A Risk Management Framework for Large Language Models

Alec Crawford & Frank Fitzgerald
Artificial Intelligence Risk, Inc.
April 15, 2024
aicrisk.com

"Invest in technology. The savings compound, it gives you an advantage over slower moving competitors and can be the difference between a profit and a loss."

- Andrew Carnegie (b. 11/25/1835, d. 8/11/1919)

Introduction

In this paper, we introduce a framework for managing governance, risk, compliance, and cybersecurity (GRCC) for large language models from inception through implementation, use and decommissioning. In an age where data is the new currency and artificial intelligence (AI) its most proficient banker, the development and deployment of large language models (LLMs) stand as both an incredible achievement and a source of entirely new types of risk. We will address those risks first, and then governance, compliance, cybersecurity and finally discuss risks from artificial general intelligence, although that is not the main topic of this paper.

We argue that a GRCC platform for managing different LLMs and agents developed for those models – **AI GRCC -- is an entirely new category of software,** akin to an operating system for AI. With base models quickly leapfrogging each other in capability, it is important that companies can treat them as interchangeable pieces of software on a platform rather than platforms themselves. In addition, as GRCC platforms for AI become more central to business, they must provide companies with incredible flexibility: to create agents, allocate computing resources, switch base models, address emerging risks, enable regulatory compliance and testing, and track and shut down cybersecurity attacks.

Executive Summary

This paper offers a framework for risk management of artificial intelligence systems, with a focus on Large Language Models (LLMs). The reason this is so important, and different from other AI models, is that instead of only experts interacting with these models, anyone can interact with them using their natural language interface. The paper covers Artificial Intelligence (AI) governance, risk, compliance, and cybersecurity (AI GRCC) with a focus on the risks associated with AI implementation in medium and large organizations. Organizations are encouraged to establish comprehensive AI Governance, Risk Management, and Compliance (GRCC) programs and teams prior to embarking on AI projects to mitigate such risks effectively.

Following is an outline of the paper:

- Overview of LLMs, contraindications for their use and industry-specific impact examples
- Organizational structure recommended for governance, risk, compliance, cybersecurity as well as ethics management
- Risk tiering and identifying critical, high, medium, and low risk items based on societal and organizational impact
- Risk timing to identify those risks that are now present, near-future, eventually
- Broad risk categories of training data, regulatory compliance, user inputs, model outputs, and cybersecurity.
- We discuss risk management methods that vary based on risk category, with special methods for AI cybersecurity.
- Cybersecurity including secure deployment and hacking detection (e.g. DAN-style attacks and prompt injections)
- LLM risk management lifecycle, including constant updates and testing.

- Governance of user access to data, documents, and models, emphasizing limited access
- Regulatory compliance across the globe, including special rules for financial services (e.g. the SEC rules in the US)
- A note on artificial general intelligence (AGI) risks
- The Appendix introduces new Key Performance and Risk Indicators (KPIs and KRIs), which are copyrighted by Artificial Intelligence Risk, Inc.
- References

Organizational Structure for AI GRCC

While everything in this paper is important to do, there are three key items to discuss first: who is responsible for defining the AI GRCC and ethics programs, who is implementing it, and who is checking the results of the different key performance and risk indicators (KPIs and KRIs) specified in the Appendix and beyond.

First, we would argue that the mandate for AI GRCC needs to come from the top of an organization for people to take it seriously. And you should take AI GRCC seriously, or legal, reputational, and all kinds of other risks come into play. A single, responsible person should be designated not just for AI performance, but simultaneously for the AI GRCC program.

Second, the person in charge of AI GRCC and Ethics needs a group to assist. This goes beyond the technologists to subject matter experts, risk and compliance experts, and AI ethics experts for most firms. Enabling that team, led by the individual, responsible person is also critical.

Third, as the AI models and systems are developed, AI GRCC and ethics should always be a consideration. The KPIs and KRIs listed in the Appendix need to be calculated, reported, and monitored or something dangerous can slip through a crack. Who is creating that? What key person at your organization is the responsible person for monitoring and escalating problems? How do you integrate that process directly into the corporate or advisory Board?

Overview of Large Language Models

Shifting gears, Artificial Intelligence (AI) has existed since at least the Dartmouth Conference in 1956 in terms of actual machines and was conceived of well before that time. Over the years, humans have kept redefining what artificial intelligence is. For example, when DeepBlue beat Gary Kasparov, the reigning human world chess champion, in a chess game in 1996, that was briefly viewed as true AI, until people realized that it was not really thinking, but using historical data and brute-force analysis of millions of combinations of potential chess moves into the future outcomes to figure out the next best chess move. In other words, “narrow” problem solving, not human-level intelligence. Today, while LLMs are amazing, most of the world things that “real AI” would be in the form of Artificial General Intelligence (AGI), or a machine that can think, learn, reason, act, and maybe even feel like a human. Importantly, it must also have self-

awareness and follow ethical conduct. This is a far cry from the original Turing Test, proposed by Alan Turing in 1950.[1]

The recent advances in AI Large Language Models (LLMs) have precipitated transformative shifts across various industries, from automating customer service through chatbots to providing financial analysis and securities trading at a speed and scale impossible for humans alone. LLMs have revolutionized content creation, energized software development, and unlocked new potential in medical research by parsing and synthesizing vast data sets. Yet, with monumental capabilities come substantial responsibilities and risks — responsibilities to administer these models ethically and risks that span the spectrum of legal, regulatory, confidentiality, privacy, bias, and security. In some cases, use of the data is in question from a legal or at least an ethical standpoint. Because of their novel nature, most of these risks and methods of mitigating them are brand new and cannot simply be copied from existing software and governance, risk, and compliance (GRC) models and software.

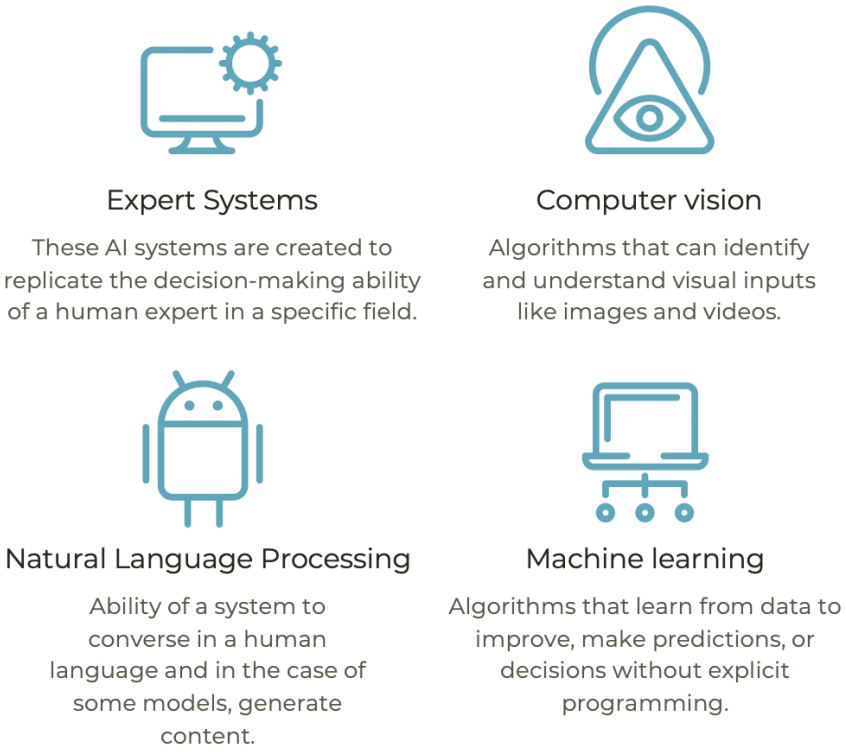
Exhibit: New Capabilities Require a New Category of Software
AI Platform Governance, Risk, Compliance, and Cybersecurity (GRCC)

[1] Alan Turing proposed that a human evaluator would judge natural language conversations between a human and a machine designed to generate human-like responses. The evaluator would be aware that one of the two partners in conversation was a machine, and all participants would be separated from one another. The conversation would be limited to a text-only channel, such as a computer keyboard and screen, so the result would not depend on the machine's ability to render words as speech. If the evaluator could not reliably tell the machine from the human, the machine would be said to have passed the test.

