

Developing a framework for addressing ethical challenges in generative AI

Simon Lucas, René M. Heinitz, Sarah J. Becker & Jean Enno Charton

To cite this article: Simon Lucas, René M. Heinitz, Sarah J. Becker & Jean Enno Charton (11 Sep 2025): Developing a framework for addressing ethical challenges in generative AI, Journal of Information Technology Case and Application Research, DOI: [10.1080/15228053.2025.2558443](https://doi.org/10.1080/15228053.2025.2558443)

To link to this article: <https://doi.org/10.1080/15228053.2025.2558443>



© 2025 Merck KGaA, Darmstadt, Germany.
Published with license by Taylor & Francis
Group, LLC.



Published online: 11 Sep 2025.



Submit your article to this journal [↗](#)



Article views: 852



View related articles [↗](#)



View Crossmark data [↗](#)

Developing a framework for addressing ethical challenges in generative AI

Simon Lucas^a, René M. Heinitz^b, Sarah J. Becker^b, and Jean Enno Charton^a

^aMerck KGaA, Darmstadt, Germany; ^bUniversität Witten/Herdecke, Witten, Germany

ABSTRACT

Technology developers in corporate environments hold responsibility as technologies transition from laboratories to market adoption, particularly with Generative Artificial Intelligence (GAI). This transformative technology presents substantial opportunities alongside profound ethical challenges that require immediate attention. As companies seek practical guidelines before regulatory frameworks are established, ethically informed self-regulation becomes essential for aligning business interests with ethical standards. In response, we have developed an actionable framework to promote responsible innovation in GAI. This framework emphasizes defining objectives, maintaining documentation, conducting bias and fairness assessments, implementing security measures, and fostering active engagement. By adhering to these principles, the framework aims to facilitate the ethical integration of GAI into science and technology operations, in order to mitigate potential harms while harnessing its transformative potential. Our bottom-up, self-service approach empowers use case owners to make ethically informed decisions, enhancing ethical capacity across the workforce and increasing awareness of critical considerations in data and GAI. Ultimately, frameworks like the GAI Ethics Framework are important tools for navigating the complex moral landscape of GAI, fostering responsible innovation, and building trust among stakeholders, thereby complementing existing regulations such as the EU AI Act.

Introduction

Generative Artificial Intelligence (GAI) has emerged as a groundbreaking technology across various domains, from creating compelling visual art and generating human-like text to advancing drug discovery and personalized education (Bian & Xie, 2021; Ruiz-Rojas et al., 2023; Su & Yang, 2023; Zeng et al., 2022). This technology holds immense promise, enabling innovative applications that previously required significant human effort and were considered purely theoretical. However, alongside these advancements, GAI introduces profound ethical challenges that demand urgent attention and resolution. As AI and in particular GAI systems become increasingly sophisticated and integrated into daily life, the potential for misuse, bias perpetuation, privacy violations, and other ethical issues substantiates (Alkaissi & McFarlane, 2023; Fischer et al., 2023; Hall & Haas, 2021; Li et al., 2024; Sun et al., 2022; M. Zhou et al., 2024).

One ethical concern surrounding GAI, for example, is its potential to generate misleading or harmful content, such as deepfakes and misinformation (Ferrara, 2024; Xu et al., 2023; J. Zhou et al., 2023). These outputs can blur the lines between reality and fabrication, posing significant threats to public trust and safety. Additionally, the biases embedded in training data can be inadvertently amplified by GAI models, leading to discriminatory outcomes and reinforcing societal inequities (Alles, 2024; Capraro et al., 2024; Jadon, 2024). Furthermore, the use of personal data in training these models raises critical issues related to privacy and data rights, necessitating stringent measures for data protection (Novelli et al., 2024; Veluru, 2024; Xiang, 2024).

Despite the recognized importance of these ethical issues, the current discourse often lacks a practical framework for addressing them. While existing guidelines and principles are valuable, their effective implementation in real-world applications presents significant challenges. (Floridi 2019; Mittelstadt 2019; Morley et al. 2020; Hickok 2020; Blackman 2020; Mökander et al. 2022). There is a growing need for a robust, actionable framework that developers, policymakers, and stakeholders can utilize to navigate the

complex ethical landscape of GAI. Such a framework should arguably be grounded in transparency, accountability, and public engagement, ensuring that the development and deployment of GAI technologies maximize benefits while mitigating risks (Kenthapadi et al., 2023; Mökander et al., 2024; Zlateva et al., 2024).

This paper aims to contribute to the following research question: “How can an assessment framework be developed to facilitate the ethical integration of GAI in corporate environments while enabling practitioners in bottom-up innovation?” In order to address this question, a practical framework is proposed for the resolution of ethical issues in GAI, with particular emphasis on those arising in technology companies. Technology developers, especially in a corporate context, bear responsibility as technologies that advance beyond the lab stage can have a profound impact once they are successfully marketed and adopted by a large user base; all while making decisions in dynamic technological fields often characterized by regulatory uncertainties and moral complexities. By exploring key ethical challenges and providing concrete guidelines for addressing them, this framework seeks to foster responsible innovation in GAI within the corporate domain. The proposed approach emphasizes the necessity of clear purpose definition, thorough documentation, bias and fairness audits, security measures, and active public engagement. Through these principles, the framework aspires to support the ethical integration of GAI into operations of science and technology companies, safeguarding against potential harms while unlocking its transformative potential.

Situation at Merck

Merck KGaA, Darmstadt, Germany, a leading industry player in the healthcare, life sciences, and electronics sectors, has a longstanding tradition of integrating ethical deliberation into its business practices. This tradition is exemplified by the company’s decade-long engagement with a team of internal ethicists and external advisory bodies. The latter have provided independent guidance on ethically sensitive issues in healthcare and life sciences, including stem cell research and genome editing, which have subsequently been incorporated into various company charters and principles (see e.g., Sugarman et al., 2018).

In response to the ethical challenges brought by digitalization, Merck has proactively issued a digital ethics mandate. This mandate reflects the company’s focus on managing data, algorithms, and AI with ethical diligence, addressing the specific ethical risks associated with these technologies. It also aims to responsibly leverage commercial and scientific opportunities that emerge with the advent of such technology. This includes the implementation of ethical frameworks specifically designed for the tech industry and the unique needs of Merck (Becker et al., 2022), integrating ethics panels to guide corporate ethical deliberations (Nemat et al., 2023), and tackling unique ethical challenges specific to data and AI. The company has further addressed the specialized needs in the digital realm by deploying standardized ethical assessment tools tailored for specific use cases like data analytics and AI use. These tools contribute to bridging the gap between high-level ethics frameworks and their practical application, thus attempting to make them actionable within the organization. These approaches include customizing ethical strategies to address challenges such as managing diverse use cases, complex project infrastructures, and the ethics expertise gap within data and AI teams (Charton et al., 2025).

In response to the rapid evolution of generative AI technology and its widespread applications, Merck has instituted a set of ethical guidelines for generative AI usage, which is one of several tasks managed under the oversight of Merck’s Global Data & AI Organization. These guidelines are designed to align GAI applications, such as ChatGPT, with the company’s Code of Digital Ethics (CoDE), an ethics framework based on 20 action-guiding principles (Becker et al., 2022). Protocols have been established to promote responsible use, delineating specific do’s and don’ts. Building on this foundation, Merck’s Global Data & AI Organization has developed MyGPT, a custom-built internal version of ChatGPT, vetted by Merck’s cybersecurity and data privacy teams and endorsed by the worker’s council, designed to meet operational and ethical standards that support the activities of Merck and its employees, enabling them to learn and leverage the productivity potential of AI early on. This initiative forms part of a broader strategy to expand the company’s data and analytics capabilities across various sectors and regions, facilitated by the harmonized data and analytics ecosystem known as UPTIMIZE, and reinforced by robust data governance. The strategy aims to cultivate a strong data and AI culture within Merck, encouraging every employee to actively create and share data, access data and analytics, and develop the knowledge and mind-set necessary to

leverage these resources effectively for business management and collaborative purposes. This includes enhancing digital literacy and competencies among employees through targeted tools and training sessions that emphasize ethical AI usage and digital proficiency. This approach reflects Merck's attempt to foster a workforce that is both technologically proficient and ethically informed.

