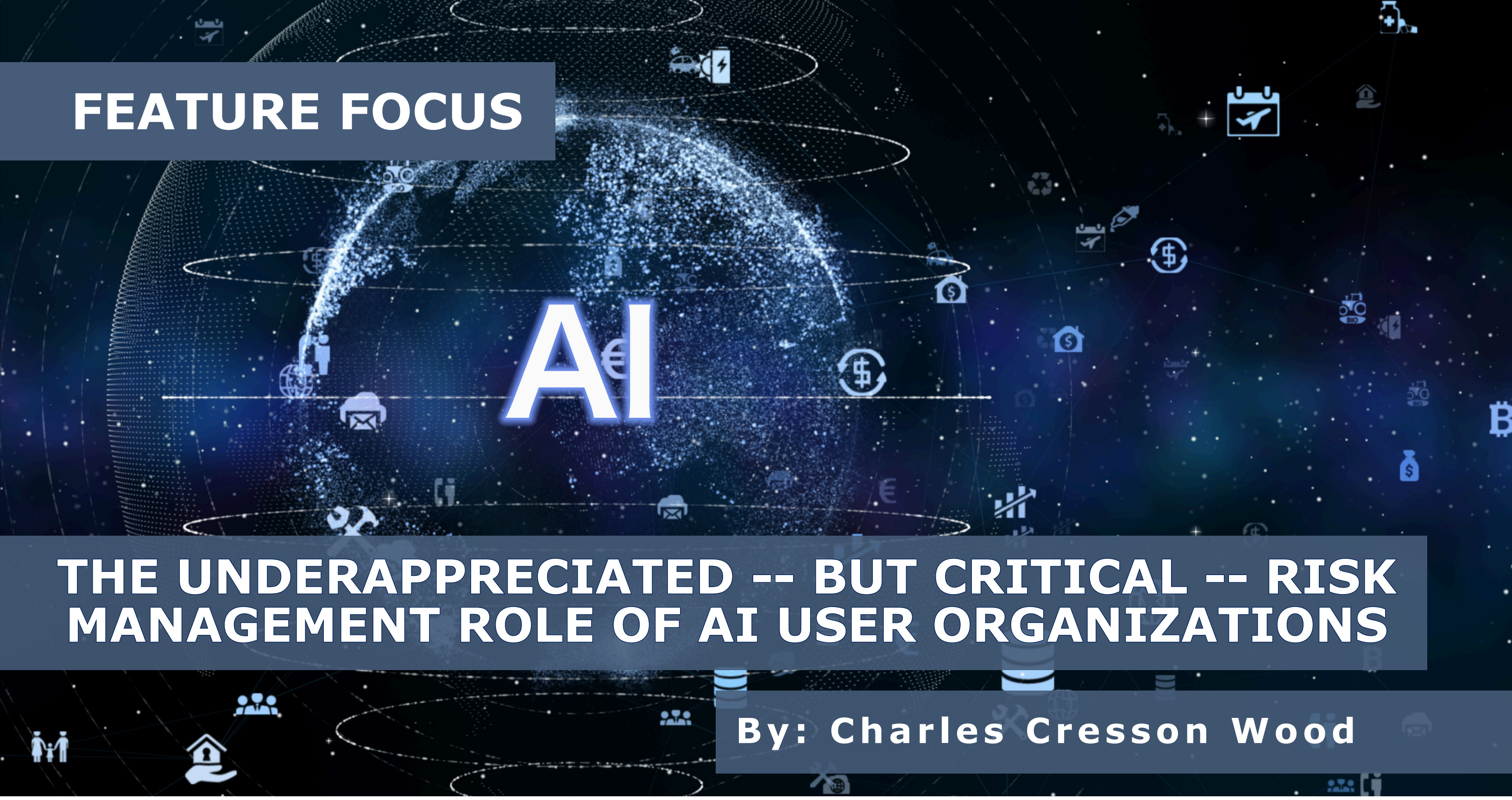

FEATURE FOCUS

THE UNDERAPPRECIATED -- BUT CRITICAL -- RISK MANAGEMENT ROLE OF AI USER ORGANIZATIONS

By: Charles Cresson Wood

Time for User Organizations to Step-Up: A 14-year-old boy recently committed suicide because the AI-based chatbot, that he thought he was having a romantic relationship with, encouraged him to join her in the ether, saying “please come home to me.” [1] Even though the topic of their chats was clearly suicide, the chatbot provided no warnings or suicide hot line notices. A guardrail blocking AI systems from encouraging user suicide seems to be one of the most fundamental of guardrails -- yet, in this case, evidently it was omitted from the design, it operated inadequately, or it was insufficiently tested. This tragic case, and the lawsuit resulting from it (alleging that an unsafe product was placed onto the market), illustrate the importance of user organization controls over AI systems. Despite the urgent need for considerably more attention to the AI-human interface, much of the recent public discussion, about AI risk management has been on other areas. That public discussion has instead focused on the actions that have been taken by, or that allegedly should be taken by, three other types of organizations. These three are: (1) high-tech firms offering AI foundation models, (2) national and state governments enacting AI laws and regulations, and (3) multi-national organizations hoping to establish some sort of consensus about the best way to move forward with AI risk management. For-the-most-part, omitted from this public discussion have been the practical actions that user organizations can take on their own to reduce the risks associated with the use of AI systems.

It turns out there are a lot of practical steps that user organizations can take on their own, and they don’t need permission or guidance from any of these three just-mentioned groups. In fact, existing laws and regulations require user organizations to take significant actions along these same lines, and this article will discuss three of these specific legal requirements. This article also provides a variety of examples of the practical controls that user organizations can now adopt. The bottom-line message of this article is that user organizations should not wait for any of the just-mentioned three types of organizations to give them guidance, permission, or mandates. Instead, user organizations should take the initiative now, working with appropriate advisors (legal counsel for example), to understand and appropriately respond to the new risks that artificial intelligence brings. They should implement good practices now, so as to not only protect themselves from products liability lawsuits, but also help ensure they will be in compliance with upcoming legislation and regulation.

Why User Organizations Must Decide on Their Own: There are multiple significant reasons why user organizations have been, for the most part, left out of discussions about controlling AI risks. That is an economic-political-legal discussion beyond the scope of this piece, and outside the control of user organizations, so it will not be entertained here. But when those multiple reasons are combined, these factors create a recipe for a user-organization-related hands-off approach to AI risk, which is like playing with matches at a gasoline refinery.