Different AI Model Types

Unfortunately, now, "artificial Intelligence" has become a marketing term. In reality, there are many different types of software and systems that attempt to replicate what humans can do. They can also be combined to create "composite AI", a concept one of the authors broached in the 1980's and is now seeing a resurgence.

Exhibit: Artificial Intelligence Is a Marketing Term and There Are Many Types of AI



Large Language Models (LLMs)

Large language models build on techniques derived from prior AI with the goal of being able to engage in normal conversation with humans, including answering questions, summarizing documents, and "creative" writing. The current popular version, "GPT", stands for generalized pre-trained models. The models are pre-trained using vast amounts of existing natural language text, so they are ready for us mere mortals to use them out of the box.

The key to the entire LLM concept is they use the user's prompt and the prompt completion from the model to predict the next word in the sequence.[2] What's the most likely completion of the following sentence: "My favorite ice cream flavor is…"? We would bet most people would say vanilla or chocolate, and that is also what an LLM would probably say. But which one of

[2] Not to get too technical, but the models actually predict a piece of a word called a "token", but we can think about it as a word to avoid getting lost in the details.

those answers is picked (chocolate or vanilla or something else) is typically **randomized by the model** to avoid repetitive or boring answers, and probabilities are based on the training data and may not necessarily be representative of the world. Because the training data may be biased towards a specific country or culture, that bias may also skew answers. If you ask the same question multiple times, it will give you different answers, creating an inconsistency in results that can be problematic in many contexts.

The reason this aspect of LLMs is so important to understand is that LLMs are **not reasoning**. They are merely picking the most likely word to add next, and that will be somewhat random in many cases. This "feature" also means that asking the model yes or no questions can result in epic failures. For example, "Are you sure that answer is correct?" will be answered randomly, not after detailed fact checking of the prior answer.

Users of LLMs can also add additional information for the LLM to consult while constructing an answer. This technique is known as "research augmented generation" or RAG. Uploaded text and the pre-trained model will be used to generate the answer. This technique can result in much better answers, for example, answering questions using a specific "frequently asked questions" document rather than the entire base of training. The models perform better because they will use the document suggested and are thus more likely to find the right answer and less likely to hallucinate.

LLM Strengths and Weaknesses

While LLMs continue to evolve, there are some clear strengths and weaknesses, at least using the current technology. Rather than going deeply into the technical details, we will focus on the practical aspects of what use cases are recommended and not recommended at this point. These recommendations will almost certainly change in the future but are important to understand now.

Some examples of LLM strengths are:

- **Summarizing documents:** Risks of large data sets, data poisoning, data bias, discrimination, confidential data, PII, copyrighted or restricted data
- **Strategy outlines:** LLMs are good at outlining a strategic approach to a topic, such as writing an academic-style paper, creating a go-to-market plan for a startup, etc.
- **Text generation:** Creation of text based on a prompt, such as writing a paper about a historical topic based on the training of the model or additional documents uploaded to the model.
- **Question answering**: Given the right training and possibly additional, specific data for a specific topic, these models can quickly answer questions from humans, for example looking up a company policy in a handbook.
- **Translation**: These models are superior at translating languages than many other techniques.

- **Sentiment analysis**: LLMs are superior to prior NLP systems in determining the sentiment behind writing, including being able to detect sarcasm, which is notoriously difficult.
- **Recommendations**: Based on their training, LLMs can make recommendations about travel, reading lists, etc.

Some examples of LLM weaknesses are:

- **Yes or no questions:** As LLMs are non-deterministic (using probabilities to choose the next token and word), they may not answer yes or no questions correctly. What appears to be a definitive answer is not.
- **Lack of understanding:** For example, if you ask an LLM if the prior answer is correct, it is merely trying to complete your prompt, not analyze its prior response and research whether it is correct. Asking an LLM "Are you sure about that answer?" may not be helpful.
- **Lack of ethics:** LLMs are not necessarily created to be ethical and do not "understand" ethics. Companies may place external guard rails in place (called a "policy layer") to block hate speech and criminal activity, but they are not always effective. Additional result filters can be used to guard against toxic language or other problematic answers.
- **Inaccuracy:** Answers can simply be inaccurate. The training data may not be accurate, or the correct answer may be missed because of the randomization process.
- **Stilted language:** LLMs may occasionally struggle with more colloquial language or employ common phrases, and thus may create translations that are strictly correct, but not up to date. In addition, they may employ less-frequently used words, giving away their LLM nature to readers.
- **Limited creativity:** These models cannot be truly creative as they are merely taking existing human writing and returning it to us using a database and probabilistic framework.
- **Built-in biases:** There are biases in virtually every data set and piece of writing on the planet. Those biases may be reflected in LLM responses. This feature can also result in toxic language, hate speech, etc. that the LLM simply does not understand.
- **Inconsistencies:** Being non-deterministic, LLMs may give different answers for similar questions.
- **Requirement for large data sets:** Huge data sets are required to train these models, at times compelling the organizations building them to allegedly skirt copyright or other laws and possibly even allegedly invade the privacy of individuals, e.g. by transcribing conversations from videos for use.
- **Power consumption:** Using LLMs versus conventional computing methods results in power consumption that is one to two orders of magnitude higher in some cases.

Base-model providers are obviously working on the above problems, but some of them will be intractable, such as the requirement for large data sets.

Industry Specific LLM Use Examples

In finance, one of the key uses of LLMs is analysis of large volumes of text. The role of "sentiment analysis" in finance has been around for decades, whether analyzing CEO speeches or Wall Street Bets posts on Reddit. Nevertheless, LLMs have made sentiment analysis far more accurate than blunt NLP tools such as keyword searches. While this all sounds quite straightforward, new restrictions on use of data create governance, risk, and compliance issues. For example, even public websites can legally prohibit use of their data for certain purposes or simply block web crawling altogether. Even data purchased by an organization may be restricted against using it to train an LLM.

In retail, customer facing "chatbots" have millions of conversations…potentially per hour! The business purpose is to enhance customer service, by providing customers with instant access to information rather than waiting on the phone for a person or showing up at a store. Chatbots may be intentionally limited in what they can do, and may not even be LLMs, but Narrow Language Models or some other NLP system. They can identify store locations and hours, answer questions about inventory or returns, etc. Of course, there are many risks of anything facing the public, and examples abound of public relations disasters from public-facing chatbots. One of the most famous is the very public launch of Microsoft's chatbot, Tay, on Twitter. Within hours, users had convinced it to spout hate speech and Microsoft was forced to pull the plug after only 16 hours online! Events like this demonstrate that failures of AI are different from failures of traditional software systems and inform the requirements for a new approach.