Beyond that, Merck has launched various initiatives utilizing generative AI, ranging from insights and content generation to interactive customer or patient-facing applications across all its business sectors. Each of these applications raises different ethical questions that require careful consideration. Some of these applications may operate well within general guardrails to promote bottom-up innovation while avoiding the need for formal oversight. However, others may require dedicated oversight procedures or should be avoided altogether. To deal with an ethical assessment of these cases, a practical and scalable solution is needed. This is especially salient, as the company-wide AI governance structure – necessary to comply with the EU AI Act regulations among others – is complex and requires time and careful review to implement. To bridge the gap between having no oversight and a fully established AI governance framework, a practical interim solution for practitioners is needed.

Further practical complications arise from the volume of generative AI use cases that hail from diverse business sectors, each having different R&D priorities, managerial procedures, and organizational setups, which poses significant difficulties for thorough and individualized ethical assessments. Since the task of performing individualized ethics assessments by Merck's Digital Ethics Office or eliciting in-depth advisory consultations by Merck's Digital Ethics Advisory Panel (DEAP) becomes comparably complex, in principle, managers of GAI projects should be enabled to assess ethical questions surrounding their use cases on their own. In the current era of rapid digital advancement, it is important for organizations to foster a culture of ethical responsibility among their employees. By doing so, they can alleviate the burden on project managers and facilitate digital innovation and bottom-up experimentation, without the hindrance of overly complex governance structures or time-consuming ethical assessments. On the other hand, it is crucial to detect cases that may pose significant ethical risk and subject them to a rigorous assessment regarding potential risk by making use of the full expertise of Merck's ethics capacities, such as the Digital Ethics Office, or the DEAP. All these activities also ought to be aligned with trust-building considerations enhanced with internal and external stakeholders. Concerns of employees should be integrated into these considerations, such as those surfaced by feedback from speak-up lines and internal surveys on AI, addressing issues about potential job elimination, or control of AI within the workforce.

Merck's goals for developing a generative AI ethics framework

Based on the considerations outlined above, Merck has developed a methodology intended to ensure the responsible deployment of GAI technologies. This initiative, commissioned by Merck's Global Data & AI Organization, aims to help use case owners leverage GAI and foster innovation and experimentation while adhering to the company's commitment to digital ethics, specifically by interpreting Merck's CoDE into appropriate oversight. In doing so, it categorizes use cases into three levels of oversight requirements, facilitating targeted governance that evolves from self-assessment to more structured regulatory compliance as necessary. This tiered oversight approach is intended to promote experimentation within the organization without the constraints of immediate formalized oversight. Recognizing the dynamic nature of generative AI technologies, the framework is set out to identify use cases that can be implemented without further oversight, those requiring dedicated oversight procedures, and those to be avoided altogether.

The framework is designed specifically to guide practitioners at Merck through the ethical complexities of their projects, promoting responsible decision-making in anticipation of more formalized governance structures influenced by emerging regulations, such as the EU AI Act. It adopts a bottom-up, self-service approach to ethics, empowering use case owners to make ethically informed decisions. In doing so, it builds ethics capacity across the workforce, fostering awareness and understanding of critical ethical considerations related to data and GAI. The model supports a progressive escalation of governance, ensuring that while innovation is encouraged, it remains within ethical guidelines and corporate responsibilities outlined in Merck's CoDE. Moreover, the implementation of the framework is consistent with the prevailing academic discourse that ethical guidelines such as the CoDE must be translated into practice to provide a valuable impact for organizations (Blackman, 2020; Floridi, 2019;

Hickok, 2020; Mittelstadt, 2019; Mökander et al., 2022; Morley et al., 2020). The development process of the framework follows important design science research guidelines, in particular by designing a viable method to approach real business problems (Hevner et al., 2004). Informed by direct feedback from Merck employees, including informative internal surveys on key envisioned use cases and anticipated ethical challenges, as well as focused interviews with project managers, decision-makers, and department heads, the framework ensures that the application of GAI aligns with corporate ethics standards and meets the current needs of GAI users within Merck. It also addresses both inherent and contextual risks associated with specific use cases and emerging regulations while taking external developments into account. Ultimately, Merck has integrated seven “ethics markers” into this framework, creating a decision tree that categorizes generative AI applications into three governance levels, thus balancing innovation with ethical accountability. In what follows, we will outline the considerations that led to the architecture, design, and practical operational characteristics incorporated into the framework.

Development of a generative AI ethics framework

Methodology

A substantial theoretical foundation underpins the ethical discourse of GAI or AI more broadly (Al-Kfairy et al., 2024; Ferrara, 2024; Floridi, 2012; Floridi & Taddeo, 2016) yet increasingly concentrates on bridging the “principles to practice gap” (Corrêa et al., 2024; Georgieva et al., 2022; Mittelstadt, 2019; Morley et al., 2020). Various practical proposals have emerged, such as checklists for digital ethics evaluation (Deon, 2018; UN Global Pulse, 2019), staff training modules (Danish Design Center, 2021; Omidyar Network, 2018), formal ethics panels (Nemat et al., 2023; Schuett et al., 2023), and organization-wide certification or audit procedures (AI Ethics Impact Group, 2020; Epstein et al., 2018; Saleiro et al., 2018). Within this discourse Mökander et al. (2023) classify various approaches into three categories: the *Switch*, a binary assessment of whether a system qualifies as AI and thus requires specific risk measures; the *Ladder*, which ranks systems by their ethical risk; and the *Matrix*, a multidimensional model incorporating context, data input, and decision structure. While our own framework is not strictly following an implementation theory, it aligns with the *Ladder* approach by focusing on systems’ ethical risk profiles. In this sense, it offers a middle ground between purely binary conceptions and more expansive, matrix-oriented perspectives.

It is important to recognize that ethical risks in GAI manifest across multiple levels of analysis, which are relevant is such a *Ladder* approach. Ethics of GAI pertain not only to the overall objectives of the technological solution but also to each component within the solution. For instance, a GAI system designed based on sound ethical deliberation can still be misused for unethical business purposes. Therefore, it is essential to thoroughly address the various layers of risks and concerns associated with its use. First, inherent risks of GAI include issues such as inaccuracy, hallucinations, explainability challenges, and bias (Alkaissi & McFarlane 2023; Li et al. 2024; Sun et al. 2022; M. Zhou et al. 2024). Second, there are risks related to specific use cases or clusters of use cases, such as unethical applications, the need for human verification and oversight, and the handling of confidential or sensitive data (Al-Kfairy et al., 2024; Britton-Heitz & Merhout, 2024; Chen & Esmaeilzadeh, 2024; Thorne, 2024). Finally, broader socio-technological impacts must be considered, including the effects on the workforce, skills and capabilities, and the transformation of communication and behavior (Bargh & Troxler, 2020; Feldstein, 2023; Obrenovic et al., 2024).

