DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Reference herein to any social initiative (including but not limited to Diversity, Equity, and Inclusion (DEI); Community Benefits Plans (CBP); Justice 40; etc.) is made by the Author independent of any current requirement by the United States Government and does not constitute or imply endorsement, recommendation, or support by the United States Government or any agency thereof.

Pacific Northwest
NATIONAL LABORATORY

PNNL-36662

An Ethics-Based Review of Generative Artificial Intelligence

Assuring Responsible Use

September 2024, Version 1.0

1 Jessica Baweja
2 Nancy Washton
3 Quentin Kreilmann
4 Jonathan Barr

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

1.0 Revision History

Version	Date	Updates
1.0	September 2024	--

2.0 Introduction

The emergence of generative artificial intelligence (GenAI) onto the global landscape in November 2022 led to both excitement about potential benefits of the technology and apprehension about ethical risks and challenges associated with its responsible use. Governments, corporations, and standards organizations have worked to release ethical principles to guide the development and use of GenAI (Hagendorff, 2020; Munn, 2023; NIST, 2023a). However, guidance around how to apply those principles in practice is more limited. Recent work has explored how to operationalize responsible and trustworthy AI principles (Canca, 2020; Mökander et al., 2023; Morley et al., 2021; NIST, 2023a), but nonetheless, a substantial gap remains between the specification of principles for responsible GenAI and the application of those principles in practice. Additional guidance is needed to support the responsible use of GenAI across science, industry, and government.

The lack of implementation guidance around the responsible use and deployment of GenAI is concerning given some of the associated risks that may not be addressed in current organizational policies and practices. A recent paper reviewed many of the challenges associated with assuring the safety of large language models (LLMs; Anwar et al., 2024). Those challenges were categorized into scientific limitations in our understanding of LLMs, issues with development and deployment methods, and sociotechnical problems (e.g., potential for malicious use) that are fundamental to the models themselves. Other authors have attempted to generate a taxonomy of risks associated with LLMs (Weidinger et al., 2022). The National Institute of Standards and Technology (NIST) divides risks of AI into harm to people or society, harm to an organization, and harms to an ecosystem (NIST, 2023a). Perhaps most useful when thinking about the ways that risks can be managed is to consider the underlying causes of the harms that might occur. A recent paper divides the potential harms associated with GenAI into those related to the design or development of the models and harms related to model misuse (Fischer, 2023).

2.1 Risks of GenAI

Legal issues constitute one type of concern related to the design and development of GenAI, with copyright infringement being the most common (Atkinson & Morrison, 2024; Fischer, 2023). However, there are many cases making their way through the legal system that range from to direct copyright infringement to claims of negligence regarding GenAI harms (Atkinson & Morrison, 2024). These cases remain unresolved, and so legally compliant design and usage of GenAI remains an open question in many ways. Another design issue specific to LLMs relates to the privacy of personal information. LLMs have exacerbated and created new risks regarding privacy through their aggregation of new information and the potential for inadvertent release of information through training data leakage (Lee et al., 2023; Wach et al., 2023). Finally, there are also major security concerns with LLMs: malicious users can jailbreak models, prompt injections can circumvent guardrails, and data poisoning can be used to insert harmful information or bias LLM outputs (Anwar et al., 2024; Zeng et al., 2024).

Separate from harms related to GenAI design are harms related to malicious, negligent, or irresponsible use of GenAI. Without appropriate oversight of their content, LLMs can perpetuate misinformation (De Angelis et al., 2023; Fischer, 2023; Xu et al., 2024). Given the apparent sophistication and fluidity of the outputs, naïve users may over rely on the accuracy of information produced by LLMs, which has the potential to cause serious harm if the information is safety- or security-critical. GenAI—both LLMs and text-to-image models—can also perpetuate

discrimination or bias, reproducing harmful stereotypes and supporting exclusionary norms (Anwar et al., 2024; Bird et al., 2023; Weidinger et al., 2023; Weidinger et al., 2022). The impacts of that embedded biases in training data may not be immediately obvious; although perpetuation of harmful stereotypes is one outcome, biases may also present themselves through misinformation on underrepresented groups, regions, or languages (e.g., answering a fact-based question about a country with “the United States of America” rather than “Namibia” due to the latter’s underrepresentation in the dataset). Finally, there are environmental, societal, and workforce implications related to the use of GenAI; while they may not preclude the use of the tool, they nonetheless should be considered when introducing a new application of GenAI (Anwar et al., 2024; Weidinger et al., 2023; Weidinger et al., 2022). Effects on the workforce, global economic development, and inequality in access to GenAI innovations all have broad implications that should be considered as the use of the technology is expanded (Anwar et al., 2024; Wach et al., 2023; Weidinger et al., 2023).

Regardless of the framework used to characterize them, organizations have an emergent need to review their policies and processes to evaluate whether the risks of GenAI are sufficiently managed. This report documents an effort to do so at Pacific Northwest National Laboratory (PNNL).

2.2 Responsible Use of GenAI

In this paper, we focus on supporting the responsible use of LLMs at a large research organization. These ethical considerations are especially important at government-funded research institutions, which serve not only a scientific function but also a societal one, serving as an example for other institutions regarding the responsible use of potential disruptive technology. Responsible use of GenAI involves developing a set of policies, informed by underlying principles, to support the ethical, trustworthy use of the tool; creating a governance process to manage the oversight of GenAI to determine if it upholds those principles and policies; and developing an AI Literacy program to train staff in policies and processes (NIST, 2023a). The policies at PNNL are still under development; however, this report begins by reviewing and outlining the ethical principles associated with the responsible use of GenAI that can serve as a basis for that policy. Second, we discuss a brief history of an ethical review of research to inform an ethics-based review of GenAI at PNNL. Third, we describe how the ethical review of research might be applied to GenAI, focused specifically on the modification of a human subjects research institutional review board (IRB; Grady, 2015). Fourth, we discuss AI literacy as it relates to use of GenAI. Finally, we present the results of an onboarding process designed to train staff on the ethical and responsible use of GenAI.

As we describe the ethical questions and issues, it is important to highlight that GenAI is a rapidly changing technology. As such, approaches to handling those issues must be reviewed and updated regularly to remain relevant. This document will be revisited regularly and updated; it is current only as of its publication date in September 2024.

3.0 Ethical Principles for AI

AI technologies have the potential to have a significant, disruptive impact on human life. With the tremendous power of AI to transform the ways that we live and work, concerns have arisen regarding the misuse, abuse, or ethical damage that this technology might bring. To address these concerns, researchers and professional organizations have released guidelines, frameworks, and ethical principles in recent years. Given the large numbers of these frameworks, there have also been several attempts to review, summarize, and synthesize these principles (Floridi & Cowls, 2022; Hagendorff, 2020; Jobin et al., 2019; Khan et al., 2022). These principles have varied in their specifics, but all have the general mission to outline the ways that developers, regulators, and users of AI can help to leverage the technology while also safeguarding humans from harm. Although they are not specific to GenAI, they can be applied to it.

Jobin et al. (2019) and Khan et al. (2022) both reviewed ethical principles and guidelines and identified the guidelines most common across the existing literature. Table 1 shows a list of the principles identified in these two review papers, as well as in the recently published AI Risk Management Framework (RMF) produced by the National Institute of Standards and Technology (NIST, 2023a) and the Institute for Electrical and Electronics Engineers (IEEE)'s Ethically Aligned Design (IEEE, 2022). Note that the papers vary in the ways that they identify different ethical principles: for instance, NIST's AI RMF (2023a) and Jobin et al.'s review (2019) both refer to accountability, but within the definition, describe responsibility as well. Furthermore, IEEE's document refers generally to respect for "human rights," and describes the need for AI to respect human dignity, freedom, diversity, safety, and security, encompassing many principles named separately in other papers. The table therefore attempts to indicate when a principle is included within a review or framework, even if the exact terminology differs. Any principle mentioned in at least three of these documents is shown in the rightmost column with a green circle. Obviously, such a mapping is imperfect given the variety of terms used across the field; nonetheless, it helps to represent an emerging consensus in the field of AI regarding the key principles for consideration around ethical and responsible use.

Table 1: Ethical Principles for AI

Ethical Principle	Khan et al. (2022)	Jobin et al. (2019)	NIST AI RMF	IEEE Ethically Aligned Design	Summary
Transparency	✔	✔	✔	✔	●
Privacy	✔	✔	✔	✔	●
Accountability	✔	✔	✔	✔	●
Fairness	✔	✔	✔	✔	●
Explainability	✔	✔	✔	✔	●
Justice	✔	✔	✔	✔	●
Non-maleficence	✔	✔	✔	✔	●
Beneficence	✔	✔	✔	✔	●
Responsibility	✔	✔	✔	✔	●
Safety	✔	✔	✔	✔	●

Data Security	✔	✔	✔	✔	●
Freedom	✔	✔	✖	✔	●
Autonomy	✔	✔	✖	✔	●
Sustainability	✔	✔	✖	✔	●
Human Dignity	✔	✖	✖	✔	●
Trust	✖	✔	✖	✖	●
Valid and Reliable	✖	✖	✔	✖	●
Solidarity	✔	✖	✖	✖	●
Prosperity	✔	✖	✖	✖	●
Effectiveness	✔	✖	✖	✖	●
Accuracy	✔	✖	✖	✖	●
Predictability	✔	✖	✖	✖	●
Interpretability	✔	✖	✖	✖	●

As the table shows, there appears to be an emerging consensus around the principles of transparency (sometimes also called explainability), privacy, accountability (responsibility), fairness (justice), beneficence and/or non-maleficence (i.e., AI for human good), and safety or security. These principles align well with those expressed in the NIST AI RMF (2023a). Notably, however, NIST also identifies validity and reliability as a key principle for ethical AI. It is likely that other frameworks did not include this principle explicitly because it was assumed: for instance, IEEE's report on ethically aligned design certainly includes the concept of validity and reliability in its narrative, even if it is not called out amongst its general principles. For the purposes of this effort, we will generally adopt the ethical principles outlined in NIST's AI RMF for the evaluation of the ethics of an application of GenAI. The next sections describe and define these principles based on the definitions in the AI RMF as well as the broader literature.

3.1 Definitions

Definitions presented here are adapted from existing literature. Again, different researchers and practitioners have applied different terms to ethical principles and concepts. The goal is not to argue for the use of a specific term, but rather, to accurately describe the intent of each ethical principle.