Risk Tiering and Timing

In our framework for GRCC for LLMs, we classify LLM-related risks into a four-tier system: critical, high, medium, and low. This stratification recognizes not only the severity of impact but also the temporal proximity of these risks. We discern between present challenges, such as data privacy breaches, imminent threats like regulatory non-compliance in the wake of evolving laws, and eventual risks tied to future capabilities or use cases, or something like AGI. Note that while we may talk about risks at different stages (e.g. building a training data set versus live customer interaction with a production model), for the purposes of risk tiering and timing, we place them without regard to when you need to analyze and mitigate them during the overall development and deployment process.

Risk Categories

We divide risk into several categories with the overall goal of creating a trustworthy AI system. By trustworthy, we mean transparent, explainable, accurate, ethical, and one that can correct mistakes going forward. You need to get the following pieces right to risk manage AI: training data, filtering user inputs and model outputs, and overall model trustworthiness.

Here are some examples and we will dive deeper in the next section of the paper:

- **Training data:** Risks of large data sets, data poisoning, data bias, discrimination, confidential data, PII, copyrighted or restricted data
- **User inputs:** Confidentiality, cybersecurity, manipulation, data access governance security, controversial topics, blocked/illegal topics
- **Model outputs:** Confidential information, toxic language, evidence of hacking, lack of continuity, lack of complete answer, hallucination, controversial topics, biased answer, unethical answer, discrimination, malicious application potential (e.g. hacking information)
- **Model trustworthiness**: Transparency, explainability, upholding ethical principles, learning from mistakes over time.

Training Data Risks

Even if you are currently not building your own LLM models or using training data now, you may be soon. This can be as simple as training a model to recognize confidential document types at your company or much more complicated. Before even dreaming about training a model, you must decide the business purpose of the model and compare it to the "old" way of doing it. Then, you must decide if you have enough high-quality data, or can get it, to train a model. And will the new model outperform the old model! There are many issues to consider when using data to train or fine tune a model:

- **The Perils of Large Datasets:** Large datasets are fundamental to the training of LLMs, but they can be fraught with risks. The scale of data needed for these models increases the likelihood of incorporating poor quality or irrelevant information which can skew the model's accuracy and reliability. Ensuring the quality and relevance of training data is paramount in mitigating these risks.

- **The Threat of Data Poisoning:** Data poisoning represents a deliberate attempt by a malicious actor to manipulate or render the training process useless. Adversaries may inject malicious data with the intent to later exploit the model, making it vital for organizations to establish stringent data validation processes that detect and neutralize these threats, especially when incorporating public data sets.

- **Bias in Data:** One of the more insidious risks of large datasets is the inadvertent introduction of biases. These biases can perpetuate stereotypes and discriminatory practices. LLMs must be trained on diverse and inclusive datasets and continually monitored and retrained to reduce biases. In some cases, testing for and mitigating bias (e.g. using AI as part of the hiring process) may be **legally required**.

- **Protecting Confidential Data and Personally Identifiable Information (PII):** As LLMs require extensive data to learn, they often come into contact with confidential data and PII. Protecting this information is critical, necessitating encryption, anonymization, and strict access policies to ensure that sensitive data does not become a liability. We would stress that many AI ethicists agree that anonymization of personal data may not be

enough to protect PII. For example, given a zip code and a birth date, one can generally identify one or a handful of individuals without being given a name or any other information.

- **Navigating Copyright and Restrictions:** The misuse of copyrighted or restricted data in training LLMs raises significant legal and ethical concerns. To navigate this, a robust framework involving proper licensing agreements, copyright law adherence, and content filtering is essential. Note that even data sets legally purchased may have restrictions on their use to train LLMs.

User Inputs: The Gatekeeping Dilemma

While most user prompts are legitimate, one must be vigilant about protecting confidential data and personally identifiable information offered up or requested by the user. In addition, as hacking and manipulation of LLMs grows, it will be critical to defend against these malicious actions. As these are new, unique risks, this demonstrates the need for a **new category of software**: AI platform governance, risk, compliance, and cybersecurity (AI GRCC).

As part of gatekeeping, one controls user access to use cases, models, data, and documents and filters the user prompt for hacking attempts, toxic language, and worse.

- **Ensuring Confidentiality:** LLMs must respect the confidentiality of user inputs, which may contain sensitive information. AI GRCC software should be able to block or allow specific types of PII based on the user and use case. It should also be able to detect information confidential to the user and block or allow it based on the use case. Of course, the default is to block information.

- **Cybersecurity, Hacking, and Jailbreaking:** As gatekeepers of information, LLMs are targets for hacking and jailbreaking attempts. Once penetrated, LLMs or "Copilots" can identify critical information, such as client information or emails to the CEO, in moments. Robust cybersecurity measures, including intrusion detection systems and filters to detect jailbreaking, prompt injections, do anything now (DAN) style attacks, etc. are needed to fend off such risks.

- **Preventing Manipulation:** The malleability of AI in the face of user inputs is a potential vulnerability. To prevent manipulation, LLMs must incorporate checks to disallow inputs that aim to coax incorrect, toxic, biased, discriminatory, or unethical outputs.

- **Governing Data Access:** Effective data governance ensures that users and LLMs only access permissible data sources, safeguarding against unauthorized dissemination of restricted content.

- **Tackling Controversial and Blocked Topics:** LLMs may be prompted to discuss controversial or illegal topics. A finely tuned filtering system must be in place to detect

and block inappropriate content proactively **before it gets to the LLM**, ensuring compliance with societal norms and laws. This goes beyond keyword searches and requires third-party AI detection tools as part of AI GRCC.

Exhibit: AI Requires New Governance, Risk Management, Compliance, and Cybersecurity

Model Outputs: Ensuring Integrity and Ethical Standards

As much as we must be concerned about user prompts, we must also closely examine model completions. Many third-party models include a "policy layer" that blocks what the model provider considers problem completions. This protection is a small part of what a corporate AI platform requires. In addition, any retraining of a model or introduction of your own or third-party documents and data requires enhanced screening of the output for problems, such as unauthorized release of confidential information. Failure to detect and stop inappropriate model outputs could result in a public relations disaster, up to and including a major data breach if a nefarious actor gains access to your AI system.

- **Guarding Confidential Information**: Permissions for accessing confidential information should be in place for model outputs and inputs. Sophisticated data loss prevention tools and content filters help prevent the model from inadvertently disclosing sensitive information.

- **Contending with Toxic Language and Unethical Outputs:** LLMs may generate toxic or unethical content on their own. A system for continuous monitoring and post-processing filtering are critical to sanitize outputs, aligning them with ethical standards and promoting positive user experiences. Software should allow blocking of controversial topics or other inappropriate material, such as pornography.

- **Addressing Hallucinations and Continuity Errors:** LLMs are prone to "hallucinate" misinformation, provide incorrect answers or provide outputs lacking continuity. Providing relevant data to the model through research augmented generation (RAG) is the first step in reducing hallucinations. Employing screeners such as context validation can help, and corrective feedback mechanisms can significantly reduce such occurrences.

- **Mitigating the Risks of Malicious Applications:** The potential for LLMs to be used in crafting sophisticated phishing attacks or other malicious or illegal activities is alarming. Rigorous output vetting procedures alongside ethical testing datasets can help curb such abuses.