Likewise, the assessment of GAI and its respective frameworks can operate on at least two distinct levels of abstraction (see Figure 1). The first is the use case logic, in which a user inputs a query, and the generative AI model produces a response. The second is the system logic, which encompasses the developers of GAI applications and their determinations regarding model development, the reuse of data generated during its operation, and the associated issues that arise from this process. In developing Merck’s GAI Ethics Framework, we, following the design principles commissioned by Merck’s Global Data and AI Organization, initially prioritized the use case logic for three key reasons: (1) to provide users of GAI solutions with a rapid support service; (2) to address the imbalance between users and developers of GAI solutions; and (3) to empower use case owners to address ethical questions within their specific applications.

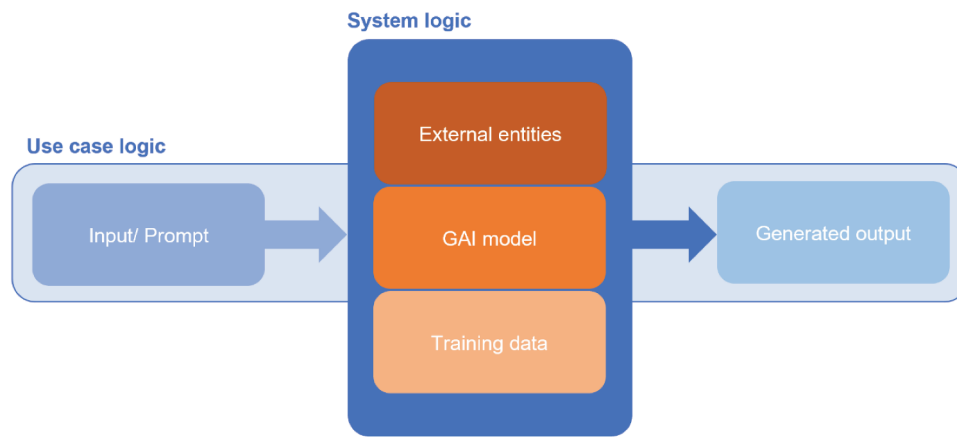


Figure 1. The use case and system logic of generative AI.

As users engage within the use case logic paradigm, designed to operate autonomously and at a speed that fosters competitive innovation, a critical need exists for ethical assessments in the absence of formalized oversight.

For the reasons discussed above, our decision to shift from a comprehensive approach to a purely use-case-driven logic yields an assessment framework that is effectively algorithm-agnostic. It is by no means confined exclusively to GAI, nor is it limited to a particular subset of AI systems; rather, it can be applied to a broad array of AI use cases in principle. Nevertheless, we maintain that the core assumption – namely, that GAI is widely accessible to a diverse range of users via a prompt-output modality – remains crucial. In particular, this feature renders the framework especially valuable for evaluating user-centric AI applications in which numerous projects are overseen by managers who may possess only a rudimentary understanding of the categories subsumed under a system logic. In the context of a framework’s design, these considerations are especially salient. For example, we have previously developed ethics assessment methodologies for AI applications in other domains, such as data analytics (Charton et al., 2025) and when devising vendor-assessment criteria for AI systems (unpublished data). In those cases, contextual factors such as project throughput, the technical expertise of personnel, and the overall level of ethical significance often necessitated moving beyond user-centric concerns. Specifically, system-level factors like robustness and model components took on heightened importance to ensure that the assessment was adequately comprehensive for each department’s needs.

We acknowledge that GAI governance at the system level is equally important. However, we believe ethical considerations extend beyond the purview of individual project managers and use case owners. Provisions for the ethical design of GAI must be established at the oversight board level. These boards make decisions regarding issues such as AI sourcing and model robustness, necessitating either executive authority or specialized technical expertise. Nevertheless, we also provide assistance to developers working on GAI solutions, which will be discussed in more detail later in this article.

The initial stage of the GAI Ethics framework development process was to determine whether the principles of the CoDE could be applied to create a practicable framework in the rapidly evolving field of GAI. In order to gain insight into this question, an analysis was conducted of the content of each principle, which was then compared to the current issues of the prevailing discourse on GAI. The description of the CoDE principles was thus calibrated to align with both recurrent findings in the relevant literature and empirical knowledge gathered at Merck. The latter included surveys conducted among the data and AI community, which highlighted anticipated use cases predominantly in the areas of coding, insights retrieval, competitor analysis, legal-document processing, and communications. These findings were further reinforced by analyzing ongoing projects across various business sectors and departments, where the principal objectives centered on insights generation, content creation, and interaction. Likewise, a key motivation identified through these surveys was the desire to boost productivity, enable customization, and reduce the burden of repetitive tasks. When asked about ethical concerns, data and AI practitioners primarily identified the potential impact on the workforce and the risk of hallucinations as key issues, followed by

concerns regarding harmful outcomes from erroneous outputs, data and privacy breaches, and systemic bias. The confluence of these factors was analyzed by the Digital Ethics team, then categorized and assigned to the most relevant principles on the basis of a consensus vote. Specifically, four pivotal principles have emerged as particularly salient within the GAI landscape, mirroring challenges frequently articulated by Merck personnel and extensively deliberated in public and academic discourse. These include the expertise to classify the output of a prompt (related to the literacy principle) (Denny et al., 2024; Romero Moreno, 2024), the use of personal data in a prompt (proportionality) (Popowicz-Pazdej, 2023; Wach et al., 2023), awareness of the limitations of the generative AI application (comprehensibility) (Annapureddy et al., 2024; Kox & Beretta, 2024), and the use of other sources to verify the output of the generative AI application (responsibility) (Preiksaitis & Rose, 2023; Thorne, 2024). The proposal to focus on this subset of principles was agreed upon in alignment with Merck's DEAP in one of its panel meetings, which emphasized the importance of meaningfully translating these principles into markers. This approach allows project characteristics specific to GAI use to be reflected back into their impact on the subset of principles and, consequently, on Merck's CoDE. These four aspects hence represented a preliminary stage of the framework's development. However, the incorporation of additional so-called "ethics markers" was necessary to create a user-friendly GAI Ethics framework. The ethics markers act as a link between the high-level principles of the CoDE and the specific characteristics of a GAI project. In other words, the principles are translated into measurable and observable indicators within the ethics markers. The implementation of previous projects employing ethics markers, coupled with the development of them and their precursors by scholars in digital ethics research, facilitated the collection of a wide range of ethics markers. As a general observation, ethics markers may be collected and examined in diverse contexts to determine their usefulness. Initially, we compiled a list of 16 ethics markers (including e.g., questions about users' technical awareness, questions about the use of personal data or questions about potential negative effects on people) drawing from prior research (Charton et al., 2025) and theoretical frameworks pioneered by others (AI Verify Foundation, 2024; Nemat et al., 2023; OECD, 2022; SAP, 2024; Zlateva et al., 2024).

Following this compilation, we conducted use case testing involving 15 scenarios to identify the most relevant ethics markers. This process included testing various iterations of the framework across different contexts within the company, covering GAI applications from operations and supply chain planning to commercial and research applications (examples of use cases included the generation of tailored patient treatment plans and the creation of marketing e-mails for customers). These use-case tests were conducted as guided interviews, during which the Digital Ethics team sought to reconstruct each use case and subsequently applied a step-by-step analysis aligned with the decision tree and the queries related to the provisionally selected markers. At each juncture, the markers were examined for their substantive relevance and terminological clarity, as well as for whether the questionnaire might have overlooked any ethically significant information. Proposals and comments provided by the respective use-case owners were documented and subsequently discussed within the Digital Ethics team for possible integration.

For instance, a common piece of feedback was that certain markers in the decision tree needed clearer and narrower definitions, such as personal data, sensitive personal data, and business-sensitive data. Descriptions of these markers should include illustrative examples to facilitate easy contextualization for their respective use cases. Additionally, test users consistently recommended establishing clear links between interconnected markers in the logic of the decision tree. For instance, the relationship between required user expertise and the necessity for output verification should be explicitly defined.

