects societal biases, but appears to amplify them." A closely related problem is that of data protection. Certain classes of data, such as Personally Identifiable Information (PII) or NIST SP 800-171 "Controlled Unclassified Information" (CUI) are similar to Protected Classes in that they must be identified in order to perform certain analyses – even to avoid bias – yet must also be protected from unwanted or malicious access.
   Use Case: Zhao et al. (2024) propose a plugin for ontology tools like Protege which could exploit subsymbolic systems to facilitate ontology development. This approach is one of several which could identify and protect CUI and bias against protected classes by more clearly identifying them – and the associated risks – when maintaining domain specific ontologies. Similarly, the Enterprise Knowledge team, in a blog post, note that ontology prompting techniques can provide the subsymbolic system with relevant ontological information about protected classes during inference.
5. A subsymbolic system integrated with an ontology built on AI ethical principles could identify or avoid potential liability by monitoring subsymbolic processes and utilization according to ethical principles.
   Use Case: In Harrison et al. (2021) a proof of concept was built which anticipates "... machine readable ethical AI principles, an agreed schema or even legally enforced standard ...[And which] could be read directly into AI entities, with the presence and implementation of the principles auditable by regulatory authorities, and adherence even standing to lower legal liability and damages apportioned to developers or the owners of AI systems." Ontologies can also identify the authoritativeness of certain sources, such as identifying the potential for sarcastic content in discussion forums (Irwin, 2024).

Individual countries, country alliances (notably the EU), nonprofits and standards organizations are in the process of advocating or developing recommendations for AI users. A partial list of these includes:

- European Commission's Ethics Guidelines for Trustworthy AI
  `https://bit.ly/4er6zGa`
- Montréal Declaration for Responsible AI Development
  `https://bit.ly/3KJVnXP`
- Organization for Economic Cooperation and Development AI Principles `https://bit.ly/45rZIIE`
- ISO/IEC TR 24368:2022 Information technology – Artificial intelligence – Overview of ethical and societal concerns `https://bit.ly/45vSLGg`
- ISO/IEC 42001:2023 Information technology – Artificial intelligence – Management system
  `https://bit.ly/3z3XzXs`
- UNESCO Recommendation on the Ethics of Artificial Intelligence
  `https://bit.ly/4bhG5of`
- US White House Blueprint for an AI Bill of Rights
  `https://bit.ly/3VqVQ5X`
- NIST Trustworthy & Responsible Artificial Intelligence Resource Center (AIRC)
  `https://airc.nist.gov/Home`
- IEEE 7000:2021 IEEE Standard Model Process for Addressing Ethical Concerns during System Design `https://bit.ly/3L89qGV`

- IEEE 7001:2021 IEEE Standard for Transparency of Autonomous Systems https://bit.ly/3RvXXED
- IEEE 7010-2020 IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being https://bit.ly/3KJVvGK
- Other IEEE working groups are drafting standards for Big Data Metadata, Security and Trustworthiness Requirements in Generative Pretrained AI Models, Ethically Aligned Educational Metadata in Extended Reality (XR) & Metaverse, Data and AI Literacy, Skills, and Readiness, and Recommended Practice for Defining and Evaluating AI Risk, Safety, Trustworthiness, and Responsibility. IEEE also sponsors its Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS).

Some, but not all, of the guidelines and standards listed above were the source material from which the AI Principles Ontology (AIPO) was built (Harrison et al., 2021). Among the AI ethics standards that have been produced, one of particular interest to the ontology community is IEEE 7007-2021 – which is also cited by Harrison et al. (2021).

Ontologies can support a standard reference model of ethical practices against which subsymbolic systems can be used to (a) fuse multiple data sources, (b) generate design-stage requirements to support ethics, (c) generate tools to collect audit data, (d) compare audit results against the reference models to assess compliance.

NeSy systems using simpler architectures in the Kautz classification as well as NeSy systems for narrow applications seem likely to be developed before systems with more complex architectures or intended for more general application domains. Consequently, there should be time to learn about and to address the risks of NeSy systems. It is hoped that experiences with such systems will lead to a better understanding of the constraints that must be imposed so that AI systems will be ethical. Unfortunately, as mentioned already, companies developing products are reluctant to reveal details about their systems.

Some authors regard technological uncertainty, incomplete data, and management errors to be the main sources of ethical risks (Guan et al., 2022). Some of these authors seek to define constraints within the software and systems. It is an open question how to define specifications for things that we never want AIs to be able to do. We do not yet have a standard way to implement this class of ethical constraints. Rules represented in textual form are open to indefinite interpretation, some of which are interpretations that humans would not agree with. Any simple attempt to build morality into machines seems subject to this problem of interpretation. For other authors the most important aspect of ethical AI is in how we humans use AI. An example which already seems to be trending out of control is the use of autonomous drones for offensive purpose that are not under the control of a human, such drones should never be able to use deadly force.