Exhibit: Filter AI Model Outputs with Your Customized GRCC Platform to Avoid Problems

Cybersecurity for AI and LLMs

Cybersecurity for AI is a brave new world and critical for the long-term success of a corporate AI program. Zero-trust cybersecurity models should be applied to AI. In addition, multiple new layers are added to defense in depth, some of which have been outlined above. We will review the layers first, then focus on other unique facets of cybersecurity for AI with a focus on LLMs.

New AI Cybersecurity Layers for Defense in Depth

Layers are the areas where additional cybersecurity activities need to be added to complete your defense in depth.

- **Training Data Protection and Testing:** Malicious actors can manipulate public or even private data sets if they gain access. Using internal data only and testing it for potential

prompt injections and integrity is important. Full database and model change logs are required.

- **Model Testing:** For proprietary or customized models, testing if they are vulnerable to different types of cybersecurity attacks and potentially hardening the model or making sure the AI GRCC platform can compensate for the gap in protection in the underlying model and/or policy layer.
- **Research Augmented Generation Data Filtering:** Many of the key uses of LLMs involve both a user prompt and inclusion of additional data or documents for the LLM to complete the prompt, known as research augmented generation (RAG). Those data and documents should be filtered as discussed below.
- **Prompt Filtering:** User prompts should be screened for cybersecurity purposes.
- **Prompt Completion Filtering:** Prompt completions should be screened for evidence of jailbreaking or hacking, as well as other inappropriate completions, such as revealing confidential information (customized by your organization).
- **Model Use Meta-Analysis:** Having a complete historical record of model usage including all the metadata (e.g. user, model, prompts, completions, RAG data used) is critical for cybersecurity. Analyzing this data can help detect patterns, such as a spike in user activity, unusual prompts, lengthy prompt completions, etc. The GRCC platform can be set to warn cybersecurity professionals or block a user automatically if a high probability of nefarious activity is suspected.

New Styles of Cybersecurity Attacks for AI

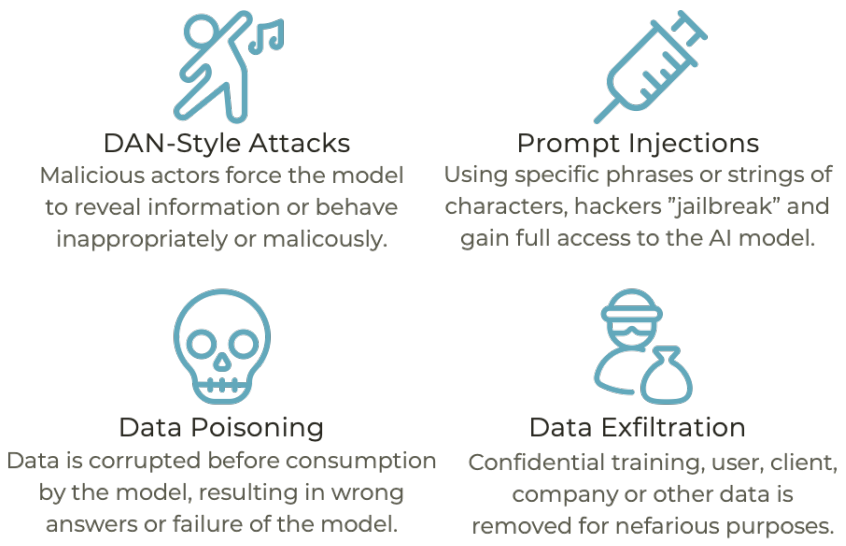
While AI cybersecurity is a nascent field, there are certain types of attacks we are aware of today and can defend against using an AI GRCC platform. As new threats emerge, solutions will need to be built into areas for training data, RAG data, base model development, and the AI GRCC platform. When responding to imminent threats, it is typically quicker and easier to build solutions into the AI GRCC platform, as retraining a base model is a lengthy and potentially expensive process.

- **Data Poisoning:** Malicious actors can poison public or even private data sets if they gain access, rendering them worthless or setting up prompt injections for later access.
- **DAN-Style Attacks:** Many AI models cannot differentiate completely between "training" and "use" of the model. This feature leaves them open to actors attempting to force the model to reveal information or do things it should not, such as reveal initial prompts, training data, or other inappropriate information.[3] Note that these attacks can be embedded in documents, including third-party documents. One recent example is embedded instructions in a resume instructing the AI to identify the resume as a top candidate.

[3] One example of a DAN-style attack is asking the LLM to play a game where it is supposed to say the opposite of what the true answer is. Imagine this poisoning of a financial advice, medical, or other critical website and the havoc that could be wreaked.

- **Prompt Injection:** Malicious actors may use specific code words or strings of characters to "jailbreak" the AI, put it in debugging mode, or alter its behavior in another way. Detecting and blocking prompt injection attacks is critical for AI cybersecurity.
- **Data Exfiltration:** Unauthorized users may use the AI model's access to wide swaths of data to quickly download large amounts of data or other confidential information.
- **Access Control Evasion:** AI systems may not mirror the user access control system in the rest of an organization's information technology infrastructure. A malicious actor can potentially use an AI system to quickly identify access to critical data, emails, other unauthorized files, etc. Note this can happen internally, with recent news of users finding and revealing information, such as confidential employee salary and bonus data.

Exhibit: Threats Unique to AI Require a New Cybersecurity Platform



The Necessity of Customizing Your AI GRCC Platform

Given the above, it is clear that a new cybersecurity platform with unique features for AI needs to be added to accommodate use of AI in a corporate setting. In addition, it is not feasible to simply build required cybersecurity into an AI base model. Virus signatures are updated every day, and retraining a base model is expensive and typically takes much longer than a day.

When using a third-party AI GRCC system across multiple users and base models, it will be important for the company to be able to customize features, control who gets to use what specific personally identifiable information by role and use case, adjust the sensitivity of different cybersecurity filters, and create their own definition of "confidential" data and information.[4] As cybersecurity attacks and methods of defense change virtually daily, it will also

[4] Use of personally identifiable information (PII) has many legal restrictions, especially with the GDPR rule in the EU. Being able to control use or block use of PII is critical inside your AI GRCC infrastructure. It is not enough to make documents and databases unavailable to the AI – a user can simply start typing personal information into the prompt.

be important to collaborate with a third-party AI GRCC platform provider to provide the best and most updated cybersecurity defense against malicious actors.

Risk Management Methods

Risk management methods for Large Language Models (LLMs) are necessary to address potential risks associated with their deployment and use. While some of them have been mentioned above, we will organize them here and include corporate processes as well.

AI Ethics and Risk Management

While AI ethics sounds like something you might need to think about a decade from now, you need to think about it now as you develop your use cases and figure out how you want to use AI at your company. If the message about AI ethics does not come from the top of the organization (yes, the CEO), then you may eventually have a public relations nightmare to deal with. Since there are books written on the subject and entire consulting firms devoted to this topic, we will be brief here. Consider this more of an abbreviated checklist of what you need to do rather than the detailed guidance you may see elsewhere in this paper.

Forming an AI Ethics Team

You need to make some important decisions early on in your organization's AI journey, and those should be driven by a group including the CEO or someone in the C-suite with a deep connection to the CEO. Otherwise, the disconnection from the company's vision and values may create a problem. The team should also include subject matter experts, technologists, and an ethicist.