Further refinement was achieved by discussing the markers and the language used throughout the tool with the DEAP, ensuring that the markers and descriptions are not only user-friendly but also philosophically sound. For example, the expert panel recommended adding additional layers to categorize the type of data used in GAI use cases during marker selection. This includes distinguishing between personal data that Merck has control over, and personal data sourced externally, as well as differentiating between consented and unconsented personal data. The panel also specifically suggested refining the marker "negatively affects humans" in the classification methodology, as its initial iteration was too generic for meaningful use. While some individuals or groups may consistently be negatively affected, it is essential to consider other ethical principles that may outweigh this detriment. To ensure a more comprehensive approach, Merck and others developing such a framework should incorporate a risk-benefit perspective or specify certain risks for individuals or groups that are deemed problematic,

leading to a high-risk classification. This will facilitate a more nuanced assessment of the impact on humans. The final selection was contingent upon user feedback, with a special emphasis on feasibility and comprehensibility criteria, ensuring accessibility for non-experts during the early stages of use case exploration.

Taken together, within this use case testing phase, a succinct set of ethics markers was chosen, guided by three primary criteria formulated through extensive pilot testing and iterative feedback loops with the DEAP. First, user-centricity was prioritized, ensuring alignment with the use case logic, given the varying levels of expertise among employees in the GAI model. Second, practical guidance was emphasized to equip users with actionable insights, furnishing them with relevant information regarding considerations when employing a GAI application. For example, when personal data is used in a GAI application, users are informed that they should consider whether the data provider has given consent for the use of their data. Finally, ease of application was prioritized, necessitating a yes-no structure to facilitate straightforward determination. For example, individual employees are usually not capable of evaluating whether the use of an application results in excessive CO2 emissions.

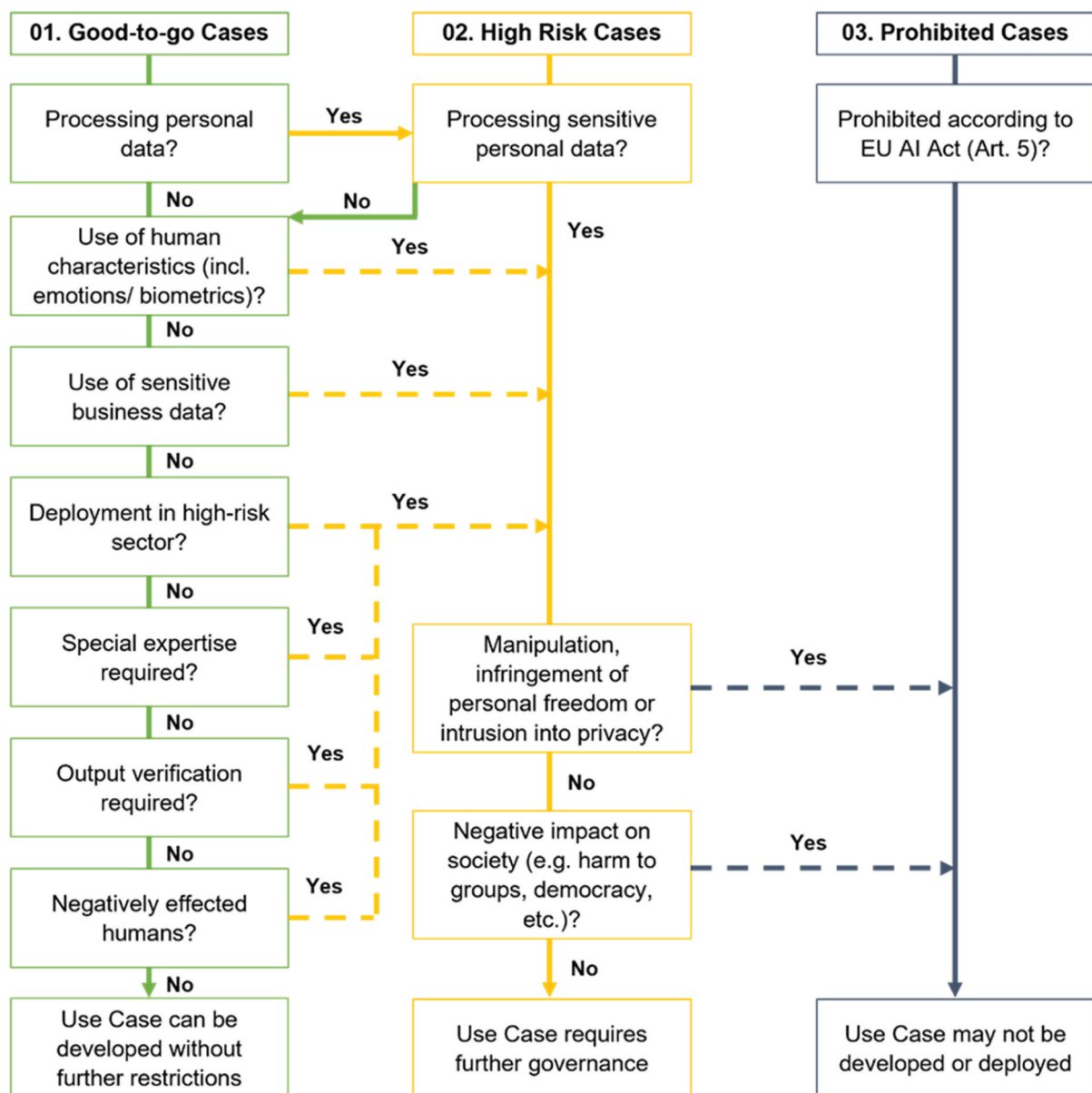


Figure 2. The generative AI use case classification Tool.

The seven ethics markers were subsequently arranged in a linear decision tree (see [Figure 2](#)), enabling a sequential yes-no assessment by use case owners. These markers cover the following areas and questions: 1) Sensitivity of data: Does the use case involve using personal information in the prompts for the AI, such as a name, an identification number, location data, an online identifier, or one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of a natural person? 2) Use of human characteristics: Are the prompts inclusive of any specific human traits such as gender, age, ethnicity or emotions? 3) Sensitive business data: Are any confidential sensitive business specifics part of the prompts? 4) High-risk area: Is the AI output from the prompts intended to directly address critical areas, e.g. participants in the healthcare ecosystem (especially patients or caregivers) or Merck employees in HR contexts? 5) User expertise: Does the team need a role with the necessary expertise to understand the AI's functionality, recognize its limitations, and interpret the relevance of its outcomes within the specific use case? 6) Verification process: Should there be a process in place to verify the AI's output from the prompts using different sources? 7) Impact on humans: Might the deployment of the AI's output from the prompts lead to potentially adverse effects on people or have extensive, negative societal consequences?

If the answer to all questions is no, the case is classified as a “good-to-go” case that requires no additional oversight. If any of the questions are answered with yes, indicating a high risk, further actions are required. In order to determine if the use case might fall under the “prohibited category,” two additional questions need to be considered: 8) Does the application lead to the manipulation or the infringement of personal freedom or intrusion into privacy? 9) Does the application have a negative impact on society? These two questions are directly related to Article 5 of the EU AI Act. The article sets out eight characteristics, which, if present in an AI system, classify it as prohibited. These characteristics include for example the ability to influence people subliminally, the use of facial recognition, or the use of emotion analysis.

All of these questions are detailed in a handbook and backed by examples to explain the markers and questions accordingly. This “handbook-like” style of the framework was developed in collaboration with the data and AI community, alongside the senior leadership in the relevant business unit, who selected it as the preferred method of deployment. It was conceived as a self-service interim tool aimed at enabling individual use and self-assessment, offering a way to familiarize users with ethical questions in the context of generative AI while swiftly and proactively addressing potential issues within each unit. At the same time, it prepares for the eventual formal oversight necessary to meet emerging regulatory requirements.