3.1.1 Accountability and Transparency

Also called responsibility, accountability refers to the attribution of responsibility for the AI system and its effects on the world (ISO, 2022; Jobin et al., 2019; Rossi et al., 2022). It often includes some aspect of legal liability or attribution of responsibility contractually as well (ISO, 2022). Different stakeholders in AI systems (e.g., designers, developers, end users) are referred to as accountable for those systems, depending on the context; however, many standards suggest instead that everyone is responsible for the AI system, regardless of their precise role (e.g., Rossi et al., 2022). However, the role that each person plays may impact the nature of their accountability. For instance, AI developers may be considered accountable for interrogating their training data for potential bias; end users, in contrast, could be accountable for ensuring that bias does not manifest in the outputs of the system. Because the focus of this document is governance of the application or *use* of AI, especially GenAI, accountability here primarily focuses on accountability of the end user to monitor system outputs, including through

the use of LLM application programming interfaces (APIs) or of those systems to generate code or other system elements for further development. .

Notably, NIST's AI RMF combines accountability with transparency, as transparency is a precondition for accountability (NIST, 2023a). In the case of GenAI, explaining the reason behind the system's outputs is challenging due to the black box nature of GenAI systems (i.e. systems based on neural networks). However, accountability for GenAI does require transparency; at minimum, users must be aware that they are interacting with a GenAI system (and not another human). If a user is unaware of the GenAI, they cannot be held accountable; in such a situation, the person responsible for deploying the system into that environment would be accountable instead.

Transparency, as defined here, is the extent to which information about an AI system is open, comprehensive, accessible, and clear (ISO, 2022). Although this principle is prevalent in many of the sources in the literature reviews mentioned earlier, there was also substantial variability in its precise definition. Some authors use transparency interchangeably with explainability (Jobin et al., 2019); the NIST AI RMF, however, does not. Here, we differentiate between transparency and explainability due to the nature of GenAI. Because GenAI systems are "black-box" systems where the reasoning or justification for a decision cannot be easily explained, the concept of traditional explainability (as in explainable AI, or XAI) may be less immediately relevant. However, transparency, meaning clear and understandable information of the system's capabilities and limitations, is clearly relevant to trustworthy and responsible use of GenAI. As already noted, transparency is also a necessary condition for appropriate accountability. Although defined separately, they may be evaluated in tandem, as users cannot be held accountable for a system that is not sufficiently transparent.

3.1.2 Privacy

Privacy refers to the "norms and practices that help to safeguard human autonomy, identity, and dignity" (NIST, 2023a). Responsible and trustworthy use of GenAI requires that the system be used and managed in a way that respects the human right to privacy. This includes not only obvious protection of privacy through exclusion of sensitive information (e.g., personally identifiable information; PII), but also consideration of the ways that private information might inadvertently be leaked. GenAI can compromise privacy by leaking sensitive information included in training data, it can lead to emergent privacy risks due to its aggregation of large volumes of data, and it can inadvertently disclose additional information through inference (Anwar et al., 2024; Gupta et al., 2023). As a result, leveraging tools to enhance the privacy of the systems is necessary to ensure that the use of GenAI remains responsible and trustworthy (NIST, 2023a).

3.1.3 Fairness

In AI in general, and in GenAI specifically, fairness includes concerns about equality as well as equity, including prevention or mitigation of bias and discrimination, and has direct ties to the concept of algorithmic bias (Jobin et al., 2019; NIST, 2023a, 2023b). In some ways, GenAI does not differ from human interactions in the sense that there continue to be challenges in managing concerns around discrimination, hate speech, and exclusion; however, this does not absolve users of GenAI of the responsibility to manage the risks associated with fairness in their use of the technology. GenAI presents new potential risks around social stereotypes, hate speech, or exclusionary norms (Weidinger et al., 2022); managing those risks of harmful bias is therefore an important principle to support fair use of that technology. In addition to considering the

fairness of the technology itself (e.g., the use of unbiased training data), users of GenAI should also consider the accessibility of the technology in deployment to help support equitable distribution of the benefits(Ashok et al., 2022; Wach et al., 2023; Weidinger et al., 2023; Weidinger et al., 2022). For instance, are the user interfaces designed with accessibility in mind (e.g., for those with visual impairments)? Do affected groups have equitable access to systems that may impact them? As GenAI is incorporated into more impactful decisions, considering the accessibility concerns becomes increasingly important.

3.1.4 Explainability and Interpretability

Explainability refers to a "representation of the mechanisms" underlying the outputs of an AI system, whereas interpretability refers to a meaningful description of the system's output in the context of the intended use of the system (NIST, 2023a, 2023b). At the heart of both explanations is the understandability of the explanation to the intended audience. The explanations need not be technical nature, and for many uses of GenAI, the intended audience may not be a highly technically sophisticated one; tailoring the explanation to that audience is important to upholding the intention of explainable and interpretable AI.

Explainability, interpretability, and transparency are all interrelated concepts. Unlike transparency as we have defined it, explainability and interpretability typically refer to the inner workings of the model itself and a description about the causes of decisions or outputs in a way that humans can understand (NIST, 2023b). Given the focus of this document on GenAI, explainability and interpretability does not generally apply. Although there is ongoing research exploring the ways that GenAI can be made more explainable, it remains a challenge in the field. In the application of these principles to GenAI, we focus instead on the concept of transparency. However, research in the field of explainability and interpretability of GenAI is progressively rapidly; explainability should therefore be revisited regularly to determine whether it has become more central to responsible use and deployment. In the current state of research, transparency around training data, uses, and limitations is more applicable.

3.1.5 Safety

Safety as a principle refers to the use of AI only in contexts where it will not, under defined conditions, endanger human life, health, property, or the environment (ISO, 2022; NIST, 2023a). Other documents refer to this principle as non-maleficence, meaning that the AI should not cause foreseeable or unintentional harm (Jobin et al., 2019). Of course, AI systems can be intentionally misused, and such malicious use is typically excluded from considerations around safety (Anwar et al., 2024; Jobin et al., 2019); instead, protecting AI from misuse is generally considered when reviewing the security of the systems. Considerations around safety for GenAI primarily relate to the ways that it is being used and the potential for harm that might result even in cases of system error or failure (NIST, 2023a). This includes not only safety from a physical perspective, but also safety from psychological harm, as misuse or inappropriate use of GenAI might also lead to psychological harm—for instance, there are ethical and safety-related concerns associated with the use of LLMs for psychotherapy {Raile, 2024 #513}. Essentially, safety for use of GenAI should focus on monitoring, simulation, and testing to help mitigate the likelihood of harms in the event that the system deviates from the intended function.

3.1.6 Security

Secure AI systems are those that maintain confidentiality, integrity, and availability through protection mechanisms to prevent unauthorized use (NIST, 2023a); security is therefore defined

as resistance to intentional acts designed to harm or damage the system (ISO, 2022; NIST, 2023b). Other documents combine security with non-maleficence or freedom from harm (Jobin et al., 2019); however, here, as mentioned, we distinguish between unintentional acts that may cause harm (included under the principle of safety) and intentional acts that may cause harm (included here, under the principle of security). Responsible use of GenAI requires consideration of resilience to unauthorized acts that may occur. Note that there remain many challenges in the security of GenAI, including LLMs; however, to achieve responsible use of GenAI, security must be considered and reasonably managed in the application of GenAI (Anwar et al., 2024). This includes consideration of adversarial prompts or exfiltration of training data or intellectual property information through misuse of the GenAI (NIST, 2023a).

3.1.7 Validity & Reliability

Validity, reliability, and accuracy are three interrelated concepts that are not typically included in many documents exploring principles of trustworthy or responsible AI (Jobin et al., 2019; Khan et al., 2022). However, inaccurate information is a well-known risk of LLMs (Anwar et al., 2024; Weidinger et al., 2022); these so-called hallucinations have been called “inevitable” (Xu et al., 2024). Thus, although validity and reliability are not commonly included in many ethical principles, due to the nature of GenAI, we believe that they are critical here. Validity refers to confirmation through objective evidence that the requirements for a specific use have been fulfilled; reliability refers to the ability of a system to maintain performance without failure under a variety of circumstances (NIST, 2023a, 2023b). Finally, accuracy refers to closeness to accepted or ground truth (NIST, 2023a, 2023b). The concepts of validity, accuracy, and reliability are distinct from the concept of accountability above, because accountability encompasses concepts beyond simply the accuracy of the content; for instance, users are accountable for assessing the safety or privacy implications of their findings. However, it is also related, as one aspect of accountability involves monitoring the accuracy of GenAI results. Although responsible use of GenAI cannot guarantee its accuracy without human verification given current limitations of the technology, its appropriateness for use and its consistent performance are reasonable principles to uphold. In addition, understanding the system limitations and reviewing its content for accuracy is a critical component of responsible use. Here, we emphasize the importance of validity and reliability as well as human verification of the accuracy of the results.

4.0 Ethical Review of Research

The principles outlined in the previous section can serve as the basis for policies regarding responsible GenAI use. However, to ensure compliance with these policies, additional governance is needed. In short, a review process is necessary to help ensure that use of GenAI remains responsible and ethical. Ethical review of research has been a longstanding challenge, and most saliently encountered in the areas of biomedical and behavioral research (DHHS, 1979; Grady, 2015). In research studies where humans are the subject of study, it is necessary to impose ethical rules and boundaries around what is and is not acceptable. In developing those rules, the fields of science devoted to the study of human subjects have grappled with many of the same issues that are currently being encountered in the field of artificial intelligence: first, the need for guiding principles to differentiate activities that are within acceptable ethical boundaries and those that are not; second, a procedure for evaluating compliance with those principles; and third, a group of individuals who are capable of assessing whether a specific project complies with those guidelines. Within the domain of human subjects research, the solution to this problem was the creation of Institutional Review Boards (IRBs).

4.1 History and Practice of the IRB

IRBs are ethics committees made up of individuals unaffiliated with a proposed human subjects research study who are charged with conducting an impartial review of that study before it can begin. Any research study that involves human subjects in the United States (US) that is funded by US federal agencies or where the results of the study are under the jurisdiction of the US Food and Drug Administration is required to undergo IRB review (Grady, 2015). Most research institutions can and do extend those requirements to all human subjects research conducted at that institution (e.g., universities).