Summary

We are currently at the outset of an AI boom which has captured the imagination of the public as well as the interest of large companies. An important consideration for any information or software based system is that its results be correct, especially for mission- and life-critical applications.

Most current AI research and development has been devoted to subsymbolic systems, which have many disadvantages. Subsymbolic systems can generate factually incorrect information that is especially problematic because it is presented in a plausible fashion. This has led to many calls for subsymbolic systems to have "guardrails" based on a reliable ontology or other semantic resource so that inconsistencies and errors can be rejected.

Symbolic techniques, such as semantic techniques, are highly developed, have many applications other than AI, and have advantages and disadvantages that are complementary to subsymbolic techniques. This suggests combining the subsymbolic techniques with symbolic techniques, leading to NeSy systems.

In this article we presented some of the most prominent examples of NeSy systems, the proposed classifications of such systems, and some of their applications. There are many risks associated with AI systems in general and NeSy systems in particular. In addition to listing some of the risks that are unique to such systems, we also gave a framework for expressing and quantifying different kinds of risks. The ethics of AI system is a very broad area so a complete survey was not possible, but broad outline was presented.

References

AllegroGraph (n.d.). Neuro-Symbolic AI with AllegroGraph. Accessed 7 July 2024. https://bit.ly/3UowUMY

K. Baclawski (2022). "What is Risk?" Ontology Summit 2022, 23 Mar 2022. https://bit.ly/3NdWJde

K. Baclawski, G. Berg-Cross, M. Bennett, L. Dickerson, T. Schneider, S. Seppälä, R. Sharma, R. Sriram, & A. Westerinen (2021). "Ontology Summit 2021 Communiqué: Ontology Generation and Harmonization," *Applied Ontology* 17(2) 233–248. IOS Press, The Netherlands

K. Baclawski, M. Bennett, & G. Waters (2024). AI Risk Panel. Ontology Summit 2024, 1 May 2024. https://bit.ly/48JeBGq

M. Bennett & R. Krishnan (2024). AI Ethics Panel. Ontology Summit 2024, 8 May 2024. https://bit.ly/48JBJo5

M. Bennett & A. Westerinen (2023). Ontologies and Large Language Models Related but Different, Ontolog Summit 2024 Fall Series, 15 November 2023, https://bit.ly/3G9wYZi

A. Bernasconi, A. G. Simon, G. Guizzardi, L. O. B. da S. Santos, & V. C. Storey (2023) "Ontological Representation of FAIR Principles: A Blueprint for FAIRer Data Sources," in Advanced Information Systems Engineering: 35th International Conference, CAiSE 2023, Zaragoza, Spain, June 12–16, 2023, Proceedings, Berlin, Heidelberg: Springer-Verlag, pp. 261–277. doi: 10.1007/978-3-031-34560-9_16.

R. Bommasani, D. Hudson, E. Adeli, et al. (2021). "On the Opportunities and Risks of Foundation Models" arXiv:2108.07258

M. J. Buehler (2024). "Accelerating Scientific Discovery with Generative Knowledge Extraction, Graph-Based Representation, and Multimodal Intelligent Graph Reasoning." Ontology Summit 2024, 27 March 2024. https://bit.ly/4bYBIzw and arXiv:2403.11996

W. C. Burns (1997). "Spurious Correlations." 1997. https://web.archive.org/web/20190925212058/ and http://www.burns.com/wcbspurcorl.htm

D. Bzdok, N. Altman, & M. Krzywinski (2018). "Statistics versus Machine Learning." *Nature Methods.* 15(4): 233–234. doi:10.1038/nmeth.4642. PMC 6082636. PMID 30100822.

K. Cagle (2023) Complementary Thinking: Language Models, Ontologies and Knowledge Graphs. Ontology Summit 2024, 18 October 2023. https://bit.ly/48OegDD

W. Chang et al. (2018). "NIST Big Data Interoperability Framework," NIST, Gaithersburg MD, June 2018. doi: 10.6028/NIST.SP.1500-10

S. Das & P. Hussey (2023). "HL7-FHIR-Based ContSys Formal Ontology for Enabling Continuity of Care Data Interoperability.," J Pers Med, vol. 13, no. 7, June 2023, doi: 10.3390/jpm13071024

F. Djoumbou, Yannick, et al. (2016). "ClassyFire: automated chemical classification with a comprehensive, computable taxonomy." *Journal of cheminformatics* 8: 1-20.