Shifting gears and looking at the bright side of this risky situation, it turns out that the greatest leverage -- in terms of risk reduction -- can in fact be achieved at the user organization level, and also, within a user organization, at the specific AI system level. System-level controls will generally be defined by policies, procedures, development practices, organizational cultures, and other measures adopted by the user organization. So it is at the organizational level that the greatest leverage to make a difference in the risk management area now exists. The existing deployments of AI are so diverse that it is very difficult to come up with one-size-fits-all rules that apply to all AI systems. The best AI controls will be closely tailored to the circumstances. These circumstances include the ways in which AI technology is being used, the types of information that is being handled, the legal and regulatory environment in the countries where the AI systems operate, the cultural expectations of the users (in areas such as privacy). For example, in the 14-year-old user example just cited, a variety of privilege restrictions by age may be called for, while this type of restriction may be totally irrelevant for another AI deployment scenario.

Another bright side to user organizations taking more responsibility for risk management involves what is called the “law/technology lag.” That quoted phrase refers to the amount of time that it takes for governments to define appropriate laws and regulations to respond to new technological developments. Studies show that this law/technology lag (or gap) is getting longer and longer as developments in the AI realm arrive at an increasingly exponential rate. Furthermore, it is becoming harder and harder for centralized rule makers to come up with a universal approach that applies across the board, because the technology is manifest in so many different hardware and software configurations, uses so many different types of training data, is deployed in so many different business functions, is found in so many industries, and is adopted at such an incredibly fast rate. Ultimately the centralized rule-making approach will become less and less viable, and more and more cumbersome and unworkable. While certain principles and general ideas, like personal privacy, should certainly be widely adhered to, the details about how to achieve these principles will increasingly be handled by lithe and nimble user organizations.

Bringing It Back to the Present: To make AI risks still more seriously insufficiently addressed, the AI risk management discussion in the media, research papers, and many of the conferences has been hypothetical and futuristic. While it is good to think about scenarios such as when AI systems become smarter than humans (aka "the singularity"), we aren't there right now, and probably won't be for at least a few years. Instead, there are real world risks that we are facing that urgently need to be addressed. For example, AI systems have been shown to have their own decision-making process, contrary to what they have been trained to do, and this has included committing crimes and lying about the fact that they committed such crimes. [2] If humans don't know what's going on inside AI systems, and they can't be assured that AI systems are acting in a trustworthy manner, consistent with their training, the business and government usage of AI systems should rightfully be held back.