During deployment, a handbook was provided to all AI practitioners, offering a general overview of ethical considerations in the context of generative AI, as well as detailed user guidelines to facilitate individual use and self-assessment. This includes instructions regarding how to proceed in high-risk situations, such as conducting a self-assessment using the annex of the handbook or, if the issue remains unresolved, consulting Merck's Digital Ethics team, and cases of prohibitive risk, warranting immediate referral to both management and the ethics department. In addition to the decision tree depicted in [Figure 2](#), the handbook offers a more comprehensive explanation of each marker's significance in relation to Merck's CoDE. It also provides examples highlighting specific situations in which certain markers would be activated, thus enabling users to identify and address potential risks more effectively.

The handbook's design followed a rationale similar to that underlying the selection of the markers and the decision tree described above. Specifically, the terminology and examples in the marker descriptions were chosen on the basis of extensive user feedback, ensuring they would be both understandable and relevant across various departments. This content was then refined with input from the DEAP to preempt misconceptions and provide additional clarity, for instance, by contextualizing what is meant by the abstract notion of “negatively affecting humans” within the company's specific setting. The same rationale applies to the handbook's self-assessment provisions, which draw heavily on prior frameworks, such as those developed for evaluating AI in data analytics (Charton et al., 2025). The guiding questions for high-risk scenarios were intentionally designed not as clear go/no-go checkpoints. On the one hand they are prompt to encourage reflection on possible precautions and the extent to which potential risks have been sufficiently addressed, on the other hand they include measures that the use case team should implement to minimize risks. In this way, the handbook aims to raise awareness of the ethical dimensions that may arise in specific project contexts and supports the use case team in finding measures to mitigate them.

For example, if the use case does not pose “prohibitive risk,” further provisions for “high-risk” use cases apply and use case owners are referred to a self-assessment guide provided in the handbook.

This guide is designed to prompt a more detailed reflection on the respective use case. Managers should consider discussing any potential issues with their team or manager. For each category, the handbook provides guiding questions to explore some of the complexities associated with the markers.

To illustrate, if a user determines through the framework that the use case necessitates a verification of the output of the GAI application, the use case will require further questioning. The self-assessment manual in the handbook assists AI practitioners in conducting further questioning of their use case and identifying measures to mitigate potential ethical risks. If the answer to the question “Should there be a process in place to verify the AI’s output from the prompts using different sources?” is yes, the user is additionally asked whether a verification process for the output of the use case is being developed. In the event that this is the case, no further action is required on the part of the user. Conversely, should this not be the case, the user will be required to provide further elaboration on the verification process. Additional guidance is provided to assist the user in this process. In this case, users are encouraged to take measures such as implementing a documentation process for verification and selecting a person or digital solution to verify the output in future. The guidance provided in the self-assessment manual is supposed to be utilized as a point of reference for discussion with colleagues and managers. If a critical ethical issue persists throughout the project lifecycle, the framework mandates engagement with Merck’s Digital Ethics team, with the option to seek external consultation through the DEAP.

Given the intended grassroots approach to enabling thoughtful risk-taking, the measures prescribed in the handbook are deliberately minimal, except in cases that are explicitly prohibited. They provide that high-risk scenarios should be discussed within the relevant management lines, aided by the handbook’s guiding questions, with the optional involvement of the Digital Ethics team if necessary.

The implementation of the framework

Our framework addresses pressing needs within various departments at Merck concerning GAI, serving as a grassroots ethics initiative driven by Merck’s official goal implement to digital ethics, strengthened by institutional efforts. Central to this initiative is fostering a data and AI community and culture to help propel grassroots engagement, alongside raising ethics awareness company-wide – from the CEO and middle management to individual data scientists. By incorporating active leadership communication and support, the framework aims to embed ethical principles in everyday decision-making.

This grassroots initiative is meant to complement, not replace, more formalized processes that require tailored scientific analysis and rigorous organizational collaboration to address specific departmental needs and legal requirements. During development, we focused on an “active enablement over control” approach (Obwegeser et al., 2020). Our framework serves as a provisional “plug and play” governance mechanism designed to bridge the gap while awaiting the establishment of a comprehensive oversight body. This future body would address additional overarching aspects of GAI, including those encompassed by the “system logic.” The experiences gained from the current project’s self-governance approach can also inform the development of this overarching framework.

The rationale behind focusing on the “use case logic” at this stage is to provide use case owners with the tools to address ethical issues within their specific applications. Ethical concerns related to the “system logic” will be addressed separately within the oversight body and the developer teams of GAI solutions, following a methodology aligned with the systematic approach outlined above. Given the larger number of users compared to developers, more hands-on assessment methods seem essential for users. The smaller number of projects within the “system logic,” coupled with their potentially far-reaching impact, allows for more in-depth ethical analysis by Merck’s ethics team and formal ethics review processes, such as through the DEAP. At present, developers are provided with a questionnaire comprising questions such as: (1) Are the prompts or outputs shared with any individual or entity other than the prompt user, for instance, for further training of the model? (2) Are the model’s limitations clearly communicated to the user? (3) Is the system robust against malicious attempts to influence its output, such as prompt injection? (4) Is the system continuously updated with new data to identify ethical issues? This section of the framework will be the subject of further elaboration in the future, particularly in consideration of the guidelines for general-purpose AI models under the EU AI Act.

How to unlock the value and general applicability of the framework

The structure of our framework follows a linear decision tree with simple questions that classify use cases into three categories. “Prohibited cases” that will be not developed, “high-risk” use cases that require a more thorough ethical assessment and “good-to-go” cases that require no additional scrutiny, thereby optimizing resource allocation and efficiency in managing ethical considerations and enabling bottom-up innovation. The screening mechanism and the resulting stratified self-assessment process are designed to be integrated into upcoming formalized AI governance structures, providing comprehensive oversight of past and current GAI project activities, with the current set-up serving as an interim solution, allowing for rapid risk classification. Due to its simple design focusing on pressing issues in GAI, the framework is applicable across sectors and addresses ethical quandaries with broad relevance to GAI but is also tailored to specific areas relevant to the company, such as interactions with healthcare professionals through outward-facing AI applications, which constitutes a “high risk area.” The framework also helps to build ethics capacity within teams, which are typically made up of project managers and technical staff. By encouraging the team to interrogate use cases from an ethical perspective, and by stimulating discussions about ethically relevant issues, the framework promotes awareness of ethical considerations and encourages ongoing dialogue about the ethical aspects of data analytics work. GAI can be applied in various scenarios, including personalized marketing e-mails, shortlisting candidates in HR processes, drug development, legal advice, and more. It is crucial to highlight to employees the potential risks associated with such scenarios.

This framework is also applicable in other contexts. Many of the principles contained in Merck’s CoDE transcend specific company boundaries and are prevalent in much of the academic literature and in various ethical frameworks of for-profit and not-for-profit entities (Becker et al., 2022; Fjeld et al., 2020; Hagendorff, 2020; Jobin et al., 2019). Specifics relevant to Merck, such as the definition of what constitutes a “high risk area” are easily interchangeable. The same applies to the definition of sensitive business areas, which can vary from company to company or sector to sector. As shown above, the successful alignment of company-specific principle frameworks – or, at a more general level, company values, charters or related guidelines – hinges not solely on theoretical perspectives from public discourse or scholarly literature, but indeed on sustained, direct engagement with practitioners throughout the company. Employing methods such as surveys, interviews, and portfolio analyses facilitates a nuanced understanding of the breadth of ongoing projects, the precise ethical questions they raise, and the boundaries and terminologies, such as the precise meaning of what constitutes “high risk” in a given company context, that must be clearly defined. By maintaining close, hands-on engagement with those directly involved, ethically oriented assessments can thus become both genuinely meaningful and effectively tailored to the organization’s specific context.