C. Fellbaum (1998) *WordNet: An Electronic Lexical Database.* MIT Press.

I. Gabriel, et al. (2024). "The ethics of advanced AI assistants." arXiv preprint arXiv:2404.16244.

S. M. García, C. V. Ayala, & I. Pineda (2018). "An overview of ontology learning tasks." *Computación y Sistemas* 22.1: 137-146.

A. Garcez & L. C. Lamb (2023). "Neurosymbolic AI: The 3^{rd} wave." *Artificial Intelligence Review* 56.11: 12387-12406.

H. Guan, L. Dong, & A. Zhao (2022). "Ethical risk factors and mechanisms in artificial intelligence decision making." *Behavioral Sciences* 12.9: 343.

H. Giglou, J. D'Souza, & S. Auer (2023). "LLMs4OL: Large language models for ontology learning." International Semantic Web Conference. Cham: Springer Nature Switzerland.

M. Glauer, A. Memariani, F. Neuhaus, T. Mossakowski, & J. Hastings (2022). "Interpretable Ontology Extension in Chemistry," *Semantic Web*. 1-22. 10.3233/SW-233183.

M. Glauer, F. Neuhaus, T. Mossakowski, & J. Hastings (2023). "Ontology Pre-training for Poison Prediction," German Conference on Artificial Intelligence (Künstliche Intelligenz). Cham: Springer Nature Switzerland.

A. Harrison, D. Spagnuelo, & I. Tiddi (2021). "An Ontology for Ethical AI Principles," *Semantic Web*, Jan. 2021, Accessed: May 20, 2024. https://bit.ly/45wcupr

J. Hastings, M. Glauer, A. Memariani, F. Neuhaus, & T. Mossakowski (2021). "Learning Chemistry: Evaluating machine learning for the task of structure-based chemical ontology classification," *Journal of Cheminformatics*

P. Hitzler, F. Bianchi, M. Ebrahimi, & Md. Sarker (2020). "Neural-Symbolic Integration and the Semantic Web," *Semantic Web* 11 (1), 3–11. https://bit.ly/4c0WeiD

M. Hutson (2023). "Rules to keep AI in check: nations carve different paths for tech regulation." *Nature* 620.7973: 260-263.

IEEE Std 7007-2021, "IEEE Ontological Standard for Ethically Driven Robotics and Automation Systems", pp. 1–119, Nov. 2021, doi: 10.1109/IEEESTD.2021.961 https://bit.ly/3KOhXhV

Kahneman, D. (2011). *Thinking, Fast and Slow*, Farrar, Straus and Giroux, ISBN 978-0374275631 OCLC 706020998.

L. Irwin (2024). "Google updates AI systems after erroneous information generated," The Hill, May 31, 2024. Accessed: May 31, 2024. https://bit.ly/3z8FZBK

ISO/IEC TR 24027:2021(en) Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making ISO/IEC JTC 1. 2024. Standards by ISO/IEC JTC 1/SC 42 Artificial intelligence

H. Kautz (2021). "Toward a Taxonomy of Neuro-Symbolic Systems," Ontology Summit 2021, Accessed April 7, 2021. http://bit.ly/2MRCWFG

H. Kautz (2022). "The Third AI Summer: AAAI Robert S. Engelmore Memorial Lecture." *AI Magazine* 43.1: 105-125.

M. Kelly & A. B. Barron (2022). "The best of both worlds: Dual systems of reasoning in animals and AI." *Cognition* 225: 105118.

H. Kotek, D. Q. Sun, Z. Xiu, M. Bowler, & C. Klein (2024). "Protected group bias and stereotypes in Large Language Models." Accessed: May 30, 2024. https://arxiv.org/abs/2403.14727

L. C. Lamb, A. Garcez, M. Gori, M. Prates, P. Avelar, & M. Vardi (2020). "Graph Neural Networks Meet Neural-Symbolic Computing: A Survey and Perspective." *IJCAI*

P. Lomov, M. Malozemova, & M. Shishaev (2022). "Extracting Relations from NER-Tagged Sentences for Ontology Learning." Computer Science On-line Conference. Cham: Springer International Publishing

J. Mao, C. Gan., P. Kohli, J. Tenenbaum, & J. Wu (2019). The Neuro-Symbolic Concept Learner: Interpreting scenes, words, and sentences from natural supervision. https://doi.org/10.48550/arXiv.1904.12584