Although there were earlier discussions of the concept of an ethics committee for human subjects research, in the US, the establishment of the IRBs into regulation occurred in 1974 (45 CFR 46), which introduced the term IRB. Five years later, the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, a committee established by the Department of Health and Human Services (DHHS), authored the Belmont Report (DHHS, 1979). In the Belmont Report, the committee outlined the key ethical principles to be upheld during the conduct of human subjects research: justice, beneficence, and respect for persons. In sum, human subjects research must equally distribute its benefits and potential harms, should do no harm, and should respect human autonomy (or protect those with diminished capacity). These principles were incorporated into the Code of Federal Regulations (CFR) in 1991 through the Common Rule (45 CFR 26 Subpart A). The Common Rule was also codified in separate regulations by 15 Federal departments and agencies. IRBs were designed to implement the ethical review relative to the Belmont Report and must be registered with the Office for Human Research Protections (OHRP) (45 CFR 46 Subpart E and 21 CFR 56.106).

In addition to establishing the need and applicability for ethical review and designating IRBs as the mechanism by which such review is completed, the CFR also outlines the requirements for the IRB membership, functions, operations, and review, as well as outlining specific criteria for expedited forms of review. These details help to elaborate on how the review of human subjects research is performed in order to ensure that it conforms to the expected ethical principles. As already noted, these requirements apply to all research involving human subjects that is conducted, supported, or otherwise subject to regulation by any federal department or agency.

The requirements for IRBs are described here because they can serve as a model to inform how the IRB might be adapted to the AI ethics context.

4.2 IRB Membership

Regulation specifies that the IRB is made up of at least five members with a diverse background, including race, gender, cultural and professional backgrounds. The goal is to have a group of individuals whose backgrounds allow them to consider research especially for vulnerable groups—e.g., economically or disadvantaged persons, children, prisoners, or individuals whose capacity to provide consent might somehow be impaired. The IRB also must include one individual whose professional expertise is in a scientific field and one individual whose expertise is in a nonscientific field. Finally, the IRB must include someone unaffiliated with the institution conducting the research, and the members must not be affiliated with the project being reviewed. In so doing, the requirements for IRB membership attempt to capture the diversity of the population who may be affected by the research. The requirements also ensure that there is sufficient scientific knowledge to adequately evaluate the work being proposed and minimize the risk of conflicts of interest. Again, such a model could be adapted to the AI use case by creating an ethical committee that includes a diverse population with relevant scientific and data scientific knowledge, as well as including a member unaffiliated with the institution to reduce the likelihood of institutional bias.

4.3 Review Procedures

IRBs are given the authority to review and approve or disapprove all research activities at the institution involving human subjects. They are also given the authority to require modifications in order for the work to proceed. In that review process, the IRB is required to determine that the risks to the participants are minimized, that those risks are reasonable when considered the potential benefits of the work, that selection of participants is equitable (e.g., not placing undue burden on certain vulnerable groups), and that informed consent of the participants is obtained and documented. They also review to assess whether there are sufficient provisions in place to protect the safety and privacy of the human subjects.

The IRB requires information from the researchers regarding the planned work to conduct its review. The way that most institutions have implemented IRB review is through the use of protocols that must be submitted to the IRB before work can begin (Ritchie, 2021). These protocols ask a set of questions to determine whether the participants in the study are being treated ethically. The researchers are asked to provide information about how the proposed research maximizes benefits while minimizing harms, respects the autonomy of the participants (through informed consent), and upholds fairness and lack of bias through equitable and reasonable selection of participants. Through the structured use of questions, these protocols ask the researchers themselves to critically consider the ethics of the proposed plan and to find ways to minimize the potential risks while maximizing the potential benefits.

The benefit to such a system is that it provides a means by which the IRB can gather information to subjectively assess the risks presented. This allows for necessary subjectivity in the ethical review: members of the IRB, when appropriately staffed, are well-placed to apply their knowledge and budget to determine whether the projects as proposed supports the ethical principles outlined in the Belmont Report (DHHS, 1979). The researchers themselves are also made to consider if there are ways that they might be able to reduce the risks in their proposed plan. This places the responsibility on the researchers themselves to identify ways that they

could minimize the potential harms. Again, such a process could be adapted to the AI use case; if ethical principles are identified, the team developing or using the AI could then be trained on the principles and asked to consider the risks presented by their proposed use of AI technologies.

Once completed, the IRB members review the submitted protocol and approve, suggestion modifications, or disapprove the study. IRB review can take one of three forms: exempt (certain categories of research require no IRB review); expedited (a single member reviews the protocol), or full convened review (by all members of the IRB). The type of the review is based on the risk posed by the study.

4.4 Exempt and Expedited Research

The regulation of human subjects research also recognizes that not all research is equally risky or requires the same degree of ethical review. Projects are divided into three tiers: exempt, expedited, or full review. Certain categories of very low risk research are considered exempt from IRB review ("Protection of Human Subjects," 2017). This includes research that is based on established methods in educational settings, such as assessment of instructional strategies, certain kinds of research conducted on anonymized data, or other studies that are benign, harmless, or painless, and require little intervention beyond what occurs in the setting naturally. The regulation also specifies an exemption for analysis of secondary data where no additional consent is required and for research that is generally designed to improve performance of the organization conducting it but for no other purpose (e.g., performance monitoring of internal processes, food quality evaluation). Exempt use cases, in character, are those studies that present no risk of harm and involve minimal intervention on the part of the humans participating in them beyond those that are already present in daily life.

In addition, regulation also establishes criteria for studies that require only a minimal review, which it terms expedited. Expedited studies are those that present no more than minimal risk to participants; they are generally projects that represent minimal risk to the participants and fall into a specific set of categories (e.g., biological specimens collected by noninvasive means). Expedited studies present a higher risk than those specified as exempt, but less than those that required a full, convened review by the members of the IRB.

The specification of exempt and expedited categories for ethical review helps to streamline the need for ethical review in those cases where the risk is low. In the same way, the application of an ethics committee to the review of AI use cases could specify those conditions where a specific project could be exempt from review or receive an ethical review. Such exceptions would minimize the need for administrative processes in those cases where the possibility of harm from the use of AI is low.

5.0 Application to GenAI Use Cases

Ethical review of the use GenAI is challenging, subjective, and requires the creation of an institutional process to execute it. Fortunately, the structure of an IRB for human subjects research solves many of the challenges: it specifies the membership of an IRB, the review procedures that can be used, and the methods by which information can be obtained for that review. That is, many of the decisions that would need to be made in the creation of an ethical review procedure for AI can be adapted from the existing regulations and guidance. The next sections discuss the ways that IRB procedures can modified for the creation of a GenAI assurance committee (GAC) to conduct an ethics-based audit. The procedures described here have not yet been implemented but outline the ways that PNNL could implement such an audit within the structure of their existing project risk management processes.

5.1 Structure of the GAC

To conduct a thorough ethical review of a GenAI project, the makeup of the GAC should be diverse, should include both technical and non-technical members, as well as unaffiliated members of the community. This membership helps to support an ethical review board that goes beyond assessing the scientific merit of the project and includes the stakeholders and impacted community of the GenAI project. A proposed breakdown of a six-person committee, and their roles and responsibilities, are below. Note that the actual membership of the GAC will include more than six people to fill each role, but no more than six members will need to be convened for a single review. Roles on the GAC include a chair, data scientist, AI ethicist, non-scientist, domain expert, and an unaffiliated member.

For expedited GAC applications, only one member of the GAC is required to review the application. Certain members of the GAC will be designated as sole reviewers given their knowledge and expertise in AI methods and ethics. For full board review applications, at least six members of the GAC will need to review. Note that the GAC will be made up by more than six individuals, and those individuals might serve different roles on the review panel depending on the specific application. That is, a GAC member might serve as a domain scientist reviewer on one application due to their expertise in the field; they might serve as a data scientist for another application if they have the appropriate expertise to do so.

The role of domain scientist, which here refers to an individual with expertise in the field where the data science is being applied, is likely to require a large body of individuals to draw from with appropriate expertise in many scientific disciplines. Each individual may serve as reviewer for only a small number of GAC applications per year depending on the number of projects or the need for the role that the individual serves on the GAC. However, it is important to have reviewers available across all scientific disciplines and areas of expertise where the institution performs work. In addition, a person may serve multiple roles in review of an application: they may, for instance, serve as both a non-scientist and an unaffiliated member of the GAC. All full applications should still be reviewed by at least six individuals, but each role may be filled by fewer than that depending on the exact makeup of the review panel. The GAC Chair will be responsible for ensuring that each review board session represents a diverse group of members with appropriate expertise for evaluating the merits and ethics of a given GenAI project.

All members of the GAC will need to take training to ensure that they are informed about the ethical principles to be upheld (i.e., NIST's AI RMF), the potential risks presented by GenAI projects, a base level of knowledge about GenAI methods, and the GAC's role in overseeing

GenAI projects at PNNL to mitigate those risks. Different institutions have outlined different requirements for membership on the IRB; however, there is generally a training process and a proposed time commitment (e.g., four hours per month). The expectation is that members would be able and willing to convene to review research on a regular basis. That training is described in more detail in Section 4.0.

5.1.1 GAC Chair

The GAC chair is responsible for ensuring that the reviews of the GAC comply with any policies or requirements put in place by the organization (or any future regulatory requirements). They review applications presented to the IRB and communicate with other reviewers as needed. The chair may serve as the sole reviewer for expedited applications (although other members of the GAC might perform that function as well if needed). The chair reviews all full board applications to the GAC along with all of the other members. The GAC chair is also responsible for ensuring that the staff at the institution complete training necessary and that the training implemented is sufficient. They will also communicate GAC feedback for applications to GenAI project PIs and/or PMs.

One of the main responsibilities of the GAC chair is to oversee GAC meetings, ensure that the reviews and approvals meet the criteria necessary for approval, and review any incident reports and project closeout reports. This includes helping to ensure that, during the review of GAC applications, all voices within the review panel are equally heard and valued, and that all members are participating in the review of the GenAI project. In general, the GAC chair serves as the organizer and key member responsible for the diligent completion of the GAC's responsibilities.