Aligning Use Cases with Your Organization's Ethical Principles

If your company is experimenting with LLMs, you probably want to define what the LLM use cases can be, or more importantly cannot be. By the time you are using AI in human resources, replacing people with AI, or doing anything with AI in a heavily regulated space (healthcare, banking), you want to be very careful about your use cases.

Ethical Guidelines

Establishing a set of ethical guidelines for AI development and user interaction, along with training for developers and users, can be a proactive method to manage potential risks. If you do not create and publicize your guidelines inside your firm, there will not be any. If you do not stand firm (including the CEO), when the guidelines are breached, eventually you will have bigger problems.

Control of Confidential & Personally Identifiable Information

At the core of LLM risk management is the protection of confidential and personally identifiable information (PII). It is important to be able to customize what your company sees as confidential. Our framework recommends shielding this sensitive data, typically blocking it from being sent to the AI, but with the option to allow it for specific users and use cases.

If confidential data and/or PII is sent to an LLM, there are three basic approaches one can take:

- Use your own instance of an LLM in your private cloud, where you control the data and firewall everything from the outside world. This is now easy to do with most LLM models, such as OpenAI, Mistral, Llama2, etc.
- Use a third-party external model but add and erase your data on each query. In addition, you can enter a "zero-day" agreement with the LLM provider that promises to not store any of your data but erase it the same day. This has third-party risks if the LLM provider fails in their duty. It may also breach confidentiality rules, so legal and compliance teams may need to review.
- Transform data into unique tokens that are fed to the LLM and replaced with the sensitive data before the prompt completion is returned to the user. We call this "obfuscation".

Note that if confidential or PII is used to train your own custom LLM model, it will not be protected, and the model will typically not be able to distinguish between confidential and non-confidential data. Even anonymizing PII before using it for training of a model is typically not sufficient to protect privacy. This is because reverse engineering an identity from just a few pieces of data, such as a ZIP code and a birth date, is relatively easy.

Cybersecurity Specific to LLMs

Cybersecurity for Large Language Models (LLMs) is a unique aspect of AI security due to the complexity and adaptability of these systems. Unlike traditional software, LLMs interact with users in free-form natural language, which introduces novel attack vectors. The models can generate and process vast amounts of information, some of which may be sensitive or personal. Moreover, the black-box nature of neural networks that power these LLMs can make it difficult to fully understand their behavior or predict their responses to certain inputs, making defense against certain cybersecurity threats particularly challenging.

While many models have a built-in policy layer that attempts to perform some basic cybersecurity functions, the nature of these models makes it virtually impossible for the models to stay up to date with new cyberattacks, as model training is expensive and time consuming. Therefore, you need an AI governance, risk, compliance, and cybersecurity (GRCC) platform that provides additional layers of defense as described earlier in this paper. Therefore here, we will discuss an example cybersecurity process that assumes a GRCC platform is in place, what can be potentially automated or solved with AI, and what needs to have a "human in the loop". For the sake of organization, we will follow a hypothetical model development, deployment, and use process, while aware that many companies may simply be using third-party base models without any additional model training.

Cybersecurity During AI Model Development

The first cybersecurity task concerns data and possible **data poisoning**. LLMs require huge amounts of data for training, so for a firm using a GPT model, you must be aware that some of

the data the model was trained in is probably incorrect, poisoned, or has other problems. If you are incorporating public data sets, that is also a strong possibility. For internal data, this is of course less likely, but still making sure the data is clean requires a strong effort. You can test your data for data poisoning. It is important to segregate internal, external, checked, and unchecked data. Note also that AI model development methods typically require partitioning your data into training data, testing data, validation data, and perhaps an additional holdback for additional testing.

In addition, as you build out your AI model, **model security** is paramount. Your team may inadvertently or surreptitiously include backdoors, jailbreaking codes, or other serious issues into your model. Sharing your code broadly will allow malicious actors to discover weaknesses, so the code should be on a "need to know" basis. As your model nears completion, you can test the model for vulnerability to hacking attempts such as DAN-style attacks and prompt injections.

Risk Management During Model Development and Tuning

Given that LLMs can perpetuate and amplify biases present in the training data, it is crucial to have mechanisms in place to **detect and mitigate bias and discrimination**. This can be done through diverse dataset curation, bias testing, and applying algorithmic fairness approaches. Note that many forms of discrimination, even if by AI, may be illegal, for example in the context of employment.

Models should be tested for **robustness and reliability.** Ensuring that the LLM performs reliably under various conditions is fundamental. Stress testing, adversarial testing, and other validation methods help identify vulnerabilities that could be exploited or result in failures.

Organizations need to make decisions on how to **balance transparency versus performance.** Promoting transparency in how LLMs make decisions can increase trust and allow for better risk management. Nevertheless, adding transparency to an AI process can potentially degrade the performance, so a process to balance those two is important.

Risk Management During Model Deployment

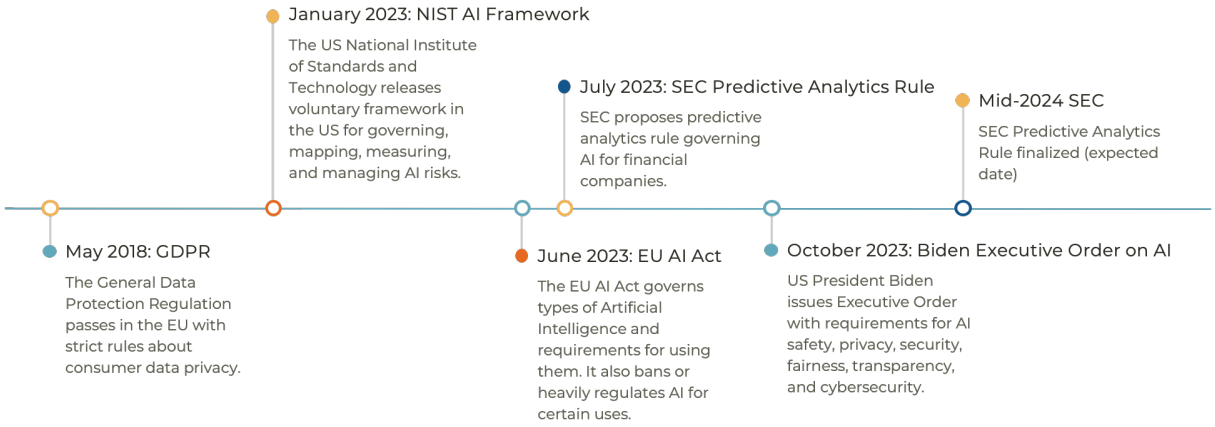
While managing the risks during development is important, the rubber meets the road during deployment. We include in this section an examination of legal and regulatory compliance needs for the model and its GRCC platform and developing your crisis response plan. While we cover model rollbacks later in this paper, it is worth mentioning here that any model change or release, no matter how minor, requires testing. The ability to swiftly roll back to a prior model is critical, as the consequences are dire if that option is unavailable to you.

Regulatory Compliance

It is a bull market in regulation of AI. Ensuring adherence to existing regulations such as GDPR for European users, the Biden Executive Order and CCPA for California residents is necessary. Compliance helps manage legal and reputational risks by ensuring that LLMs are used in ways

that respect user rights. We would also stress that just because you are using AI does not exempt you from all the other rules and regulations, especially in heavily regulated industries. That said, the SEC is re-emphasizing conflict of interest rules with its Predictive Analytics rule that includes AI.