V. Mann, S. Viswanath, S. Vaidyaraman, & V. Venkatasubramanian (2023). Accelerating Drug Discovery and Development Using an Ontology-Based Machine Learning Framework 2023 AIChE Annual Meeting

R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, & L. De Raedt (2018). Deepproblog: Neural probabilistic logic programming. https://arxiv.org/abs/1805.10872

G. Marcus (2020). "The next decade in AI: four steps towards robust artificial intelligence." arXiv:2002.06177

G. Marcus (2024). "No AGI (and no Trustworthy AI) without Neurosymbolic AI," Ontology Summit 2024, 28 February 2024. https://bit.ly/42Y5T5X

D. L. McGuinness (2023). The Evolving Landscape: Generative AI, Ontologies, and Knowledge Graphs, Ontology 2024 Pre-Summit, 11 October 2023. https://bit.ly/3LYsfgh

X. Meng, et al. (2024). "The Application of Large Language Models in Medicine: A Scoping Review." *iScience*

M. Missikoff, R. Navigli, & P. Velardi (2002). "Integrated Approach to Web Ontology Learning and Engineering." *Computer* 35(11) 60–63.

T. Mossakowski (2024). Neuro-symbolic integration for ontology-based classification of structured objects, Ontolog Summit 2024, 20 March 2024. https://bit.ly/4c0fP2N

J. Mousavi & A. Termehchy (2023). "Towards Consistent Language Models Using Declarative Constraints." Joint Workshops at 49^{th} International Conference on Very Large Data Bases (VLDBW'23) – Workshop on LLMs and Databases (LLMDB'23). https://arxiv.org/abs/2312.15472

F. Neuhaus (2023). "Ontologies in the era of large language models – a perspective." *Applied Ontology* 18.4: 399-407.

H. Naveed, et al. (2023). "A comprehensive overview of large language models." arXiv preprint arXiv:2307.06435

NIST AI 600-1 (2024). "AI Risks and Trustworthiness," Artificial Intelligence Risk Management Framework, Section 3. https://bit.ly/4fnczR5

Ontology Summit 2020. "Knowledge Graphs," https://bit.ly/33NydJ9

Ontology Summit 2024. "Neuro-Symbolic Systems with and for Ontologies and Knowledge Graphs," https://bit.ly/3RDw4LD

Open AI (2024). "Product safety standards." https://openai.com/safety-standards/

OSTP (n.d.) Blueprint for an AI Bill of Rights. The White House Office of Science and Technology Policy (OSTP). https://bit.ly/3VqVQ5X

K. Roy (2024). "Healthcare Assistance Challenges-Driven Neurosymbolic AI" Ontolog Summit 2024, 24 April 2024. https://bit.ly/44fIbCH

Rutesic, Hochschule, Pfisterer, Fischer, & Paulheim (2023). Ontology-Based Models of Chatbots for Populating Knowledge Graphs October 2023 DOI:10.1007/978-3-031-47745-4_17 In Knowledge Graphs and Semantic Web (pp. 228-242)

T. P. Sales, N. Guarino, G. Guizzardi, & J. Mylopoulos (2017). "An ontological analysis of value propositions," in 2017 IEEE 21st International Enterprise Distributed Object Computing Conference (EDOC), IEEE, pp. 184–193.

R. Schaeffer, B. Miranda, & S. Koyejo, (2024). Are emergent abilities of large language models a mirage?. *Advances in Neural Information Processing Systems* 36.

I. Schlag & J. Schmidhuber (2018). Learning to reason with third-order tensor products. CoRR, abs/1811.12143

M. Schreiner (2023). "GPT-4 architecture, datasets, costs and more leaked," The Decoder. July 11, 2023. https://bit.ly/4e6gZK9

A. Sheth (2024). Forging Trust in Tomorrow's AI: A Roadmap for Reliable, Explainable, and Safe NeuroSymbolic Systems, Ontolog Summit 2024, 17 April 2024. https://bit.ly/48TZ0nL

J. Sowa (2024). "Without Ontology, LLMs are clueless," Ontology Summit 2024, 28 February 2024. https://bit.ly/3uHQv10 https://bit.ly/3TahUli

C. Steinbeck, et al. (2006). "Recent developments of the chemistry development kit (CDK)-an open-source java library for chemo-and bioinformatics." *Current pharmaceutical design* 12.17: 2111-2120.