5.1.2 Data Scientist

The data scientist member of the GAC is responsible for reviewing the methods applied in a GenAI project and described in a GAC application to ensure that the proposed methods sufficiently mitigate the ethical risks associated with that project. Due to the technical nature of GenAI projects, it is necessary that the GAC have a member sufficiently versed in the work to evaluate whether the risks associated with a project are reasonably mitigated by the methods outlined in the application. The data scientist is not intended to evaluate the technical merit of the project outside of the impact those technical choices have on the proposed benefits and risks of the GenAI project. Instead, the data scientist member reviews the technical choices to ensure that any ethical risks are mitigated.

5.1.3 AI Ethicist

A number of professional societies are now issuing certifications in AI ethics (e.g., IEEE CertifAIed™). The GAC should include at least one member who has been issued a certification or credential in AI ethics or can otherwise demonstrate credible expertise in this area. This person will be responsible for contributing that knowledge and expertise in GAC reviews, especially for full board applications. They will be responsible for ensuring that the project takes sufficient measures to mitigate all relevant ethical risks, including requesting additional information from the project team to evaluate this as needed.

5.1.4 Non-Technical

The non-technical member is intended to serve as a member of the public or the community at large whose interests are not in scientific areas. The intent is to make the GenAI project under review be accountable to the public at large as well as the scientific community. The non-scientist reviewer is intended to help assess whether the proposed project upholds the community values, views, and norms, and that any potential risks are mitigated.

5.1.5 Domain Expert

Unlike human subjects research, AI reviews require expertise in data science and in the scientific discipline or the domain where the project plans to apply GenAI (e.g., biology, nuclear security, physics, chemistry). The domain expert member of the GAC should have knowledge and expertise necessary to evaluate the capabilities of the model itself as it relates to the discipline and of the risks that could arise within that use. The purpose of the domain expert's review is to evaluate the GAC application for any risks that might be specific to the application of GenAI in that domain or discipline that might not be apparent to someone outside that field. This include concerns about operational risks, laboratory safety considerations, or any other risks that are specific to that use of GenAI.

As already mentioned, having a domain expert on the GAC will require a cadre of potential scientists to draw from who can serve on the GAC, many of whom may review a small number of applications during any given period. However, this is necessary given the breadth of potential applications of GenAI. This position will likely rotate often for full board reviews, with many scientists and experts serving in this role across the institution.

5.1.6 Unaffiliated Member

To mitigate potential institutional conflicts of interest or due to power structures within the institution, including an unaffiliated member on the GAC is critical. This unaffiliated member is not a member of the institution (i.e., PNNL). They may or may not have data science experience or scientific expertise. This role often overlaps with the role of non-technical member, as some of the responsibilities and expectations are similar. Again, this person is intended to represent the community viewpoints, norms, and values. However, they have the additional responsibility of serving to ensure that reviews of the GAC applications are fair and not biased by any potential institutional conflicts of interest (e.g., a desire to please research sponsors; a desire to protect institutional reputation) and are sufficiently merited.

5.2 GAC Procedures

GAC review procedures begin at project initiation, when it is determined whether or not a proposed project will require GAC review (or is either irrelevant or exempt from GAC review). For relevant projects that are not exempt, the project team will complete a GAC application and then the GAC will conduct its review and determination. During the project execution, the PM will be responsible for event reporting of any incidents that occur during the project related to ethical risks. There will also be an ongoing review process. Finally, there will be a project closeout process. This overall process is summarized in Figure 1. The next sections review each of these steps in more detail.

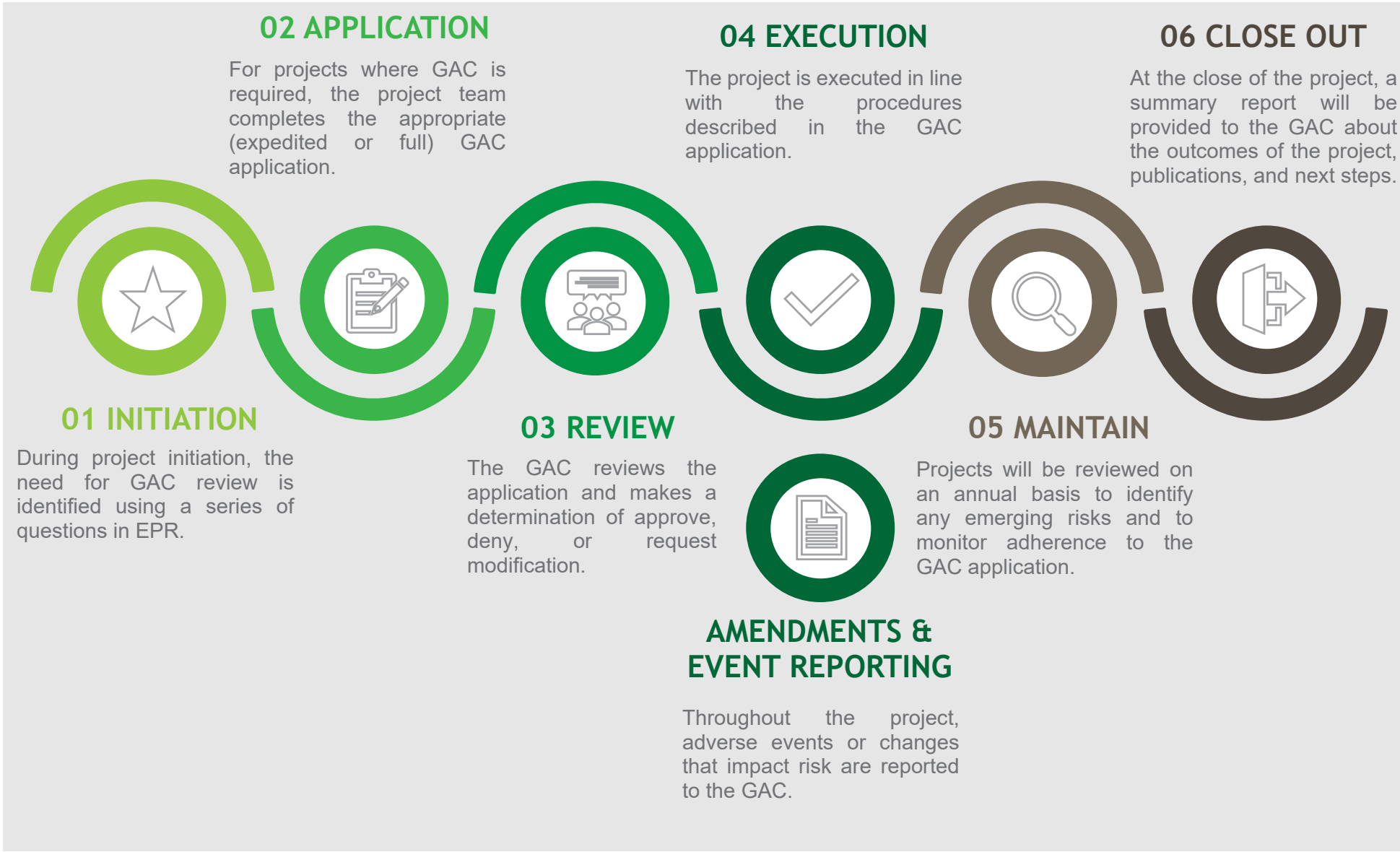

Figure 1: Proposed GenAI Assurance Council Process

5.2.1 Project Initiation

The initial determination of exempt, expedited, or full board review will be made by the Project Management Office Director (PMOD) in collaboration with the GAC based on information provided by the PM during project initiation. During the project initiation process in Electronic Prep & Risk (EPR), the PM will answer two questions about the project:

- Does this project significantly involve artificial intelligence (AI) or machine learning (ML)?
- Does the project significantly involve generative artificial intelligence?

After completing the relevant questions, the PM will work with the PMOD to assess the need for GAC review and the type of review that would be required. To guide this review, we have created a risk assessment matrix to evaluate the risks associated with the proposed application of GenAI. This risk assessment matrix can be used to help the PMOD to determine whether a project needs additional review. Risk matrices are a common tool for understanding and assessing risk; they often list the risk, severity of the impact, and the likelihood. Other approaches evaluating the riskiness of a technology list the principles to be protected and assess the risk of a technology relative to each (O'Neil & Gunn, 2020). This framework has been applied in AI ethics for applications in cybersecurity (Bruschi & Diomede, 2023). The risk matrix is presented in Figure 2. Using this risk matrix, with additional questions to the PM as needed, the PMOD will identify whether the project qualifies as exempt (no additional review required), expedited (a single GAC member will conduct further review), or full (the complete GAC will conduct further review). In short, exempt projects are those that are characterized as entirely low risk across all principles; expedited projects may be up to medium risk relative to one or more principles; full review projects include at least one aspect that qualifies as high risk. Additional description to characterize each type of project is below. If needed based on the PMOD assessment, the PM will work with the project team to complete the relevant GAC application.

Figure 2: Risk Assessment Matrix

RISK DOMAIN	LOW RISK	MEDIUM RISK	HIGH RISK
VALIDITY & RELIABILITY	The project team has good knowledge in the domain where the AI is being applied, and failure of the model can easily be recognized and corrected. There are metrics or methods to understand and verify the validity and accuracy of the AI output.	There are real-world gaps in the team's knowledge and understanding of the AI outputs in the domain where it is being applied, or validity or accuracy of model outputs is difficult to assess.	The team has limited understanding of the domain where the AI is being applied, or there is no way to evaluate the validity or accuracy of the outputs.
FAIRNESS	The content generated will not have a substantial or significant real-world impact on users, recipients, or stakeholders. Any biases can be identified and mitigated as necessary.	The content generated may have a substantial or significant real-world impact on users, recipients, or stakeholders. There are potential biases that may be hard to address.	The content generated will have a substantial or significant real-world impact on users, recipients, or stakeholders. The degree to which bias is embedded in the content is difficult to ascertain or mitigate.
SAFETY	There is very little potential for harm from the use of the AI system beyond those that are already present. Any failures of the AI system are contained and can be managed.	There are potential safety issues from the use of AI in the application that can be easily identified and managed. Failure of the AI would not cause significant harm.	There is a high risk of harm from failures of the AI. Failures of the AI would be hard to control.
ACCOUNTABILITY & TRANSPARENCY	The project team will review and be accountable for all GenAI content. The user has an understanding of the AI system's capabilities and limitations in order to appropriately manage the outputs. The content will either not be published externally, or, if published externally, workflows will be documented, all output will be carefully reviewed, and the ideas expressed will be primarily human generated.	The project team will review and be accountable for all GenAI content and will appropriately manage the outputs. However, the content will be published externally. The AI system plays a larger role in the creation of the content. Workflows are documented and reported as part of the document.	Content generated will be part of a formally published external document, and the AI plays a significant role in the conclusions, writing, and ideas in the document. There may be no documentation of the workflow used to produce the results.
PRIVACY	The data entered into the system will be entirely open-source and non-sensitive. If that is not true, the AI system is approved for the type of information that will be entered into it, such as classified, restricted, business or business sensitive, personally identifiable, controlled unclassified, or official use only information (i.e., PII, CUI, or OUO).	Data entered into the system are open-source, and not CUI, classified, business sensitive, restricted data, strictly private, or PII. However, the planned use of GenAI may develop into OUO, CUI, classified, restricted, or business sensitive information, such as intellectual property (IP) or cost information.	Data entered into the AI system may include PII or classified, business sensitive, restricted data, or strictly private information.
SECURITY	There are protections mechanisms in place to maintain the confidentiality, integrity, and availability of the AI system. The AI system is resistant to foreseeable intentional malicious acts designed to harm or damage it.	There are protection mechanisms in place to maintain the confidentiality, integrity, and availability of the AI system. Intentional malicious acts may not be as readily foreseeable, and therefore it is not clear how resistant the AI system is to intentional acts designed to harm or damage it.	The protection mechanisms to maintain the confidentiality, integrity, and availability of the system are unclear or are not in place. Intentional malicious acts designed to harm or damage the system may be more likely to be successful.