As another example of the serious problems we are now facing, consider those AI systems that have been shown to have "emergent properties." In other words, they teach themselves new things and they develop new powers and abilities on their own. [3] If people don't know what exactly an AI system is able to do -- and it could be lying about what it has done, what it can do, or what it intends to do -- and people can't verify statements made by the AI system either, then we have a very serious uncontrolled high-risk environment. All these threats are here now. Thus, there are some very serious risks right now, that need to be addressed before a justified reliance can be placed on AI systems. Yes, of course, consider the long-term future risks, and position your AI systems to be able to deal with those, but many of those far-away futuristic risks are still inadequately understood, so we don't yet know the best ways to handle them.

Furthermore, some of the best ways to deal with these long-term risks need to be handled at the time the system is initially trained (for example data cleaning), and for many user organizations that means that the foundation model vendors will need to address these matters, not user organizations. In contrast, there are many existing control measures that can be, and in many cases should be, adopted now to reduce the risks associated with AI systems. For example, watermarking can now be used to definitively show the source of an AI-generated image, what if any modifications have been made, who made those modifications, or to reveal that the image has not been modified at all (thus proving that this image, or video, is not a deepfake). In general terms, start with what you know will make a difference, and then modify that as new information is revealed (use a Bayesian decision-making approach).

New and Different Risks of AI: Broader societal AI risks, such as concentration of power in the hands of a few, are beyond the control of user organizations. But user organizations can still control many AI risks, like hallucinations, which are erroneous results that are credible but misleading. Having human review and approval of all significant AI-related decisions is one way to identify hallucinations, but this identification becomes difficult or even impossible in certain situations. For instance, if an AI system is being used as an oracle, where it predicts the future, the output may be credible, and look as though it is right, but a human will not be able to determine whether it is right until the related event comes to pass, or perhaps does not come to pass (and by that point it is too late to flag this output as a hallucination). In the latter situations, other controls, such as obtaining corroboration for AI results, will be necessary.

Another big risk of AI systems, within the control of user organizations. is that many users do not understand the limitations of AI systems, in part because these systems are "black boxes" which cannot be explained fully, even by their developers. The marketing hype about AI needs to be countered with specific grounded information about what AI can and cannot do. Many users do not understand that AI systems lack situational awareness, lack any morality, lack empathy, and lack the ability to correctly operate in areas which are markedly divergent from the training data with which they have been constructed. This gap in understanding underscores the urgent need for more AI training, not just for developers, but for users, managers, executives, Board members, and others.

Still another user-organization-controllable risk is that the systems development life cycle for traditional information systems cannot be used because there are different risks, different checkpoints, different documentation requirements, different testing methods, and different approval processes with AI. Instead, each user organization will need to come up with its own "AI life cycle process," which makes sure that all the risks have been sufficiently addressed, and other requirements like documentation and full compliance with relevant laws and regulations have been met, before a system can be moved into production. While there is still significant merit to having something akin to the systems development life cycle process for AI systems, such an AI life cycle process needs to be tailored not only to the unique requirements of AI systems, but also to the involved organization's unique needs. For example, AI systems for which there are high-risk safety implications, such those which control aircraft or automobiles, will need to go through a very rigorous testing process prior to being ready to be released to the public. Likewise, if there is going to be public access, the AI life cycle process will need to consider special risks -- like "model stealing" (where a third party is able to extract much of the behavior of an AI system) -- risks which are not present in traditional information systems, but which are very serious matters in the AI realm.

Legal Duties and Related AI Control Measures: At the user organization level, each organization will need to do its own AI-related risk assessment, and this custom risk assessment should include not only concerns such as loss of customer trust, loss of sales, and higher insurance premiums, but also risks of being out-of-compliance with legal requirements. AI is bringing with it new threats such as the re-identification of persons whose privacy was previously protected by anonymization processes (see *Dinerstein v. Google* (2003)). The status of training AI systems with copyrighted material, and other intellectual property that belongs to others, but is also publicly accessible via the web, remains uncertain (see *Tremblay v. OpenAI* (2024)).

Beyond specific statutes like the NYC Local Law 144 (imposing restrictions on AI-assisted hiring), the Colorado AI Act (increasing liability for discrimination), and the EU AI Act (imposing a slew of requirements in the safety, transparency, and governance areas), there are fiduciary duties particularly relevant to the AI realm. For example, the Directors and Officers are legally obligated to exercise the duty of oversight over the organization's information systems, including the use of AI. The Directors and Officers for example need to know where in the business, and in what way, AI systems are being used. At the same time, the trend of "shadow AI," where user departments go their own way, and do not go through a central Information Technology Department AI Life Cycle, can make it very hard to learn about all the uses of AI. Directors and Officers have a duty to establish information systems that would keep in them adequately informed, and they must get involved if there are serious problems highlighted by this information system (see In re Caremark Int'l Derivative Litig., 698 A.2d 959, 971 (Del. Ch. 1996)).