Z. Susskind, et al. (2021). "Neuro-symbolic ai: An emerging class of ai workloads and their characterization." arXiv:2109.06133

A. Thirunavukarasu, et al. (2023). "Large language models in medicine." *Nature medicine* 29.8: 1930-1940.

V. Venkatasubramanian (2024). "Ontology-based Machine Learning in Pharmaceutical Engineering" Ontolog Summit 2024, 24 April 2024. https://bit.ly/44g9X26

J. Wei, Y. Tay, R. Bommasani, et al. (2022). "Emergent Abilities of Large Language Models," arXiv:2206.07682 https://bit.ly/4cEJT3R

B. Yohsua, et al. (2024). International Scientific Report on the Safety of Advanced AI. Diss. Department for Science, Innovation and Technology.

D. Yu, et al. (2023). "A survey on neural-symbolic learning systems." *Neural Networks*

Y. Zhao, N. Vetter, & K. Aryan (2024). "Using Large Language Models for OntoClean-based Ontology Refinement." Accessed: May 30, 2024. https://arxiv.org/abs/2403.15864

KENNETH BACLAWSKI is an Associate Professor Emeritus at the College of Computer and Information Science, Northeastern University. Professor Baclawski does research in data semantics, formal methods for software engineering and software modeling, data mining in biology and medicine, semantic collaboration tools, situation awareness, information fusion, selfaware and self-adaptive systems, and wireless communication. He is a member of the Washington Academy of Sciences, IEEE, ACM, IAOA, and is the chair of the Board of Trustees of the Ontolog Forum.

MIKE BENNETT is the Director of Semantics and Standards at the Enterprise Data Management Council, an industry association in the financial services industry. Mike is the editor of the EDM Council Semantics Repository where I have created a framework for editing semantically rich material and presenting it in a technology neutral way for business subject matter experts. This is a long way from being perfect but I am committed to presenting business meaning to business people, but within a rigorous framework.

DR. GARY BERG-CROSS is a cognitive psychologist (PhD, SUNY-Stony Brook, and is now semi-retired from a professional life that included R&D in applied data & knowledge engineering, collaboration, and AI research. He has taught at several colleges and universities (SUNY, Widener, University of Delaware, George Washington, George Mason University, and others). He serves on several professional Boards and is on the Board of Discipline Editors for the Washington Academy of Sciences. Formerly a co-chair of the Research Data Alliance work-group on Data Foundations and Terminology he now co-chairs the Earth Science Information Partner's Semantic harmonization cluster.

TODD SCHNEIDER is an ontologist and semantic architect, extending enterprise and systems architecture with explicit representations of the semantics of the domain modeled. He is co-chair of the Technical Oversight Board of the Industrial Ontology Foundry (IOF) and on the Board of Trustees of the Ontolog Forum. Previously, he was a Senior Principal Engineer and architect at Raytheon and representative to the Network Centric Operations Industry Consortium (NCOIC) where he chairs the Semantic Interoperability Framework and Net-Centric Attributes Working Groups. He is an active member of the Ontolog forum participating in the development of the Open Ontology Repository and was on the ISWC 2009 organizing committee. Through his work at Raytheon, NCOIC, and the Ontolog Forum he is advancing the notions of semantic clarity and derivation, aided by the use of semantic technologies, architecture and modeling.

RAVI SHARMA is a nuclear particle physicist. He received the NASA Apollo Achievement Award, ISRO Distinguished Service Awards and several others. He is Fellow of IETE, General Motors and has published in nuclear, space and ontology Journals. Leadership experience in Enterprise

Architecture in US Federal, State, DoD agencies, and General Motors. He is voting member of SAE Fuel Cell Standards Committee, Board of Trustees Ontolog Forum, and continues research in Vedic and Modern Particle Physics. He has held Leadership Positions in IT Industries in US and India and transformed organizations such as ISRO, Earth Observation agencies and several industries.

MARK UNDERWOOD is the co-founder and CEO of Krypton Brothers, a New York based boutique consultancy. He chairs the IEEE P2957 Working Group to develop a standard for a Reference Architecture for Big Data Governance and Metadata Management. Mark lists these lessons: be an avid researcher, bookmarker, critic. Inhale.

ANDREA WESTERINEN is an independent software engineer and systems architect, and CTO of OntoInsights, LLC (http://ontoinsights.com). She specializes in ontology development and knowledge engineering, and has extensive software development experience. Ms. Westerinen has strong interests in semantic and linguistics technologies, and has worked in the computer industry since 1979, at places like Raytheon/BBN, Two Six Labs, SAIC, CA Technologies, Microsoft, Cisco, Intel and IBM. Her responsibilities have included researcher, strategist, program manager, personnel manager, software developer, ontologist and enthusiast, as needed. She has a B.S. in Physics and Mathematics from Marquette University, and an M.S. in Computer Science from Nova Southeastern University.