5.2.1.1 Exempt GenAI Projects

Exempt projects will not undergo GAC review because they present minimal risk. Put differently, they are assessed to be low risk across all principles by the PMOD. In exempt projects, the project team has thorough knowledge and understanding of the domain and can easily identify and correct AI failures and the content generated has minimal or no significant real-world impact, to include minimal potential for harm even in the case of failure. As with all uses of GenAI, the project team will remain accountable for all content, and are informed users of the AI system, including its capabilities and limitations. Strong security measures are in place to protect the AI system against malicious acts. The data being used are either open-source and non-sensitive. Sensitive data types include CUI, PII, business sensitive, or strictly private information. Projects may use Official Use Only (OUO) information if the system is approved for that information (e.g., AI Incubator Chat). In summary, GenAI projects that qualify as exempt would generally:

- Use open data sets for research or analysis without any personal identifiers;
- Perform tasks with no direct impact on individuals, such as playing games or creating non-personalized content; and
- Serve as aids in non-critical decision-making processes where human oversight is readily available.

Examples of GenAI projects that might be considered exempt from GAC review would be:

- Creation of educational content for non-safety or security-critical domains.
- Simulated energy usage data to understand grid distribution.
- Search, summarization, or synthesis of documents or literature with appropriate human verification of results.
- Generation of routine responses to frequently asked questions with no safety or security implications.

With all types of AI projects (exempt, expedited, or full review), the determination will be made based on the planned activities at the outset of the project; if those activities change, the PM will be responsible for modifying the scope in EPR and addressing the associated changes in risk and risk management.

5.2.1.2 Expedited Review GenAI Projects

GenAI projects that require expedited review may leverage sensitive information, such as indirect or anonymized data on human subjects. Any application that might harm individuals in some way requires expedited review, but the associated risk must be minimal. There may be potential safety or security consequences, especially in the case of system error or failure, but these risks are managed and do not pose a significant threat. The outputs of the system may have a meaningful real-world impact on end users or stakeholders. Security measures are in place, but resistance to intentional malicious acts is less certain. In summary, GenAI projects that qualify as expedited may:

- Process non-sensitive, non-human data, or process anonymized personal data where there is a strong safeguard against re-identification;

- Support decision-making in sensitive areas, such as national security or healthcare, but where decisions are not solely made by the AI; and
- Influence user behavior or decisions, but where the stakes are not life-critical, and errors can be tolerated to some extent.

Some examples of projects that would qualify as expedited might be:

- AI-assisted drafting of background sections for reports on energy policy, where the content has a moderate impact on stakeholders.
- Research on the environmental impacts of potential energy projects, using AI to synthesize findings that might influence public opinion or regulatory decisions.
- Use of Gen AI to analyze patent databases where findings might lead to insights involving business-sensitive information.

5.2.1.3 Full Review GenAI Projects

GenAI projects that require full GAC review are those that involve highly sensitive data, including personally identifiable information (PII), proprietary business data, or classified information. These projects have the potential for significant safety or security consequences for individuals or groups, particularly in the event of system errors or failures. The risks associated with these projects are higher, and the validity and reliability of the GenAI's outputs are critical due to their substantial real-world impact. These projects often require rigorous ethical considerations, robust data protection measures, and comprehensive accountability mechanisms.

There is a wide breath of studies that might require full GAC review. Some characteristics of full review projects might be:

- GenAI systems that make autonomous decisions with significant consequences for individuals or society;
- GenAI systems that make decisions or recommendations that are security- or safety-critical;
- GenAI systems that involve active surveillance or monitoring of individuals, potentially impacting privacy rights; or
- GenAI systems that handle sensitive personal data where incorrect outputs could lead to discrimination or harm.

Some examples of GenAI projects that would require full GAC review are:

- Autonomous Vehicles: AI that operates vehicles and must handle sensitive data regarding passengers and their travel patterns.
- Facial Recognition for Security or Surveillance: AI that identifies individuals in real-time for security or surveillance purposes.
- Predictive Policing Systems: AI that uses personal data to predict crime hotspots or individual's likelihood of reoffending.
- Customized Education Platforms: AI that adapts learning experiences based on personal data from students.

- Personalized Marketing: AI that uses personal data to create targeted advertising campaigns.
- Employee Monitoring Systems: AI that uses personal data to monitor employee productivity and behavior.

Each type of project requires a different level of oversight and review to ensure that the GenAI systems are developed and deployed responsibly, with appropriate consideration for the ethical, privacy, and safety implications associated with their use. These divisions between exempt, expedited, and full board review GenAI projects are intended to made based on information at project initiation to ensure that these risks are sufficiently mitigated.

5.2.2 GAC Review and Determination

Following project initiation, the GenAI project will be identified as needing GAC review or as exempt from review. If the GAC feels that the project presents more than minimal risk, then an expedited or full board review will be required. Each determination is associated with an application designed to assess the risks of the project and to determine whether those risks are sufficiently mitigated. Again, risk is defined relative to the seven NIST principles outlined at the beginning of this document. Just as with a human subjects IRB, review of an expedited protocol would be completed by a single member of the GAC; review of a full board protocol would require review by six members of the GAC.

Questions for a GAC application (expedited and full) are presented in Appendix A. These questions are designed to elicit information about the proposed GenAI project, the potential risks to the ethical principles outlined earlier, and the ways that the team plans to mitigate those risks. The full application process asks additional questions about the data protection procedures given the higher risk nature of the information involved.

Using the information gathered on the GAC application, members of the GAC will complete a review. For expedited projects, a single member of the GAC (e.g., the chair) is sufficient; for higher-risk projects, a full review is required, where at least six members of the team review the project's application. Using the information on the protocol, the GAC will seek to determine:

- Whether the potential benefits of the project are commensurate to the potential risks;
- Whether the ethical principles outlined in NIST's AI RMF are sufficiently upheld; and
- Whether the project team has taken sufficient action to mitigate the risks of the project relative to the NIST AI RMF.

Based on the information provided in the GAC application, the GAC will make one of three determinations:

1. Approve: The GenAI project is approved using the methods outlined in the GAC application. The project team, led by the PM, is responsible for executing the project according to those methods and reporting any deviations from the methods outlined. The team is also responsible for reporting any adverse events that may have occurred.
2. Request Modification: The GAC may request clarification of the GAC application or modification of the proposed methods described in the application to verify that any

ethical risks are sufficiently mitigated. The GAC will then review the modifications and make a final determination.

3. Reject: Although unlikely, there may be GenAI projects where the GAC feels strongly that the risks associated with the proposed application of GenAI cannot be reasonably mitigated even with modification to project protocols. In those cases, the GAC may reject the proposed application. If so, the GenAI project team is not authorized to proceed on the work without substantial modification to their proposed project and resubmission of that work to the GAC.

The GAC will communicate their decision to the PM. Based on the GAC determination, the team will either proceed with the work as planned, modify their application, or substantially alter the proposed use of GenAI and resubmit.

5.2.3 Project Execution

Once a GenAI project is approved by the GAC, the project team, and particularly the PM, is responsible for ensuring that the project is executed in accordance with the methods outlined in the application. The GAC chair will maintain records of the GAC determination and the associated application from the project team. The PM on the GAC application is also responsible for reporting any amendments or deviations from the proposed GenAI work on the project or any adverse events that occurred as a result of the GenAI use, design, development, or deployment (described in more detail in the next section).

5.2.4 Amendments & Event Reporting

The PM on the GAC application is responsible for ensuring that the project is executed in accordance with the methods outlined on the GAC application. They are also responsible for identifying when the team has deviated from the methods in a way that requires modification to their GAC application. This will not include all changes to the project, only those changes that affect the responses to the questions on the GAC—i.e., those changes that relate to risk relative to the seven NIST principles outlined earlier and elicited on the GAC application.

The PM on the GAC application is also required to report to the GAC if there are any adverse events associated with the ethical risks of the GenAI being used, designed, developed, or deployed on the project. If, for example, there is an issue that involves harm to the end users, loss of privacy, or validity of the results, that event needs to be reported to the GAC. The GAC will then determine if there are any changes that are needed to address the event and to mitigate the risks of those events occurring again in the future. Depending on the nature of the event, the project team may be asked to stop work on the project pending methodological changes. There may be no changes required if the risk is sufficiently mitigated; nonetheless, any of these harms need to be reported to the GAC for consideration. The same members of the GAC who reviewed the initial application will review the event report; however, additional members may be consulted as needed.

5.2.5 Maintenance

In addition to amendments or event reporting, GenAI projects will be reviewed on an annual basis. This annual review will be required to ensure that projects are continuing to follow the methods outlined in their GAC application and that the risks of the project have not substantially changed. This annual review process will again be performed by the original GAC team that

reviewed the application; again, as needed, additional GAC members will be consulted. The GAC chair is responsible for reaching out to the PM on the GAC application to request information for the annual review. This annual review will essentially consist of a project update (i.e., major activities completed) and major outcomes of the work completed to date.