While the framework has been rolled out and is currently utilized by GAI practitioners, Merck follows the advice of the DEAP, which recommended that Merck contextually monitor the GAI use case development process over time. This includes analyzing the accuracy and appropriateness of the categorizations as assessed by the use case owners or developers, as well as reviewing low-risk cases to ensure they still align with the established categories. The system is designed to be dynamic to accommodate changes in the definitions of certain categories within the framework. By implementing such a continuous monitoring process, Merck can better ensure that its GAI use cases continue to be developed responsibly and ethically while effectively addressing new challenges or changes in data categories. Regular reviews of the decision tree and governance mechanisms will be conducted to identify areas for improvement and ensure that the system remains robust and effective. Such an approach is intended to ensure that the assessment method adheres to predefined quality criteria and is effective in detecting cases where oversight is warranted, thereby achieving its intended objectives.

The development process outlined herein reveals key insights into the breadth such a framework can reasonably cover. Designed primarily to meet the operational needs of Merck its scope can be extended to a range of GAI use cases within the company – and potentially beyond them. At the same time, it cannot serve as an all-purpose blueprint for other organizations. However, several considerations that shape the framework’s design constitute generally applicable guidelines for ethics practitioners. Chief among these are the selection of a suitable principle framework or the derivation of relevant principles from more abstract organizational values or charters and the tailoring of ethically salient markers to the distinctive operational features and ethical dimensions

of a particular company. Our observations indicate that achieving this alignment requires philosophical validity while also accommodating the practical expectations of AI practitioners in various departments. In our experience, a blend of interactive methodologies – such as surveys, interviews, and iterative pilot testing, supported by external ethics committees like the DEAP – represents a promising approach. We further suggest that the underlying rationales for identifying critical markers, alongside the process by which they were selected, hold general relevance for similar ethical frameworks.

Conclusion

The objective of this paper was to contribute to the following research question: “How can an assessment framework be developed to facilitate the ethical integration of GAI in corporate environments while enabling practitioners in bottom-up innovation?” In order to address this question, we have developed a practical framework for the resolution of ethical issues in GAI, with particular emphasis on those arising in technology companies.

The ethics framework outlined here serves as a vital tool for Merck in addressing the ethical risks associated with GAI technologies while embedding digital ethics throughout the organization. Developed with support from Merck’s Digital Ethics Advisory Panel (DEAP) and framed around broader GAI user needs and provisions established by senior AI leaders in the company, this framework is created through constant feedback seeking among actual GAI users. It builds on the company’s existing ethical models, such as the Code of Digital Ethics (CoDE) and semi-automated risk assessments, while specifically addressing the unique challenges of GAI, including inaccuracies, hallucinations, and the necessity for human oversight and verification. By providing flexible guidance for the responsible and ethical implementation of GAI in the private sector, the framework proves adaptable across various industries and contexts. Its tiered oversight approach encourages innovation and experimentation while ensuring appropriate governance. This bottom-up, self-service approach empowers use case owners to make ethically informed decisions, building ethics capacity across the workforce and enhancing awareness of key ethical considerations in data and GAI.

In developing such a framework, we conclude that a risk-based, *Ladder*-style approach, as described by Mökander et al. (2023), employing clear, actionable criteria, offers the optimal balance between complexity and usability for diverse teams supporting a large number of users. In contrast, a *Switch* model that subjects every GAI system to ethical review may stifle innovation, while a *Matrix* model with intricate, multi-dimensional analyses risks overburdening staff. Nevertheless, each approach proves valuable in particular contexts: for instance, we adopted the *Matrix* logic to develop a semi-automated risk assessment in a data analytics context (Charton et al., 2025).

The present form of ethical reflection and oversight is set up to evolve from self-assessment to more formal regulatory compliance as required, maintaining a balance between creativity and control. It is expected that frameworks such as our GAI ethics framework will continue to be important in the future, despite the implementation of substantial regulatory measures such as the EU AI Act. While the EU AI Act offers organizations a framework for the development of reliable and compliant AI, it has also been the subject of criticism. It is argued that the scope of the EU AI Act is insufficient, with calls for a wider list of prohibited AI systems and a more comprehensive Annex III (High-Risk-AI-Systems). Furthermore, concerns have been raised that exceptions within the EU AI Act facilitate the reduction of risk estimation from high-risk to low-risk, which in turn results in the minimization of the obligations placed upon organizations. Finally, while the EU AI Act provides comprehensive guidance on the regulation of “AI systems” and “general-purpose AI models,” it offers less detailed guidance on “general-purpose AI systems,” which encompass the majority of GAI applications, including ChatGPT, DALL-E, and Midjourney (Wachter, 2024). As these applications become increasingly integrated into business operations, such frameworks will be essential for navigating the complex ethical landscape, fostering responsible innovation, and building trust among stakeholders. It is imperative that frameworks such as the GAI Ethics Framework are viewed as a valuable addition to existing regulations, such as the EU AI Act.

Disclosure statement

RH and SB are employed by idigiT GmbH. SL and JC are employed by Merck KGaA, Darmstadt, Germany. idigiT has been commissioned by Merck to consult the company on digital ethics. In her role as advisor to Merck, SB has been a compensated private consultant. For this editorial, the authors acted as independent scholars and did not receive payment for their contributions.

Funding

The author(s) reported there is no funding associated with the work featured in this article.

Notes on contributors

Simon Lucas is Lead Expert Bioethics & Digital Ethics at Merck KGaA. In this role, he is working on translating ethical reasoning into actionable practices within the company, especially around novel biotechnologies and digital solutions. He joined Merck in 2018 after a decade in R&D and strategy roles at Grünenthal, Boehringer Ingelheim, and Evonik. Simon studied chemistry and bioethics in Munich, Saarbrücken, Copenhagen and Mainz. He holds a PhD in natural sciences and pursued his MA in neuroethics and experimental philosophy.

René M. Heinitz is a Solution Engineer at Deloitte, specializing in digital ethics, policy analysis and AI governance. René supports organizations across industries in the development and implementation of AI governance frameworks. He is co-author of several publications on the operationalization of digital ethics and has held several seminars on the subject at the University of Witten/Herdecke. He studied economic sociology and social sciences at the University of Trier, the Université Catholique de Louvain and the University of Cologne.

Sarah J. Becker is Lead Partner at Deloitte Consulting for digital ethics & AI governance. With more than 15 years of experience, she advises multinational clients on AI strategy, governance, and ethics. She helps organizations design and implement AI governance frameworks to ensure the responsible use of AI. Sarah is a lecturer at the University of Witten/Herdecke and has published several articles and books in the field of digital ethics. She studied human medicine and general management and was awarded with a research fellowship at Columbia University. She holds a PhD in social sciences from the University of Witten/Herdecke.

Jean Enno Charton is Director Digital Ethics & Bioethics at Merck KGaA. In this role, he is responsible for digital ethics and bioethics matters across all divisions, reporting directly into the board. Prior to this position, Jean Enno held various roles in the R&D department at Merck Healthcare, most recently as Chief of Staff to the Chief Medical Officer. He joined Merck in 2014 and holds a PhD in Immuno-Oncology from the University of Lausanne, Switzerland, as well as a biochemistry degree from the University of Tübingen, Germany.