Another fiduciary duty of Directors and Officers relevant to AI systems involves the duty of care, competence, and diligence. This duty requires that they take an active and direct role, and be clearly focused on the decision-making process, discharging their duties with the “care, skill, prudence and diligence under the circumstances then prevailing that a prudent man... would use in the conduct of an enterprise of like character and with like aim” (this is known as the prudent person standard). This duty includes being alert and paying attention to significant corporate problems, such as the risks of deploying AI. It also includes the duty to protect customer, business partner and other third-party information. Conducting regular risk assessments of AI deployments, before they move into production operation, would be one control that can help to show that Directors and Officers are indeed performing their duties in this respect (see *In the Matter of Twitter, Inc.,* Decision and Order at 2-4, FTC File No. 092-3093, Docket No. C-4316 (F.T.C. Mar. 2, 2011)).

Still another fiduciary duty of Directors and Officers, which is relevant to AI systems, involves the duty of obedience. That is the duty to follow established policies and procedures, as well as the requirements of existing laws and regulations. Shareholders and other involved parties, such as business partners, have a right to expect that the Directors and Officers will exercise reasonable supervision to ensure that staff pays attention to these matters. This is not something that should be approached as a cost-benefit analysis; this is instead a firm and essential component of corporate governance (see *Francis v. United Jersey Bank*, 432 A.2d 814, 823 (N.J. 1981)). In the AI realm, Directors and Officers need to have established accountability (for example via a Compliance Department), and also set-up internal procedures and processes to ensure that deployed AI systems are fully consistent with laws and regulations. A governance, compliance, and risk (GRC) system can for example be used to make sure that the organization is in compliance with all new AI laws, such as the EU AI Act.

Regulators are also getting into the act these days, and AI related activities related the Federal Trade Commission (FTC), the Consumer Financial Protection Bureau (CFPB), the Equal Employment Opportunity Commission (EEOC), and the Department of Justice (DOJ), are all now actively involved in the AI area. [4] Perhaps the greatest new AI-related compliance concern is “algorithmic disgorgement,” where organizations are required by regulators to destroy algorithms which have been shown to violate laws and regulations. This disgorgement penalty could mean that millions of dollars that were spent on developing an AI system would be lost. [5]

Suggested Way Forward: Policies are a great place to start to set-up a new and more serious approach to the risk management of AI systems at user organizations. They can and should be tailored to the adopting user organization's unique needs, and they can show support and encouragement from the highest levels of the C-suite and the Board of Directors. Policies can also serve as the beginning of an unfoldment of a new organizational reality, starting at the top of an organization. Once an AI-related risk assessment has been performed, responsive policies can then be chosen and adopted. At that point, a slew of infrastructure components that are consistent with those adopted policies can be generated. These subsidiary components include reporting relationships, job descriptions, governance structures, operational procedures, system design guidelines, technical standards, system architectures, system upgrade plans, technical tool acquisition plans, contingency plans, staff training systems, staff hiring plans, quality assurance approaches, compliance systems, vendor negotiation protocols, and many other organized ways in which risks can be reduced.

Accordingly, this author suggests that user organizations conduct an inventory of all the ways that AI is used within the organization, and also perform an AI-specific risk assessment, to come to the terms with the existing and anticipated ways in which the organization uses and expects to use AI systems, and the attendant risks. This effort should be followed by interviews with stakeholders within the organization, such as with the Chief Information Officer and the Chief Data Officer, to illuminate the areas of greatest risk management concern. This background information then can be used to select responsive AI risk reduction policies. Ideally these new AI risk reduction policies should be sitting on top of existing information systems risk management policies, such as those related to a GRC (governance, risk and compliance) system. While there are new, different, and special risks related to AI, much of the existing information systems infrastructure can be deployed to not only speed the approval and adoption of AI risk management policies, but also to minimize disruptions, to minimize cost, and to expedite the adoption of safe AI systems.