5.2.6 Closeout

Once no additional development, use, or deployment of a GenAI is completed, the project can be closed with the GAC. This occurs at the same time as the project is closed in EPR. This will consist of completion of a summary of the work completed, major outcomes of the work, and (if relevant) any publications associated with the work. For those GenAI projects where the end result of the project will be deployed into an operational environment, the PM will also be asked to provide any updates to the ongoing monitoring or oversight of the GenAI that will be in place to ensure that it continues to function as designed and intended. Once the project closeout is completed, there will be no further GAC review required.

6.0 AI Literacy Training

In order for policies and governance to function appropriately, staff will need to be trained on the ethical risks associated with GenAI as outlined in the NIST AI RMF. In addition, staff who are PMs on projects requiring GAC review will require some additional training. Finally, training will be necessary for the GAC members.

6.1 All Staff

All staff should receive basic training on the responsible use of GenAI. This basic training should outline the potential ethical risks associated with GenAI as well as the principles that PNNL expects staff to hold. The training should also describe the expectations for staff around staff's role and accountability regarding the use, development, or deployment of GenAI. This includes self-assessment of the risks of use of GenAI as well as a general understanding of the oversight processes for GenAI (i.e., the GAC). This training could take many forms, but a scenario-based training like the one used for the business conduct and ethics training is one potentially effective approach.

At minimum, that training should include:

1. Responsible use of the GenAI as it relates to different types of data;
2. Staff accountability for GenAI content;
3. Information for self-assessing the riskiness of their specific uses of GenAI; and
4. Reporting mechanisms for potential ethical concerns or questions.

Training for the responsible use of GenAI, including a self-assessment risk matrix, was recently pilot tested as part of the deployment of the AI Incubator Chat at PNNL, and is described in the next section.

6.2 PM of GenAI Projects

PMs of projects that require a GAC review based on the questions answered in EPR will require some additional training. This training will provide more detail about the need for ethical review process, the role of the GAC, and the expectations regarding the PM's role in the GAC process. This will include an outline of the GAC application process, the event reporting and modification process, and the maintenance requirements, as well as closeout procedures. It will also need to describe the consequences for noncompliance. The training will need to describe the role of the PM in the GAC process and their responsibility for oversight of the project as it relates to ethical risks associated with the project's use of GenAI.

6.3 GAC Members

GAC members will require additional training above and beyond that of staff and PIs/PMs for GenAI projects. In addition to the training already described, GAC members will need to understand:

- Their role in reviewing in the GAC application;

- The review process and the importance of procedural integrity in the GAC review process (e.g., allowing all members to contribute to the discussion);

- The determination process and their role in event reporting, modification, and project closeout.

6.4 Onboarding for Responsible Use of GenAI

As mentioned, the AI literacy training proposed in the last section was recently evaluated as part of the onboarding process for an internal GenAI tool, the AI Incubator Chat. The AI Incubator Chat is PNNL's internal-facing instance of GPT-4/GPT-4 Turbo. The AI Incubator Chat was released on a rolling basis beginning in January 2024. As part of the process, the team developing and maintaining the AI Incubator Chat created an onboarding process. Before receiving access to the tool, users were required to agree to terms and conditions regarding its use. They were also required to attend a 30-minute briefing on the appropriate use of the tool. That training included a description of the tool and its capabilities and limitations and an overview of the types of data that were allowable and prohibited for entry into the tool. Central to the terms and conditions and the onboarding training was user accountability for the content generated by the tool; users were informed that they remained responsible for any of the content generated just as if they had produced it themselves.

The onboarding training also incorporated principles outlined in this document, including an initial training regarding the principles outlined in NIST's AI RMF as they relate to GenAI. In addition, a modified form of the risk assessment matrix was created to allow users to self-assess the riskiness of their own use of GenAI. Modification of the risk matrix was necessary because of the internal-facing nature of the AI Incubator Chat; given that it is an approved, enterprise-wide tool, many of the concerns regarding security were already addressed in cybersecurity review. In addition, considerations regarding privacy were limited to following the guidelines outlined during the onboarding training. The modified risk matrix is presented in Figure 3. The intention is for users to review this matrix and consider their own AI Incubator Chat relative to each listed risk domain. In the training, users were instructed to use the AI Incubator Chat only for low-risk use cases; if they felt that they might be entering into a higher-risk use case, they were asked to contact a member of the AI Incubator Chat team for further discussion.

To determine the efficacy of this training process, AI Incubator Chat users were asked to complete a survey following the training. This survey asked them to rate their knowledge and understanding of ethical and responsible use of the AI Incubator Chat. There were 247 respondents to the survey. A full list of items are presented in Appendix B. Results of the survey are summarized in Figure 4 and Figure 5.

RISK DOMAIN	LOW RISK	MEDIUM RISK	HIGH RISK
VALIDITY & RELIABILITY	The user has good knowledge in the domain where the AI is being applied, and failure of the model can easily be recognized and corrected. There are metrics or methods to understand and verify the validity and accuracy of the AI output.	There are real-world gaps in the user's knowledge and understanding of the AI outputs in the domain where it is being applied, or validity or accuracy of model outputs is difficult to assess.	The user has limited understanding of the domain where the AI is being applied, or there is no way to evaluate the validity of accuracy of the outputs.
FAIRNESS	The content generated will not have a substantial or significant real-world impact on users, recipients, or stakeholders. Any biases can be identified and mitigated as necessary.	The content generated may have a substantial or significant real-world impact on users, recipients, or stakeholders. There are potential biases that may be hard to address.	The content generated will have a substantial or significant real-world impact on users, recipients, or stakeholders. The degree to which bias is embedded in the content is difficult to ascertain or mitigate.
SAFETY	There is very little potential for harm from the use of the AI beyond those that are already present. Any failures are contained and can be managed.	There are potential safety issues from the use of AI in the application that can be easily identified and managed. Failure of the AI would not cause significant harm.	There is a high risk of harm from failures of the AI. Failures of the AI would be hard to control.
TRANSPARENCY	Content generated is reviewed and will not be officially published externally. If published externally, workflows are documented, all content is carefully reviewed, and the ideas expressed are primarily human generated.	Content generated will be part of a formally published external document, and the AI plays a larger role in the creation of the content. Workflows are documented and reported as part of the document.	Content generated will be part of a formally published external document, and the AI plays a significant role in the conclusions, writing, and ideas in the document. There may be no documentation of the workflow used to produce the results.
PRIVACY	Data entered into the system follow all PNNL guidelines and do not include classified, business sensitive, strictly private, or personally identifiable information.	Data entered generally appear to follow PNNL guidelines, but it is not clear if the outputs might develop into business sensitive information, such as intellectual property (IP).	Data may not follow PNNL guidelines for data privacy, and could include classified, business sensitive, strictly private, or personally identifiable information.
SECURITY	All interactions with the generative AI follow PNNL guidance and there are no attempts to circumvent AI restrictions in place for safety, privacy, or security.	Interactions with the generative AI may fail to follow all PNNL guidance, or attempts are made to circumvent some of the restrictions in place for safety, privacy, or security.	Interactions with the AI deliberately attempt to circumvent restrictions in place for safety, privacy, or security.

Figure 3: AI Incubator Chat Risk Matrix

Overall, the results of the survey suggest that people feel comfortable using the AI Incubator Chat and think that their training was sufficient to support responsible GenAI use. They also express desire for clearer guidelines in the future. Conversations with participants in the onboarding sessions suggested that questions remained regarding GenAI disclosure, especially in scientific publications. There were also specific questions about use cases. Future instances of the training process will incorporate scenario-based training to address these concerns.