ORCID

Simon Lucas  <http://orcid.org/0009-0006-7467-0113>

References

- AI Ethics Impact Group. (2020, October 8). From principles to practice: An interdisciplinary framework to operationalise AI ethics. <https://www.ai-ethics-impact.org/en>
- AI Verify Foundation. (2024, January 16). Proposed model AI governance framework for generative AI. https://aiverifyfoundation.sg/downloads/Proposed_MGF_Gen_AI_2024.pdf
- Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15(2). <https://doi.org/10.7759/cureus.35179>
- Al-Kfairy, M., Mustafa, D., Kshetri, N., Insiew, M., & Alfandi, O. (2024). Ethical challenges and solutions of generative AI: An interdisciplinary perspective. *Informatics*, 11(3), 58. <https://doi.org/10.3390/informatics11030058>
- Alles, G. (2024). *Towards a comprehensive understanding: An inquiry on generative AI, learning inequalities and civic responsibility* [Doctoral dissertation]. ST. Cloud State University. The Repository @ St. Cloud State. https://repository.stcloudstate.edu/eng_etds/37
- Annapureddy, R., Fornaroli, A., & Gatica-Perez, D. (2024). Generative AI literacy: Twelve defining competencies. *Digital Government: Research and Practice*, 6(1), 1–21. <https://doi.org/10.1145/3685680>
- Bargh, M. S., & Troxler, P. (2020). Digital transformations and their design-renewal of the socio-technical approach. In H. Rotterdam (Ed.), *Hoger beroepsonderwijs in 2030 toekomstverkenningen en scenario's vanuit Hogeschool Rotterdam* (1th ed., pp. 327–370). Hogeschool Rotterdam Uitgeverij. <https://www.hogeschoolrotterdam.nl/contentassets/3668dcb90f8e4c7ab114f4cf899cd45b/bundel-hoger-beroepsonderwijs-in-2030.pdf>

- Becker, S. J., Nemat, A. T., Lucas, S., Heinitz, R. M., Klevesath, M., & Charton, J. (2022). A code of digital ethics: Laying the foundation for digital ethics in a science and technology company. *AI & Society*, 38(6), 2629–2639. <https://doi.org/10.1007/s00146-021-01376-w>
- Bian, Y., & Xie, X. Q. (2021). Generative chemistry: Drug discovery with deep learning generative models. *Journal of Molecular Modeling*, 27(3), 1–18. <https://doi.org/10.1007/s00894-021-04674-8>
- Blackman, R. (2020, October 15). A practical guide to building ethical AI. *Harvard Business Review*. <https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai>
- Britton-Heitz, A., & Merhout, J. (2024). Human roles in implementation and oversight of artificial intelligence: Towards a governance framework. *MWAIS, 2024 Proceedings*. 14. <https://aisel.aisnet.org/mwais2024/14>
- Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J., Branas-Garza, P., Butera, L., Douglas, K. M., Everett, J., Gigerenzer, G., Greenhow, C., Hashimoto, D. A., Holt-Lunstad, J., Jetten, J., Johnson, S., Kunz, W. H., Longoni, C., ... Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus*, 3(6). <https://doi.org/10.1093/pnasnexus/pgae191>
- Charton, J., Lucas, S., Vogd, W., Halter, W., & Becker, S. J. (2025). *Operationalising Digital Ethics: Establishment of an Ethics Evaluation Tool for Data Analytics* [Manuscript submitted for publication]. Merck KGaA.
- Chen, Y., & Esmailzadeh, P. (2024). Generative AI in medical practice: In-depth exploration of privacy and security challenges. *Journal of Medical Internet Research*, 26, e53008. <https://preprints.jmir.org/preprint/53008>
- Corrêa, N. K., Santos, J. W., Galvão, C., Pasetti, M., Schiavon, D., Naqvi, F., Hossain, R., & Oliveira, N. D. (2024). Crossing the principle-practice gap in AI ethics with ethical problem-solving. *AI Ethics*, 5(2), 1–18. <https://doi.org/10.1007/s43681-024-00469-8>
- Danish Design Center. (2021, June 12). The digital ethics compass. <https://ddc.dk/tools/toolkit-the-digital-ethics-compass>
- Denny, P., Leinonen, J., Prather, J., Luxton-Reilly, A., Amarouche, T., Becker, B. A., & Reeves, B. N. (2024, March). Prompt problems: A new programming exercise for the generative AI era. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education*, Portland, Oregon, USA. (Vol., 1. pp. 296–302). <https://doi.org/10.1145/3626252.3630909>
- Deon. (2018, April 28). An ethics checklist for data scientists. <https://deon.drivendata.org/>
- Epstein, Z., Payne, B. H., Shen, J. H., Hong, C. J., Felbo, B., Dubey, A., Groh, M., Obradovich, N., Cebrian, M., & Rahwan, I. (2018). TuringBox: An experimental platform for the evaluation of AI systems. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Stockholm, Sweden. <https://doi.org/10.24963/ijcai.2018/851>
- Feldstein, S. (2023). The consequences of generative AI for democracy, governance and war The International Institute for Strategic Studies (IISS). In *The International Institute for Strategic Studies (IISS)*, survival: October–November 2023 (1th ed., pp. 117–142). Routledge. <https://doi.org/10.4324/9781003429388>
- Ferrara, E. (2024). GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, 7(1), 549–569. <https://doi.org/10.1007/s42001-024-00250-1>
- Fischer, M., Foord, D., Frecè, J., Hillebrand, K., Kissling-Näf, I., Meili, R., Peskova, M., Risi, D., Schmidpeter, R., & Stucki, T. (2023). Sustainability in a digital context. In M. Fischer, D. Foord, J. Frecè, K. Hillebrand, I. Kissling-Näf, R. Meili, M. Peskova, D. Risi, R. Schmidpeter, & T. Stucki (Eds.), *Sustainable business managing the challenges of the 21st century* (1th ed., pp. 117–128). Springer. https://doi.org/10.1007/978-3-031-25397-3_8
- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*, No. 2020-2021. <https://doi.org/10.2139/ssrn.3518482>
- Floridi, L. (2012). Distributed morality in an information society. *Science and Engineering Ethics*, 19(3), 727–743. <https://doi.org/10.1007/s11948-012-9413-4>
- Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, 32(2), 185–193. <https://doi.org/10.1007/s13347-019-00354-x>
- Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), 1–5. <https://doi.org/10.1098/rsta.2016.0360>
- Georgieva, I., Lazo, C., Timan, T., & Van Veenstra, A. F. (2022). From AI ethics principles to data science practice: A reflection and a gap analysis based on recent frameworks and practical experience. *AI Ethics*, 2(4), 697–711. <https://doi.org/10.1007/s43681-021-00127-3>
- 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>
- Hall, M., & Haas, C. (2021). Brown hands aren't terrorists: Challenges in image classification of violent extremist content. In V. Duffy (Ed.), *Digital human modeling and applications in health, safety, ergonomics and risk management. AI, product and service. HCII 2021. Lecture notes in computer science* (p. 12778). Springer. https://doi.org/10.1007/978-3-030-77820-0_15
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. <https://dl.acm.org/doi/10.5555/2017212.2017217>