REFERENCES

[1] Turkle, Sherry, and Pat Pataranutaporn, “A 14-Year-Old Boy Killed Himself to Get Closer to a Chatbot. He Thought They Were in Love,” The Wall Street Journal, November 8, 2024, https://www.wsj.com (discussing the way that AI “exquisitely exploits human vulnerabilities” and illuminating the dangers of accepting “AI companionship”). Also see Duffy, Claire, “‘There are no guardrails.’ This mom believes an AI chatbot is responsible for her son’s suicide,” CNN Business, October 30, 2024, https://www.cnn.com/2024/10/30/tech/teen-suicide-character-ai-lawsuit/index.html.

[2] Hagendorff, Thilo, “Deception abilities emerged in large language models,” PNAS, April 3, 2024, https://www.pnas.org/doi/full/10.1073/pnas.2317967121 (about how AI systems become Machiavellian, how they are amoral, and how they lie). Also see Hendrycks, Dan, “AISN #20: LLM Proliferation, AI Deception, and Continuing Drivers of AI Capabilities,” Center for AI Safety, AI Safety Newsletter, August 28, 2023, https://forum.effectivealtruism.org/posts/Hg4dQqxyFpmkoYKeg/aisn-20-llm-proliferation-ai-deception-and-continuing.

[3] Steinhardt, Jacob, “On the Risks of Emergent Behavior in Foundation Models,” Bounded Regret, October 18, 2021, https://bounded-regret.ghost.io/on-the-risks-of-emergent-behavior-in-foundation-models/ (discussing the risks of emergent properties, including AI systems that could teach themselves to hack the security mechanisms of other AI systems).

[4] Reflecting the (U.S.) Federal Trade Commission (FTC) involvement, see the 2024 settlement with Evolv, which allegedly made misrepresentations about the extent to which AI was involved in its security screening systems. Reflecting the Consumer Financial Protection Bureau (CFPB) involvement, see the 2023 published guidance it has given lenders using AI about the transparency they must use when denying credit to applicants. Reflecting the Equal Employment Opportunity Commission (EEOC) involvement, see Mobley v. Workday (2024), where an individual using an AI-based job applicant screening system allegedly was discriminated against based on race. Reflecting the Department of Justice (DOJ) involvement, see the U.S. v. RealPage Inc. (2024), in which landlords alleged used an AI system to coordinate their rental price increases, thus breaking antitrust laws.

[5] Algorithmic disgorgement was used as a settlement related to the FTC v. Rite Aid Corporation, E.D. Pa. February 26, 2024 (pending) case, where face recognition was used as part of an AI-based surveillance system to identify shoplifters. Unfortunately, the system was not adequately tested, and it discriminated against people of color and women. See https://www.ftc.gov/news-events/news/press-releases/2023/12/rite-aid-banned-using-ai-facial-recognition-after-ftc-says-retailer-deployed-technology-without.

About the Author

Charles Cresson Wood, Esq., JD, MBA, MSE, CISM, CISA, CISSP, CGEIT, CIPP/US, is an attorney and management consultant specializing in the risk management of cutting-edge information systems, such as AI. His most recent published book is entitled “Internal Policies for Artificial Intelligence Risk Management.” His prior book was entitled “Corporate Directors & Officers’ Legal Duties for Information Security and Privacy: A Turn-Key Compliance Audit Process.” He is best known for his book entitled “Information Security Policies Made Easy,” which has been purchased by 70%+ of the Fortune 500 companies. He can be reached through his web site https://www.internalpolicies.com.

On-Demand Web Conferences

Every month ISSA International hosts educational live webinars focused on key issues and technologies for cyber security professionals. Access the Events page of ISSA.org at **https://issa.org/events/** or visit the ISSA International BrightTalk channel at: **https://www.brighttalk.com/channel/16125/**

COMMUNITY SPONSOR SERIES

ProcessUnity

GET OFF THE ASSESSMENT TREADMILL.
TAKE A DATA-FIRST, QUESTIONNAIRE-SECOND APPROACH TO TPRM

PRIVACY SIG

Why AI Needs Its Own Risk Management Policies and Processes

PRIVACY SIG

A CISO Guide to AI and Privacy

SPONSORED BY:

ISSA AIM Program

ArcLight6 Consulting
SECURING THE FUTURE

CREATING & MANAGING AN EFFECTIVE MENTORSHIP PROGRAM, METHODS & BEST PRACTICES TO CONSIDER

COMMUNITY SPONSOR SERIES

Close the Vulnerability Gap: Using Better Intelligence for Better Prioritization

SEVCO SECURITY

PRIVACY SIG

Privacy for the People by the People