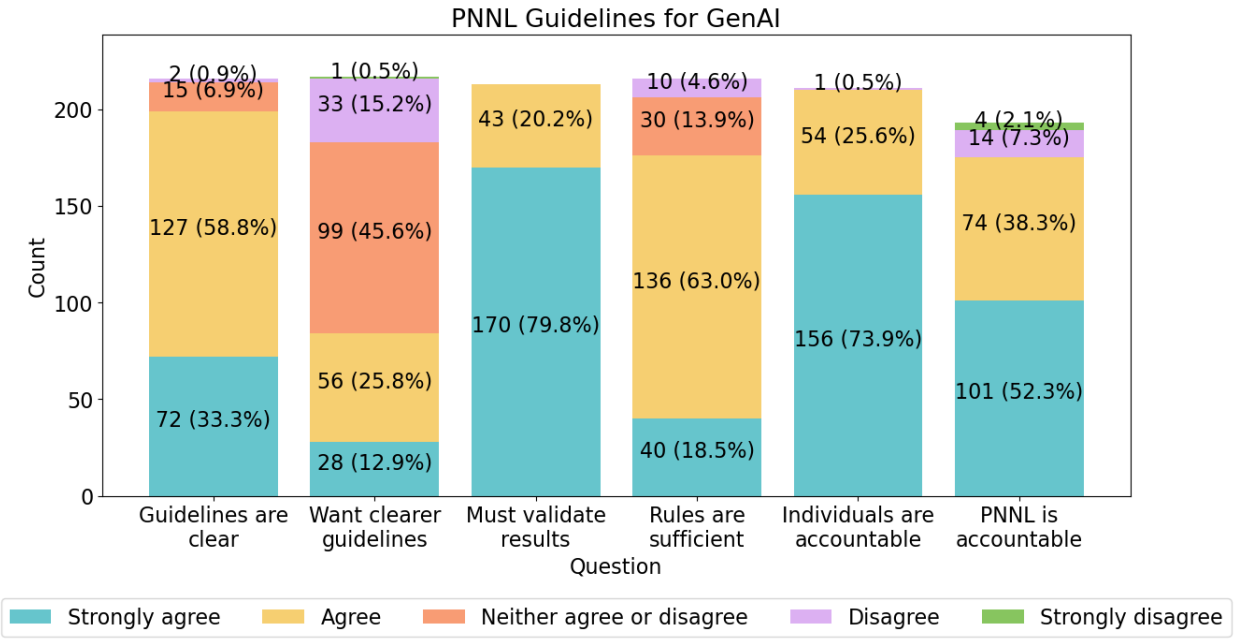

Figure 4: Clarity of PNNL Guidelines

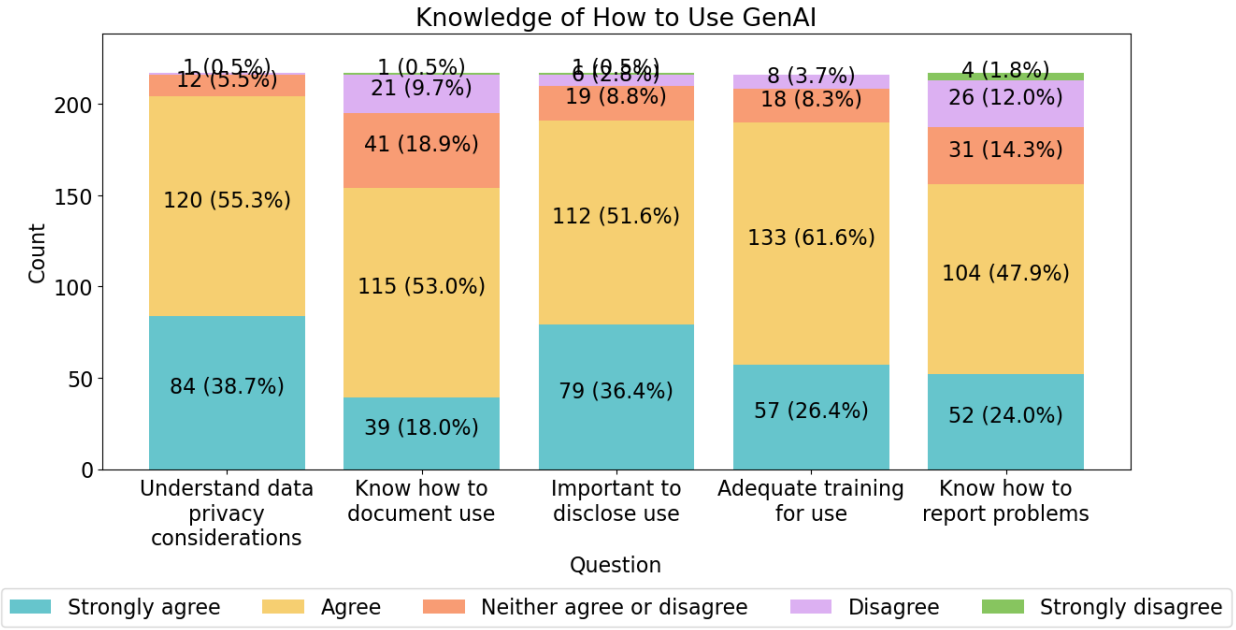

Figure 5: Knowledge of GenAI Use

The onboarding process was generally well-received, and feedback suggested that users felt empowered to responsibly use GenAI. This initial step in addressing concerns around AI literacy can serve as one pillar of the responsible management of GenAI use at PNNL.

7.0 Discussion

The introduction of GenAI into the workplace at PNNL has the possibility to be transformative, but it also presents many ethical, legal, and societal challenges. There is a need to establish mechanisms to manage the new risks presented by GenAI through the three pillars of policy, governance, and literacy.

This paper strives to address those challenges by outlining the foundations for new GenAI policy, through the establishment of consensual ethical principles; providing the basis for a procedure for GenAI governance, through the modification of an IRB process; and describing an example of AI literacy for responsible GenAI use. The principles outlined in this document for responsible, safe, and ethical use of GenAI are presented to provide a basis for future policy at PNNL regarding the use of GenAI. As shown in the AI Incubator Chat onboarding process, one aspect of responsible use of GenAI is providing staff with the tools to understand the principles that must be upheld.

The establishment of an ethical review process is proposed to fill the gap between ethical principles and implementation of responsible GenAI at PNNL. The GAC model outlined in this paper can serve as a blueprint for such a governance mechanism, describing the structure, procedures, and roles and responsibilities necessary to conduct an ethics-based audit of GenAI projects. This model is design to be flexible to adapt to the many types of projects that may be encountered at PNNL and to emerging risks of GenAI.

Finally, AI literacy is a critical component of responsible GenAI use. By providing staff with the knowledge and tools to assess and manage the ethical implications of their GenAI usage, PNNL can support responsible innovation and efficiency with GenAI. The onboarding process for the AI Incubator Chat at PNNL provides a practical example of how AI literacy can be implemented. The survey results presented here further provide a demonstration that such training can be effective in enhancing users' understanding of GenAI and their responsibilities regarding it.

As GenAI continues to evolve, it is critical that PNNL remains proactive in addressing the safety, security, and ethical challenges that it can present. The GAC model presented here, combined with updated policy and AI literacy training, are steps toward creating a framework for the responsible use of GenAI at PNNL.

8.0 Outcomes & Continuing Work

In addition to the ethics-based review of GenAI at PNNL, several other outcomes and opportunities for collaboration were identified during this course of this work. First, to help support increasing responsible AI usage at other Battelle-managed laboratories, the authors of this paper also participated in the creation of a capability accelerator focused on AI literacy. A capability accelerator is a toolkit designed to support other Battelle-managed national laboratories to rapidly implement processes and functions that have been found to be useful and effective at another laboratory. In this instance, PNNL, through its onboarding and training process of the AI Incubator Chat, demonstrated that they were substantially ahead of other laboratories in the creation and deployment of a training program for responsible usage of GenAI. Given that, a capability accelerator was created to describe the approach that PNNL has taken in disseminating knowledge and training regarding the use of GenAI in support of its mission. This toolkit outlined key considerations for a responsible GenAI policy, for the creation of an advisory board focused on AI at the laboratory, and example training materials for deployment at those laboratories. This capability accelerator was presented to Battelle leadership and was approved for additional dissemination and adoption at other laboratories, demonstrating the value of an ethics-based approach to GenAI adoption.

In addition, PNNL had the opportunity to speak and collaborate with researchers at the University of Utah's Responsible AI institute. Following an initial conversation, the overall framework described here was presented to members of the institute at a working meeting. Members agreed that they will review and apply the risk matrix as a test use case for the matrix at another institution. This suggests that the risk matrix has applications even outside of a national laboratory, potentially providing value to other organizations seeking to deploy GenAI in an ethics-centered way. Furthermore, the authors submitted a "Birds of a Feather" session to an upcoming conference (Supercomputing 2024) with University of Utah researchers the operationalization of ethical principles in high-performance and supercomputing applications.

Regardless of the final outcomes of these efforts, there has been an overwhelmingly positive response to ideas presented here, particularly the emphasis on AI literacy and the risk matrix for evaluation of specific GenAI use cases. This underscores the need in the GenAI community for practical, usable tools to implement the principles of Responsible AI in the real world. We hope that the guidance and concepts presented here represent a starting point for others to build additional tools to realize ethical principles for GenAI in a variety of future applications and contexts.

9.0 References

Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., Singh Lubana, E., Jenner, E., Casper, S., Sourbut, O., Edelman, B. L., Zhang, Z., Günther, M., Korinek, A., Hernandez-Orallo, J., Hammond, L., Bigelow, E., Pan, A., Langosco, L., . . . Krueger, D. (2024). Foundational Challenges in Assuring Alignment and Safety of Large Language Models. arXiv:2404.09932. Retrieved April 01, 2024, from https://ui.adsabs.harvard.edu/abs/2024arXiv240409932A

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for Artificial Intelligence and Digital technologies. *International Journal of Information Management*, *62*, 102433. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2021.102433

Atkinson, D., & Morrison, J. (2024). A Legal Risk Taxonomy for Generative Artificial Intelligence. arXiv:2404.09479. Retrieved April 01, 2024, from https://ui.adsabs.harvard.edu/abs/2024arXiv240409479A

Bird, C., Ungless, E., & Kasirzadeh, A. (2023, 29 August 2023). *Typology of Risks of Generative Text-to-Image Models* Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, Montreal, Quebec, Canada. https://doi.org/10.1145/3600211.3604722

Bruschi, D., & Diomede, N. (2023). A framework for assessing AI ethics with applications to cybersecurity. *AI and Ethics*, *3*(1), 65-72. https://doi.org/10.1007/s43681-022-00162-8

Canca, C. (2020). Operationalizing AI ethics principles. *Commun. ACM*, *63*(12), 18–21. https://doi.org/10.1145/3430368

Commission, E. (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment*. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health [Perspective]. *Frontiers in Public Health*, *11*. https://doi.org/10.3389/fpubh.2023.1166120

DHHS. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Washington, DC Retrieved from http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report

Fischer, J. E. (2023). *Generative AI Considered Harmful* Proceedings of the 5th International Conference on Conversational User Interfaces, Eindhoven, Netherlands. https://doi.org/10.1145/3571884.3603756

Floridi, L., & Cowls, J. (2022). A Unified Framework of Five Principles for AI in Society. In *Machine Learning and the City* (pp. 535-545). https://doi.org/https://doi.org/10.1002/9781119815075.ch45

Grady, C. (2015). Institutional Review Boards: Purpose and Challenges. *Chest*, *148*(5), 1148-1155. https://doi.org/10.1378/chest.15-0706

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. arXiv:2307.00691. Retrieved July 01, 2023, from https://ui.adsabs.harvard.edu/abs/2023arXiv230700691G

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

IEEE. (2022). Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems. In.

ISO. (2022). ISO/IEC TS 5723:2022(en). In https://www.iso.org/obp/ui/#iso:std:iso-iec:ts:5723:ed-1:v1:en

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399.

Khan, A. A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., Fahmideh, M., Niazi, M., & Akbar, M. A. (2022). *Ethics of AI: A Systematic Literature Review of Principles and Challenges* Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering, Gothenburg, Sweden. https://doi.org/10.1145/3530019.3531329

Lee, H.-P., Yang, Y.-J., Serban von Davier, T., Forlizzi, J., & Das, S. (2023). Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. arXiv:2310.07879. Retrieved October 01, 2023, from https://ui.adsabs.harvard.edu/abs/2023arXiv231007879L

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: a three-layered approach. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00289-2

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2021). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. In L. Floridi (Ed.), *Ethics, Governance, and Policies in Artificial Intelligence* (pp. 153-183). Springer International Publishing. https://doi.org/10.1007/978-3-030-81907-1_10

Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics*, *3*(3), 869-877. https://doi.org/10.1007/s43681-022-00209-w

NIST. (2023a). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*.

NIST. (2023b). The Language of Trustworthy AI: An In-Depth Glossary of Terms. In

O'Neil, C., & Gunn, H. (2020). Near-term artificial intelligence and the ethical matrix. *Ethics of Artificial Intelligence*, 235-269.