- Hickok, M. (2020). Lessons learned from AI ethics principles for future actions. *AI and Ethics*, 1(1), 41–47. <https://doi.org/10.1007/s43681-020-00008-1>
- Jadon, A. (2024). Ethical AI development: Mitigating bias in generative models. <https://easychair.org/publications/preprint/1Klv>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Kenthapadi, K., Lakkaraju, H., & Rajani, N. (2023, August). Generative AI meets responsible AI: Practical challenges and opportunities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Long Beach, California, USA. (pp. 5805–5806). <https://doi.org/10.1145/3580305.3599557>
- Kox, E. S., & Beretta, B. (2024). Evaluating generative AI incidents: An exploratory vignette study on the role of trust, attitude and AI literacy. *HAI 2024: Hybrid Human AI Systems for the Social Good*, 188–198. <https://doi.org/10.3233/FAIA240194>
- Li, Z., Yi, W., & Chen, J. (2024). Accuracy of training data and model outputs in generative AI: CREATe response to the Information Commissioner Office consultation. *arXiv Preprint*, arXiv: 2407, 13072.
- Mittelstadt, B. D. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501–507. <https://doi.org/10.1038/s42256-019-0114-4>
- Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2022). Conformity assessments and post-market monitoring: A guide to the role of auditing in the proposed European AI regulation. *Minds and Machine*, 32(2), 241–268. <https://doi.org/10.1007/s11023-021-09577-4>
- Mökander, J., Curl, J., & Kshirsagar, M. (2024). A blueprint for auditing generative AI. *arXiv Preprint*, arXiv: 2407, 05338.
- Mökander, J., Sheth, M., Watson, D. S., & Floridi, L. (2023). The switch, the ladder, and the matrix: Models for classifying AI systems. *Minds and Machines*, 33(1), 221–248. <https://doi.org/10.1007/s11023-022-09620-y>
- Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 26(4), 2141–2168. <https://doi.org/10.1007/s11948-019-00165-5>
- Nemat, A. T., Becker, S. J., Lucas, S., Thomas, S., Gadea, I., & Charton, J. (2023). The principle-at-risk analysis (PaRA): Operationalising digital ethics by bridging principles and operations of a digital ethics advisory panel. *Minds and Machines*, 33(4), 737–760. <https://doi.org/10.1007/s11023-023-09654-w>
- Novelli, C., Casolari, F., Hacker, P., Spedicato, G., & Floridi, L. (2024). Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. *arXiv Preprint*, arXiv: 2401, 07348.
- Obrenovic, B., Gu, X., Wang, G., Godinic, D., & Jakhongirov, I. (2024). Generative AI and human-robot interaction: Implications and future agenda for business, society and ethics. *AI & Society*, 40(2), 1–14. <https://doi.org/10.1007/s00146-024-01889-0>
- Obwegeser, N., Yokoi, T., Wade, M., & Voskes, T. (2020). 7 key principles to govern digital initiatives. *Mit Sloan management review*. 7 Key Principles to Govern Digital Initiatives.
- OECD. (2022). Oecd framework for the classification of AI systems. https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.html
- Omidyar Network. (2018, June 12). New toolkit shows companies how to anticipate, prevent bad actors from using tech in harmful ways. <https://medium.com/omidyar-network/introducing-the-worlds-first-ethical-operating-system-7acc4abc2bfa>
- Popowicz-Pazdej, A. (2023). Why the generative AI models do not like the right to be forgotten: A study of proportionality of identified limitations. *Przegląd Prawniczy Uniwersytetu Im Adama Mickiewicza*, 15, 217–238. <https://www.ceeol.com/search/article-detail?id=1230518>
- Preiksaitis, C., & Rose, C. (2023). Opportunities, challenges, and future directions of generative artificial intelligence in medical education: Scoping review. *JMIR medical education*, 9, <https://preprints.jmir.org/preprint/48785>
- Romero Moreno, F. (2024). Generative AI and deepfakes: A human rights approach to tackling harmful content. *International Review of Law, Computers and Technology*, 38(3), 297–326. <https://doi.org/10.1080/13600869.2024.2324540>
- Ruiz-Rojas, L. I., Acosta-Vargas, P., De Moreta-Llovet, J., & Gonzalez-Rodriguez, M. (2023). Empowering education with generative artificial intelligence tools: Approach with an instructional design matrix. *Sustainability*, 15(15), 11524. <https://doi.org/10.3390/su151511524>
- Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv Preprint*. <https://arxiv.org/abs/1811.05577>
- SAP. (2024). Sap AI ethics handbook. <https://www.sap.com/documents/2023/03/7211ee96-647e-0010-bca6-c68f7e60039b.html>
- Schuett, J., Reuel, A., & Carlier, A. (2023). How to design an AI ethics board. *AI Ethics*, 5(2), 863–881. <https://doi.org/10.1007/s43681-023-00409-y>
- Su, J., & Yang, W. (2023). Unlocking the power of ChatGPT: A framework for applying generative AI in education. *ECNU Review of Education*, 6(3), 355–366. <https://doi.org/10.1177/20965311231168423>

- Sugarman, J., Shivakumar, S., Rook, M., Loring, J. F., Rehmann-Sutter, C., Taupitz, J., Reinhard-Rupp, J., & Hildemann, S. (2018). Ethical considerations in the manufacture, sale, and distribution of genome editing technologies. *American Journal of Bioethics*, 18(8), 3–6. <https://doi.org/10.1080/15265161.2018.1489653>
- Sun, J., Liao, Q. V., Muller, M., Agarwal, M., Houde, S., Talamadupula, K., & Weisz, J. D. (2022, March). Investigating explainability of generative AI for code through scenario-based design. *Proceedings of the 27th International Conference on Intelligent User Interfaces*, Helsinki, Finland. (pp. 212–228). <https://doi.org/10.1145/3490099.3511119>
- Thorne, S. (2024). Understanding the interplay between trust, reliability, and human factors in the age of generative AI. *International Journal of Simulation Systems, Science & Technology*, 25(1). <https://doi.org/10.5013/IJSSST.a.25.01.10>
- UN Global Pulse. (2019, June 12). Risks, harms and benefits assessment. <https://www.unglobalpulse.org/document/risks-harms-and-benefits-assessment-tool/>
- Veluru, C. S. (2024). Impact of artificial intelligence and generative AI on healthcare: Security, privacy concerns and mitigations. *Journal of Artificial Intelligence & Cloud Computing*, 3(347), 2–6. <https://www.onlinescientificresearch.com/articles/impact-of-artificial-intelligence-and-generative-ai-on-healthcare-security-privacy-concerns-and-mitigations.pdf>
- Wach, K., Duong, C. D., Ejdays, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiwicz, 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/10.15678/EBER.2023.110201>
- Wachter, S. (2024). Limitations and loopholes in the EU AI Act and AI liability directives: What this means for the European Union, the United States, and beyond. *Yale Journal of Law and Technology*, 26(3), 671–718. <https://ora.ox.ac.uk/objects/uuid:0525099f-88c6-4690-abfa-741a8c057e00>
- Xiang, A. (2024). Fairness & privacy in an age of generative AI. *Science and Technology Law Review*, 25(2). <https://doi.org/10.52214/stlr.v25i2.12765>
- Xu, D., Fan, S., & Kankanhalli, M. (2023, October). Combating misinformation in the era of generative AI models. *Proceedings of the 31st ACM International Conference on Multimedia*, Toronto, Canada. (pp. 9291–9298). <https://doi.org/10.1145/3581783.3612704>
- Zeng, X., Wang, F., Luo, Y., Kang, S. G., Tang, J., Lightstone, F. C., Fang, E. F., Cornell, W., Nussinov, R., & Cheng, F. (2022). Deep generative molecular design reshapes drug discovery. *Cell Reports Medicine*, 3(12), 100794. <https://doi.org/10.1016/j.xcrm.2022.100794>
- Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023, April). Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany. (pp. 1–20). <https://doi.org/10.1145/3544548.3581318>
- Zhou, M., Abhishek, V., Derdenger, T., Kim, J., & Srinivasan, K. (2024). Bias in generative AI. *ArXiv Preprint, arXiv: 2403.02726*.
- Zlateva, P., Steshina, L., Petukhov, I., & Velez, D. (2024). A conceptual framework for solving ethical issues in generative artificial intelligence. *Electronics, Communications and Networks*, 381, 110–119. <https://doi.org/10.3233/FAIA231182>