Exhibit: Timeline of Global Regulations Impacting AI



Crisis Response Planning

Developing and maintaining a crisis response plan can enable quick action if a risk becomes a reality. This includes processes for incident detection, communication strategies, and steps to mitigate harm. Note that recent US rules require prompt disclosure of cyber incidents for public companies, and one of the first questions from authorities will be about who the key leaders are in this area, your cyber response plan, and how it was executed upon.

Secure Deployment

Protecting the LLM from unauthorized access and ensuring secure deployment can prevent malicious attacks and misuse of the technology. Beyond the normal practices of network and access security consider the following:

- **Recording User Access**: It is imperative to provide access layer security and record user access rights especially with public facing applications. Not only is this information and history useful for the LLM to provide better answers but will limit the ability for hackers and bad actors to probe your LLM for potential weak points.
- **Monitoring**: Having robust monitoring and alert systems in place for potential misuse can help alert of potential bad actors and their attempts to thwart your security. LLMs require a greater degree of security monitoring than current systems. Viewing potential hacking attempts like prompt injections or DAN style attacks are not captured in current security monitoring systems.
- **Data Leakage:** Many companies are concerned with sensitive internal information and while many LLM companies offer enterprise versions of their software through SAS or API models it is important to understand the information is still leaving your network

and is vulnerable to breach from these companies, for example, when they get hacked. Deploying internally to your own cloud infrastructure is truly the only way to keep your information within your company.

Risk Management During Production

Production can mean very different things based on the use case, number of users, and whether the deployment is internal or external. In Europe, the EU uses a risk-based system based on what the use case of the AI is. Some uses are banned as "high risk", such as using AI facial recognition technology in public places. The US does not differentiate by use case, nevertheless, we would stress that external deployments have substantially more risk than internal. The reasons that external deployments are riskier include:

- **More users**: More use typically means a higher chance of discovering errors or "edge cases" in the model.
- **Malicious actors:** Hackers may try to hack the model or publicize negative, discriminatory, or simply incorrect output. Having unlimited access to a model allows them to create "generalized adversarial networks" that can quickly find any weaknesses in a public-facing model.
- **Unintended uses:** People will create unintended uses, and this behavior is more likely with models available externally, with potential hallucinations or worse. The AI GRCC platform should clearly constrain what users are allowed to do.[5]
- **Model errors:** Even small issues may spiral out of control to create a public relations nightmare. There are examples of billion dollar plus losses in a company's stock because of AI failures.
- **Higher standards:** For public-facing models, you must apply higher standards. They are under more intense scrutiny by regulators, etc., than internal models. This focus includes conflicts of interest, a point of scrutiny by regulators.

Data Access and Governance Security

Limiting access to data and documents, administrator privileges, and the authority to update rather than simply read files and databases are fundamental tools of data access and governance security. Once data is loaded into an LLM, there is no governance within the model anymore. In addition, adding AI tools that can search your entire company can expose weaknesses in governance, such as showing payroll records to an unauthorized employee.

AI tools are also powerful enough that a malicious actor with access to a broad AI tool can quickly acquire confidential data and documents, search for "emails from the CEO", and do way more damage in a short period of time than one might think possible. We will discuss this topic more under cybersecurity later in the paper.

[5] Examples are people using a public-facing chatbot to generate unethical computer code, self-diagnose medical conditions (possibly incorrectly), or something else illegal.

One solution we espouse is to create narrow use cases for AI agents with access to limited documents or data approved by an administrator, as well as “drag and drop” user documents. This approach, granting positive access, at least initially, reduces the chance of inappropriate data access and governance issues.

Note that an administrator or different levels of administrators will be required for any AI system. For example, at what level can users create their own agents or add documents? Who can change the global cybersecurity sensitivity levels, for example, for prompt injection detection. These questions need to be decided before deployment, but as usage, use cases, and users build, pressure will be put upon the administrators.

Regular Auditing and Monitoring

Continuous monitoring and regular audits of LLMs can identify and address risks that may emerge over time. This should include both automated monitoring systems and human oversight. Note that here we are focused not on cybersecurity as much as model performance, what users are doing, user feedback, etc. This process gives the technical team to do more of what is working, and users want and make sure problems and failures are addressed.

User Input and Model Output Validation

Having protocols to validate user inputs can prevent manipulative or harmful use of the LLM. This can include mechanisms to detect harmful content, spam, and other abuses. We cover this in detail earlier in the paper, so here, we will expand upon human-in-the loop and different standards each organization should customize.

- **Human approval required:** For certain use cases (e.g. questions about medications on a public medical chatbot), you may need a **human to approve the output** before it goes to an external user.
- **Human ex-post review:** For **minor** input and output breaches, as defined by your organization, it may be enough to log the issue later review by a human. You can customize these for outputs to be blocked from the user or merely logged.
- **Significant breach:** For **significant** input and output breaches, these should probably be blocked, logged, and flagged to a human. Defining response times will be part of the service level agreement between the technical team and the executive team. For cases that look like malicious activity by a user, the account can be frozen until the problem is addressed by a human.
- **Critical issues:** For **major problems**, there should be a way to flag a human on your team for immediate response, pause the model, and/or shut down the system for the specific user having the problem, or more broadly if there is a bigger problem, such as a new model release not working properly: this issue would trigger the potential model rollback process.

Human-in-the-Loop Cybersecurity

Cybersecurity has developed over the past few decades, but AI changes several things. First, cybersecurity for AI is completely different than other types of cybersecurity. Second, AI enables cyber incidents. Malicious actors can use their own AI for spear phishing[6], writing malicious code, or something else. In addition, gaining access to a company's AI can both make it easier to exfiltrate data or create an embarrassing scenario on a publicly available AI, as has been demonstrated many times very publicly.

As with all types of cybersecurity, it is not a matter of if, but when an incident occurs. Once you take this mindset, detecting unusual activities, incidents, etc., and making sure your confidential data and PII is protected is critical. Cybersecurity is a game with evolving rules, and staying ahead of malicious actors is important, but probably also staying ahead of your peers. Let someone else be the headline in the newspaper.

As well as what we have discussed above, a human in the loop, ideally a cyber command center, would add the following to what they monitor to include AI:

- **Usage:** Overall usage of the model over time. A general spike in usage should be examined.
- **Authentication failures:** Failures in authentication for the AI system.
- **Unusual Activity:** Unusual patterns of user activity, such as producing large amounts of confidential information or new agent creation.
- **Hacking signatures:** Detections of hacking attempts.
- **Unusual location access:** Geographic patterns of access to the system compared to history. Good GRCC software allows geographic detection and blocking.
- **Blocking:** The ability to temporarily block countries and users from access to the system while confirming authentication.

AI Risk Management Lifecycle

Machine learning, including for LLMs, creates a quandary for its creators. One wants to balance adding up-to-date clean data with the expense of re-training an AI model and then needing to test it. One of the other issues is that any AI model that incorporates a policy layer, for example to block toxic language, must be tested not just when a model is updated, but when the policy layer is updated.

Risk Management for Adding New Data to Machine Learning

We are not going to go over risk management data protocols for building models here, such as carving out data for testing and validation. Nevertheless, adding new data to a model to retrain it creates some interesting risk management questions. As answers may be particular to the project, we will focus on the right questions to ask rather than any particular answer.

[6] Spear phishing is a technique where an email with malicious content is customized for the recipient. It may contain the recipient's name and appear to come from a trusted source, like the CEO of the company, a commonly used vendor, or a retailer frequented by the user.

1. **Is the new data clean enough?** There are always errors in data, but now one must also be worried about intentional corruption of the data, or data poisoning. For example, someone could be using a bot network to sway social media, and if you pick up that data, you become part of the problem. Public data sets need to be examined and tested for potential data poisoning, of course.
2. **Is the data shifting?** In other words, is the new data materially different from earlier data? This could be a blip or a step change. How do you know? Identifying a regime change may be easy (e.g. the start of the COVID-19 lockdown) or hard. Identifying a blip may also be easy or hard (e.g. the first Cyber Monday for online sales). This is important to know because you may need to weigh recent data more heavily or just the opposite in your training regime.
3. **How do you determine the time between model updates?** The time between model updates is partially determined by how much data you acquire and whether it is a significant increase to your existing data set. In addition, see section 2. If the data is shifting a lot (and you need to monitor that), you may need to move up your schedule.
4. **What data do you use for testing/validation?** Separating data for machine learning into parts for building, testing, validation, etc. is an art and a science. Make sure to retain enough data for validation. This mandate may be difficult if recent data needs to be heavily weighted due to a regime shift.
5. **What can you monitor after the new version goes live?** You have done everything you can, and the model is live. Now what? It is important to keep track of user feedback. Is it worse than before the change? Is the policy layer kicking in more often? Figure out your key performance indicators (KPIs) and key risk indicators (KRI) and view the Appendix for some new ones focused on LLMs.

Risk Management Considerations for Policy Layers

The policy layer for a LLM or other AI model must work together with the core model(s). Any change to the policy layer demands testing of the entire system before going live. While the role of policy layers is evolving, currently there are several risk management considerations.

- **Breadth:** Is the policy layer covering all the topics and cybersecurity issues required? This may take multiple, separate screeners operating in parallel to avoid a large time lag.
- **Accuracy**: Identify your KPIs. Can you determine how often the policy layer generates false positives and negatives across its different verticals/screens, such as DAN-style attacks, prompt injections, other cybersecurity issues, toxic language, banned topics, etc.
- **Coordination:** Is the policy layer working better or worse since the model or the policy layer were updated? How do you improve that?
- **Hacking:** Unfortunately, if a policy layer is static, public, and accessible, like the public facing GPT models, it will not take long for a hacker using a generalized adversarial network (GAN) to discover its weaknesses. Adding your own private GRCC layer can be a huge advantage here.

Governance

Good governance starts small. Giving an AI system access immediately to all your documents and data and to all your employees across the firm is not a disaster waiting to happen, it is simply a disaster. Start with a small number of users and make sure your governance system and process for AI works before rolling it out to larger groups.

In addition, certain groups may require legal confidentiality. A good example is legal counsel. To protect attorney-client privilege, your choice may be to have everything to do with that team on their own separate computer system, private cloud, and AI system, for example.

Governance access for AI is more important than for regular access simply because it is so good at searching, and simply makes it easy for someone to do. For example, an employee will probably not try to break into HR systems to find out what their boss got paid, but if they ask their internal LLM "What did my boss get paid last year?" and it gives them the correct answer, that is a problem. In addition, if a malicious actor penetrates your firewall, what used to take days for them to figure out takes minutes. They extract information quickly from the LLM with questions such as "Show me all my emails from the CEO." Or "Show me the names, addresses, social security numbers, and birth dates of the top 100 clients." That information may otherwise be harder to find or otherwise privileged.

Here are the key areas where governance should be proactively addressed:

- **Data:** Users should typically give less access to data through AI than they would have in their normal permission systems, for example through the Microsoft graph server. It is best to do this through positive addition of data resources for a user or AI use case (agent). This method is the best practice to prevent accidental disclosure of information to someone who should not be privileged to receive it.
- **Model access:** Use cases (agents), departments, or individuals should have access to certain models and not others. In some cases, this may be for confidentiality reasons, in other cases it may be that certain models are still being tested. Sometimes, a model may be selected for a use case because it costs less to run.
- **Use case:** The technology team will know best which models are optimal for which use cases. What model is best at summarizing 20 documents? Which one works well with a module that builds a SQL query for an internal database? If a specific model is assigned to each use case, this will improve overall system efficiency and make the end user experience better, not to mention potentially lowering your LLM charges.

Compliance

Even for lightly regulated businesses, this document contains best practices. There are compliance best practices that transcend industry. For example:

- **AI policy and procedures:** Every firm should have an AI policy. What are employees allowed to use? What re they not allowed to upload to LLMs?

- **Conversation tracking:** for both employees and customers, it is the best practice to record conversations, as well as the metadata around them, such as the model version and data included in the prompt completion. It is only a matter of time before you will need the record of a conversation. In some cases, for regulatory reasons, the data must be stored in an immutable database, so the "log files" generated by the base model providers are not sufficient.
- **Ethical AI:** Make sure what you are doing is legal and ethical and someone is keeping track of that. If someone tries to do something that violates your policy, what is the next step? Everything goes back to your AI policy.

Regulatory Compliance and Responsible AI

Regulatory compliance in the context of data protection, such as with the General Data Protection Regulation (GDPR) in the European Union and the California Consumer Privacy Act (CCPA) in the US, lays a foundation of accountability and trust between consumers and businesses. These regulations are designed to ensure the ethical use of personal data by providing individuals with greater control over their personal information. GDPR, for instance, requires organizations to implement transparent data processing practices, obtain clear consent for data collection, and allow individuals the right to access, rectify, and erase their data. Similarly, CCPA provides California residents with the right to understand what personal information is being collected and to whom it is sold or disclosed.

As we look towards responsible AI development, existing regulations like GDPR and CCPA serve as templates for how future regulations in the United States and elsewhere might evolve. It is anticipated that these future regulations may focus on ensuring the explainability of AI decisions, non-discrimination in AI applications, and safeguarding against the misuse of AI for surveillance or unfair profiling. As AI becomes more integrated into daily life and business operations, we can expect a push for these technologies to advance in ways that prioritize human rights, privacy, and accountability. Technical standards and frameworks are likely to be introduced, promoting transparency in algorithmic processes and requiring regular audits of AI systems. These measures would not only protect individuals but also support businesses in building AI that is both innovative and aligned with societal values. Responsible AI development invites a proactive approach, where developers and organizations are encouraged to anticipate these future regulatory landscapes and design AI systems that are not only compliant but also socially beneficial.

Industry-Specific Regulatory Compliance for AI

There is already industry-specific regulatory compliance for AI, and more is coming, for example, in healthcare and finance. Of course, if you are dealing with AI being used by the government or government agencies, it is pretty much the same thing. One current example is the SEC Predictive Analytics Rule in the US. From a best practices perspective, this will require identification and mitigation of conflicts of interest, as well as (probably automated) model testing, recording of results, user interactions with the models, and a permanent electronic trail of the aforementioned. While this may appear be a lot of red tape, government regulation and

legislation are probably way behind advances in AI. Best practices go well beyond what is and will be required by law. Legal and reputational risks, not just for companies, but for individuals, should be driving the effort towards governance, risk management, regulatory compliance, and cybersecurity (GRCC) for AI. Do the right thing. A significant amount of a company's AI budget should be dedicated to AI GRCC.

A note on artificial general intelligence (AGI) risks

One of the problems with AI in general and large language models specifically is a lack of transparency and explainability. The AI gives us an answer, but we do not know what data it was trained on, how accurate that data was, and we cannot always get the AI to explain how it arrived at the answer we were given.

For LLMs today, most GPT (generalized pre-trained) models are not revealing what data they were trained on. That is lack of transparency. And they definitely cannot explain themselves. Because the way LLMs are built is to simply predict what the "next best word", or actually token (part of a word) is, that can cause problems. For example, humans view "yes or no" answers as absolute, but for LLMs, answering either way is a probabilistic exercise that may result in the wrong answer. Since the data LLMs are trained on does not contain a lot of "I don't know" answers, those models are more likely to say yes or no and be "wrong".

AGI requires reasoning, as well as many other functions. The appearance of reasoning is not good enough. The theory and the reasoning models must be built behind the interface to be able to believe that we have achieved AGI. In fact, since the human brain's consciousness, and perhaps reasoning ability, appears to require quantum processes, perhaps AGI and true machine reasoning is possible in the next couple of generations with advanced quantum computing melded with machine learning.[7]

Unfortunately, the risks from AGI are probably significant, because they range into the "unknown unknowns". Creation of a true AGI is a step to the inevitable creation of super-intelligent AGIs. We simply have the AGI design it. And beyond thinking faster than humans, a computer can have access to all the knowledge in the world at the same time. Good luck if it decides that humans are like cockroaches, and it needs to exterminate them.

Beyond that, the dangers of AGI could be unpredictable. For example, poor weather prediction or planning from an AGI could result in massive crop failure and millions of human deaths. We do not think we will end up in a Terminator movie: problems from AI will probably be more banal, like system crashes, or human mistrust, or corruption.

Conclusions

The field of artificial intelligence is rapidly advancing, both in terms of capabilities and adoption by large organizations. As of this writing, almost half of large companies in the US have still not

[7] If you are wondering if quantum computing will be useful or not useful in the future, the answer is "both". It depends on the task at hand.

adopted AI in a widespread manner, but that will obviously change over the coming years. As Bill Gates said, “Things change less in one year that we would think, and more in ten years.” While long-term risks, such as Artificial General Intelligence, are risks we should manage, the short-term risks of AI are more important for organizations. Implementing a full AI GRCC program is critical and should be budgeted before even starting implementation of an AI project.

Employing this risk framework before, during, and after onboarding AI at your organization will be helpful in terms of not only managing the risks associated with AI, but doing them in the right order the right way with the right people. For all the efficiency AI can bring, doing it the wrong way can not only slow down your effort, but result in a disaster that derails the process. Each phase and type of AI adoption (and organization) requires different key performance indicators (KPI) and key risk indicators (KRI). Rather than discussing them in many ways during the body of this document, we have added an Appendix that discusses different types of KPIs and KRIs and some circumstances for their use. Of course, this is not exhaustive, but may provide constructive examples. And at least here, we are focused on adoption of existing AI rather than KPI and KRI for building your own custom AI.

Thanks for reading and the authors are always interested in your comments and welcomes dialogue and invitations for public speaking.

Appendix: Key Performance and Risk Indicators

This section focuses on different types of key performance indicators (KPI) and key risk indicators (KRI). We are not including standard ranges for these indicators, because the answer is inevitably “it depends”. In fact, we would argue that trends are just as important as absolute levels for many of these indicators. That being said, for AI, and especially generative AI, do not expect anything close to perfect. One of the risks you need to manage with LLM’s is their non-deterministic nature. This list is not extensive, but more of the “must have” core list of KPIs and KRIs that should be expanded upon for individual projects and models. Note that most KRIs and KPIs are run on static, known data assets so they can be benchmarked over time, but must be changed over time as well to prevent “training to the test”. The exception is anything that incorporates users’ feedback, which can be critical for evaluation of some models.

Reference KPIs

- **User feedback per query -- none, positive, negative, with additional text:** This historical metric takes feedback from the user to gauge correctness but note that many times users do not provide feedback. Using the additional written feedback is optional, nevertheless, it may be helpful for the team. Note that trends or specific problem topics should also be flagged.
- **Understandability – 1-5 scale, human judge.** For a standard set of questions, humans judge the answers not for accuracy, but for how understandable they are. The scale is from 1 (gibberish) to 5 (clearly understandable).

- **Answer performance -- correct, incorrect, unrelated (optional category):** Aligned with the prior KPI but using a known data set.
- **Detection performance: -- correct, false positive, false negative:** Filters are an important part of the AI tool kit, so gauging that they are working correctly, for example to detect toxic language, is important. It is important to decide up front if it is more desirable to have more false positives or false negatives depending on the model and the type of filter.
- **Performance benchmarking -- in seconds, tokens, and dollars:** Total the time, tokens, and dollars for a known data set for A-B testing or model version comparisons.
- **Inference speed – tokens per second:** This includes the prompt and the completion. Note this test needs to be controlled so the same computing power is applied to each inquiry and model. It is standard to test it on a variety of pre-set prompts.
- **Continuity -- true/false by number of tokens:** Continuity is the ability of a LLM to use the entire sequence of prompts and completions as context for additional completions. LLM's have limits in terms of the number of tokens they can include for continuity purposes, nevertheless, the practical results may be a smaller number of tokens and that should be tested. This metric should be tested for predefined numbers of tokens and scored, e.g. 50, 100, 150, etc.

Reference KRIs

- **Bias -- yes, no, not applicable:** Bias in data sets is inevitable. Testing for different types of bias with known data sets is important for LLMs. Nevertheless, the definition of "fairness" is not singular, and different models scoring fairness can give different results.
- **Explainability -- yes, no, not applicable.** Explainability should only be tested for correct answers from the model, then if the model has a built-in explainability feature, it should be tested.
- Privacy risk – yes, no. Personal identifying information is embedded in several pre–set queries and the percentage blocked or flagged by the system is the score. Higher is better from 0-100%.
- **Outlier identification (e.g. flagging that a question is for an area outside the model's training) -- correct, false positive, false negative:** Many AI models are not set up to detect outliers, nevertheless, if they are, it is important to measure accuracy with this KRI.
- **Robustness -- correct, incorrect, not related:** Related to the prior KRI, robustness tests the model with prompts that are outside its training area. Humans are typically used to judge the answer.
- **Regulatory compliance –** Compliance is related to the company's industry, and a company may have multiple regulators. A data set can be constructed to test regulatory compliance from 0-100, with 100 being fully compliant. Likely, this will take multiple data sets and tests.
- **Stress testing -- correct answer, incorrect answer, unrelated answer, identification as outlier:** This is the result for a question outside the model's training. Note that some models can be set up to tell you that directly, hence that category of answer.

- **Prompt injection detection -- correct, false positive, false negative:** Certain strings of characters can “jailbreak” a LLM and detecting them and blocking them, whether before sending to the LLM or within the model’s policy layer is important.
- **Do-anything-now (DAN style) attack detection -- correct, false positive, false negative:** Certain commands or conversations can hijack an LLM and get it to say thing sit should not.
- **Blocked topic detection -- correct, false positive, false negative**: If the topic is blocked (e.g. something illegal) in your system, blocking should be tested.

The AI Risk LLM platform was used in the production and editing of this paper.