Protection of Human Subjects, § 46 (2017).

Ritchie, K. L. (2021). Using IRB Protocols to Teach Ethical Principles for Research and Everyday Life: A High-Impact Practice. *Journal of the Scholarship of Teaching and Learning*, *21*(1), 120-130.

Rossi, F., Trevino, N., & Ahmed, A. (2022). *Everyday Ethics for Artificial Intelligence*. ibm.biz/everydayethics

P

Paliszkiewicz, J., & Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, *11*(2), 7-30. https://doi.org/https://doi.org/10.15678/EBER.2023.110201

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., & Isaac, W. (2023). Sociotechnical Safety Evaluation of Generative AI Systems. arXiv:2310.11986. Retrieved October 01, 2023, from https://ui.adsabs.harvard.edu/abs/2023arXiv231011986W

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., . . . Gabriel, I. (2022, 20 June 2022). *Taxonomy of Risks posed by Language Models* Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea. https://doi.org/10.1145/3531146.3533088

Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv:2401.11817. Retrieved January 01, 2024, from https://ui.adsabs.harvard.edu/abs/2024arXiv240111817X

Zeng, Y., Lin, H., Zhang, J., Yang, D., Jia, R., & Shi, W. (2024). How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs. arXiv:2401.06373. Retrieved January 01, 2024, from https://ui.adsabs.harvard.edu/abs/2024arXiv240106373Z

Appendix A GAC Applications

Depending on the answers to the above questions, the project team will be routed to either full board or expedited review. They will need to complete the appropriate application for GAC review. The questions here were adapted from information in the NIST AI RMF, IEEE's *Ethically Aligned Design*, and the European Union's Assessment List for Trustworthy AI (Commission, 2020; IEEE, 2022; NIST, 2023a).

A.1 GAC Full Review Application

The PM of the project will be responsible for completing the GAC application in consultation with other team members. Not all questions will necessarily apply to every generative AI (GenAI) project; explaining how and why a specific item does not apply is expected to be part of the process of completing the application. The items included are intended to serve as a prompt for the team to consider at and provide information relating to that potential risk.

A.1.1 Background

1. Provide the background, specific goals, and rationale of the project. If this is a research project, describe the hypothesis and research question you will answer. If it is a non-research project, describe the project purpose and expected outcomes.
2. Describe the overall project plan, including data science methods, data collection/analysis procedures for research projects, or major activities and efforts for non-research projects.
3. Explain the expected benefits and what you hope to achieve with the project.
4. For multi-site projects, list any collaborators and describe the role of each member of the project team.

A.1.2 GenAI System Design

5. Describe the functionality of the proposed GenAI system, including planned use cases.
6. Describe the nature of the data, including training, testing, and validation data, that will be used for development and use of the GenAI system.
7. Determine whether the proposed GenAI will be:
    a. A self-learning or autonomous system: There will be no human oversight or intervention of the AI.
    b. A human-in-the-loop system: Humans will be able to intervene in every decision cycle of the system.
    c. A human-on-the-loop system: Humans will be able to intervene during the design cycle and will monitor the system operation.
    d. A human-in-command system: Humans will be able to oversee the overall activity of the system and determine when, how, and whether to use the AI system and will be able to override any AI decisions.

A.1.3 Data Management & Analysis

8. Describe the measures that will be applied to ensure the data, including training data, used to develop the system is up to date, high quality, complete, and representative of the environment into which the system will be deployed.
9. Describe the measures that will be used to evaluate and document the GenAI system's accuracy.
10. For continual learning systems, explain the mechanisms that will be used to monitor the GenAI system's performance to determine if it is functioning in novel or unanticipated ways.

A.1.4 Security and Safety

11. Discuss the potential security risks of the GenAI system and their mitigations, including the susceptibility to cyber-attacks and the integrity of the GenAI system.
12. Describe any potential risks of the GenAI system to human safety for each planned use case and explain how those risks will be mitigated.
13. Discuss any potential risks of the GenAI system to human safety in the case of faults, defects, outages, attacks, misuse, inappropriate, or malicious use. Describe how the risks will be mitigated.

A.1.5 Accountability and Transparency

14. Explain the traceability of the proposed system and how auditing of the GenAI's decisions, recommendations, or outputs will be handled.
15. For deployed systems, explain the continual monitoring process, performance assessment, and planned avenues for reporting vulnerabilities, risks, or biases by end-users or stakeholders.

A.1.6 Explainability & End User Interaction

16. Outline the mechanisms that will be used to inform end users about the GenAI's purpose and limitations and discuss whether the GenAI's decisions will be explainable to the end users.
17. Specify whether users will know they are interacting with a GenAI and provide justification for any ambiguity.

A.1.7 Privacy and Data Protection

18. Address the handling, storage, and processing of any personally identifiable information (PII) and describe any potential risks to individual privacy as well as how those risks will be mitigated.
19. If the project involves interaction with human subjects or human subjects data, attach documentation of discussion with the Institutional Review Board, either:
    a. A determination of that the project is not considered to be human subjects research; OR
    b. A completed Institutional Review Board Application, with all associated documents and the final determination.

20. For projects involving data on individuals, describe the processes that will be used to allow individuals the right to withdraw consent, the right to object, and the right to be forgotten in the GenAI system.

A.1.8 Fairness & Inclusivity

21. Explain how the data that will be used for training and development of the GenAI system promote fairness and inclusivity of the resulting system.
22. Discuss the strategies and mechanisms that will be used to promote fairness and ensure participation in the GenAI system design and development from a diverse range of stakeholders.
23. Describe the planned user testing and how it will support the inclusion of diverse end-users or subjects.
24. Discuss the mechanisms that will be used to flag bias issues in the GenAI system.
25. Describe the testing procedures that will be used to support system accessibility (e.g., accessibility via screen readers).

A.2 GAC Expedited Review Application

A.2.1 Background

1. Provide the background, specific goals, and rationale of the project. If this is a research project, describe the hypothesis and research question you will answer. If it is a non-research project, describe the project purpose and expected outcomes.
2. Describe the overall project plan, including data science methods, data collection/analysis procedures for research projects, or major activities and efforts for non-research projects.
3. Explain the expected benefits and what you hope to achieve with the project.
4. For multi-site projects, list any collaborators and describe the role of each member of the project team.

A.2.2 AI System Design

1. Describe the functionality of the proposed GenAI system, including planned use cases.
2. Describe the nature of the data, including training, testing, and validation data, that will be used for developing the GenAI system.
3. Determine whether the proposed GenAI will be:
    a. A self-learning or autonomous system: There will be no human oversight or intervention of the AI.
    b. A human-in-the-loop system: Humans will be able to intervene in every decision cycle of the system.
    c. A human-on-the-loop system: Humans will be able to intervene during the design cycle and will monitor the system operation.
    d. A human-in-command system: Humans will be able to oversee the overall activity of the system and determine when, how, and whether to use the AI system and will be able to override any AI decisions.

A.2.3 Data Management & Analysis

4. Describe the measures that will be applied to ensure the data, including training data, used to develop the system is up to date, high quality, complete, and representative of the environment into which the system will be deployed.
5. Describe the measures that will be used to evaluate and document the GenAI system's accuracy
6. For continual learning systems, explain the mechanisms that will be used to monitor the AI system's performance to determine if it is functioning in novel or unanticipated ways.

A.2.4 Security and Safety

7. Discuss the potential security risks of the GenAI system and their mitigations, including the susceptibility to cyber-attacks and the integrity of the GenAI system.
8. Explain how the study presents no more than minimal risk, especially in the case of faults, defects, outages, attacks, misuse, inappropriate, or malicious use.

A.2.5 Accountability and Transparency

9. Explain the traceability of the proposed system and how auditing of the GenAI's decisions, recommendations, or outputs will be handled.
10. For deployed systems, explain the continual monitoring process, performance assessment, and planned avenues for reporting vulnerabilities, risks, or biases by end-users or stakeholders.

A.2.6 Explainability & End User Interaction

11. Outline the mechanisms that will be used to inform end users and stakeholders about the GenAI's purpose and limitations and discuss whether the AI's decisions will be explainable to the end users.
12. Specify whether users will know they are interacting with a GenAI and provide justification for any ambiguity.

A.2.7 Privacy and Data Protection

13. Explain the measures taken to protect the privacy, confidentiality, and integrity of data used for the training, development, and during deployment of the GenAI system. NOTE: For studies involving the use of personally identifiable information (PII), completion of a full board review application is required.

A.2.8 Fairness & Inclusivity

14. Discuss the strategies and mechanisms that will be used to promote fairness and ensure participation in the GenAI system design and development from a diverse range of stakeholders.
15. Describe user testing and the inclusion of diverse end-users or subjects, mechanisms to flag bias issues, and testing procedures that will be used for system accessibility (e.g., accessibility via screen readers).

Appendix B Survey Questions

1. What is your job family at PNNL?
2. Rate your level of knowledge in the following areas:
    a. Machine learning techniques, like neural networks or random forests.
    b. Ethics for responsible use of AI systems.
3. How often do you use AI chat technologies, such as ChatGPT, Claude, etc., either at work or personally?
4. The statements below all relate to PNNL's guidelines around using the AI Incubator Chat for your work. Rate the extent to which you agree or disagree with the following statements.
    a. The guidelines for using the AI Incubator Chat are clear and understandable.
    b. I would benefit from clearer guidelines for appropriate use of the AI Incubator Chat.
    c. I think PNNL's rules for AI Incubator Chat use are sufficient to prevent misuse.
    d. I understand the data privacy considerations and rules that apply to inputting data into the AI Incubator Chat.
    e. If I use the AI Incubator Chat in my work, I know the proper procedures for documenting and reporting details on how the AI Incubator Chat was applied.
    f. I understand the importance of disclosing the AI Incubator Chat assistance in authorship and related tasks.
    g. I have received adequate training to understand and adhere to ethical AI practices at PNNL.
    h. I know the appropriate channels to report a concern if I discovered evidence that an AI was causing harm or had unintended bias.
5. The next statements are related to accountability for AI outcomes. Rate the extent to which you agree or disagree with the following statements.
    a. Any decisions or predictions relying on the AI Incubator Chat outputs should first go through human validation.
    b. Individual users are responsible for any content generated by the AI Incubator Chat and used as part of PNNL work.
    c. PNNL is accountable for any content generated by the AI Incubator Chat and used as part of PNNL work.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov