Ethics of Foundation Models in Computational Pathology: Overview of Contemporary Issues and Future Implications

Rui Fei Du, Eduard Lloret Carbonell, Jiaxuan Huang, Sheng Liu, Xiaohang Wang, Dinggang Shen, and Jing Ke

***Abstract*— Artificial intelligence (AI) has profoundly transformed our lives, reshaping industries and impacting nearly every aspect of society over the past few decades. It has recently become even more influential, primarily due to the rise of foundation models representing a new paradigm in AI development. These models, characterized by their large-scale training on vast datasets, have unique capabilities such as emergence and transference, enabling them to generalize across diverse tasks. Since their introduction, foundation models have been increasingly applied in fields such as autonomous driving, computer vision, marketing, finance, industrial robotics, and healthcare. Pathologists worldwide use computational methods to analyze diseases that profoundly impact human well-being, including cancer diagnosis and staging, genetic mutation prediction, and treatment and prognosis forecasting. In this article, we discuss how, despite the promise of foundation models in various applications, their development and application in computational pathology remain challenging due to inherent characteristics such as emergence, homogenization, hallucination, transference, compositionality, and explainability. While powerful, these traits introduce numerous ethical concerns and challenges, impacting safety and reliability, patient privacy, accountability, and equity and fairness in healthcare access. We examine these ethical issues, focusing on key concerns like algorithmic discrimination and misuse, accuracy, privacy breaches, transparency, public accessibility, and accountability. Furthermore, potential solutions to these challenges are analyzed, offering future perspectives on promoting the development and application of more ethical AI and foundation models in computational pathology. These insights aim to guide foundation models toward responsible integration of AI in healthcare.**

***Index Terms*— Foundation model, computational pathology, artificial intelligence ethics, AI trustworthiness.**

Received 3 January 2025; revised 18 February 2025; accepted 13 March 2025. This work was supported in part by the Natural Science Foundation of Shanghai under Grant 23ZR1430700; in part by the National Natural Science Foundation of China under Grant 82441023, Grant U23A20295, Grant 62131015; in part by the Medical Research Project of Health Commission of Shanghai Hongkou District under Grant HW2202-33; in part by Shanghai Municipal Central Guided Local Science and Technology Development Fund under Grant YDZX20233100001001; in part by High-performance Computing (HPC) Platform of ShanghaiTech University; and in part by the Start-up Scientific Research Project of Shanghai Fourth People's Hospital Affiliated to Tongji University under Grant SYKYQD03901. *(Corresponding authors: Jing Ke; Dinggang Shen; Xiaohang Wang; Sheng Liu.)*

Rui Fei Du and Eduard Lloret Carbonell are with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: rdu123@sjtu.edu.cn; edu.lloret@sjtu.edu.cn).

Jiaxuan Huang is with the Dulwich College Shanghai, Shanghai 201111, China (e-mail: christine.huang28@stu.dulwich.org).

Sheng Liu is with the Department of Thyroid and Breast Surgery, Shanghai Fourth People's Hospital, School of Medicine, Tongji University, Shanghai 200434, China (e-mail: tocter@msn.com).

Xiaohang Wang is with the Department of Radiation Oncology, Ren Ji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China (e-mail: wxh1991@126.com).

Dinggang Shen is with the School of Biomedical Engineering and the State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai 201203, China, also with Shanghai United Imaging Intelligence Company Ltd., Shanghai 201807, China, and also with Shanghai Clinical Research and Trial Center, Shanghai 200231, China (e-mail: dgshen@shanghaitech.edu.cn).

Jing Ke is with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: kejing@sjtu.edu.cn).

Digital Object Identifier 10.1109/TMI.2025.3551913

I. Introduction

PATHOLOGYis a data-driven discipline that leverages both clinical and phenotypic data to enhance the diagnosis of diseases. Traditional pathology is conducted by directly examining tissue-bearing slides under a microscope (also known as *histopathology* slides), allowing pathologists to evaluate nuclear and cytoplasmic compositions of tissues in fine detail. As relevant technology matured in the past two decades, a novel process for digitizing histopathology slides using whole-slide scanners emerged [1], laying the groundwork for digital pathology to thrive. Digital pathology is a sub-field of pathology that involves the examination of digitized high-resolution whole-slide images (WSI), potentially aiding pathologists in the identification of elusive or very focal abnormalities [2], [3], [4], [5], [6], [7], [8]. With the rapid advancements in artificial intelligence (AI) over the past decade, computational pathology, a sub-field of digital pathology, has reemerged and gained momentum [9], [10]. Computational pathology involves applying advanced AI techniques, such as machine learning, deep learning, and data analytics, to analyze and interpret pathology data, including digital images, genomic information, clinical records, and more [11], [12], [13], [14]. Computational pathology aims to enhance traditional pathology by incorporating computational tools to support pathologists in diagnosis, prognosis, and

treatment planning, ultimately leading to improved patient outcomes [15], [16], [17], [18], [19], [20], [21].

There is an increasing need to expand pathology diagnoses by integrating additional patient data, such as lifestyle and socioeconomic factors, into the routine diagnostic workflow to enhance research categories like cohort description, applied methods, and patient outcomes within the realms of pathology and precision medicine [22]. Such complex, cross-domain data handling and integration are well-suited for AI frameworks, particularly foundation models [23].

Before exploring foundation models, it is essential to examine the evolution of traditional AI models and understand the advancements that render them superior to their predecessors. When the first AI systems were developed, they relied heavily on explicitly programmed human logic and rules, trained on small datasets of labeled data. As a result, traditional AI models are typically designed to perform a single, specific task with little to no flexibility beyond that task. This concept, known as task-specific AI, significantly limits its applicability across various use cases [24], [25]. For example, traditional AI models designed to distinguish between an apple and a banana may exhibit outstanding accuracy but would fail at distinguishing other fruits.

Foundation models are developed to solve the inherent shortcomings of traditional AI systems, such as poor adaptation to new situations, limited scope of application, and over-reliance on human guidance. Foundation models, also known as large AI models, are typically built using an established neural network architecture called a *transformer* [26]. Transformers have been pivotal in foundation models, enabling them to understand unlabeled data. In conjunction with the ability to be trained on massive datasets, made possible by the advancement in computational hardware and parallelism (i.e., large clusters of GPUs), foundation models learn the underlying patterns of a given dataset and thus generalize to new tasks and objectives, a characteristic known as emergence.

Fig. 1 illustrates a generic comparison of structural and application differences among traditional AI models, general medical foundation models, and computational pathology foundation models, with the computational pathology example representing a typical vision-language model (VLM). This figure illustrates one of the key differences between traditional and foundation models: Traditional AI models are typically designed to perform one specific task at a time. In contrast, foundation models are capable of handling multiple tasks simultaneously, adapting to new tasks without requiring task-specific training. This figure also illustrates a multimodal CPATH foundation model, while other types, such as unimodal foundation models, vision-only models, and language-focused models (LLMs), also exist within the domain.

Foundation models are applied across diverse domains due to their flexibility and adaptability to various downstream tasks [27]. In healthcare, they enhance diagnostics, facilitate medical research, and support personalized medicine by integrating multimodal data such as patient records, medical imaging, and genetic information [28], [29], [30], [31], [32], [33]. In computer vision, they enable advancements in image recognition, object detection, and scene understanding, with applications ranging from autonomous driving to medical imaging and multimedia creation [34], [35]. Furthermore, these models play a vital role in marketing, finance, and robotics, providing scalable solutions to complex challenges across industries [23].

However, a direct consequence of utilizing AI technology is the introduction of many complex ethical risks and challenges, particularly in medicine and healthcare. For example, repeatedly training medical AI models on relatively homogeneous data or biased patient samples, such as those lacking diversity in gender, demographics, or age, potentially results in overgeneralized outcomes and biased AI-driven decisions and diagnoses [36]. Consequently, a model designed to predict diabetic retinopathy (DR) using clinical trial data from a small, homogeneous urban population in the U.S. may lead to misdiagnoses when applied to a cohort of patients from another country [37].

The main contributions of this article are summarized as follows.

1) We introduce the diverse and complex ethical issues emerging from foundation model applications due to their inherent characteristics, including emergence, homogenization, hallucination, transference, compositionality, and explainability [23], [24], [25]. These issues are discussed along with their risks and potential consequences, established guidelines for mitigation, and solutions for addressing them. These challenges and their potential consequences are thoroughly analyzed and discussed.
2) We offer a comprehensive overview of several state-of-the-art implementations of foundation models in pathology, detailing the diverse scenarios, goals, achievements, and methods of employing AI in tackling contemporary problems and challenges in computational pathology. We then examine the relationships between each foundation model application and the associated ethical challenges while exploring potential solutions to mitigate these issues.
3) Lastly, we explore the implications for the future of AI ethics and provide insights on ensuring that AI and foundation models develop ethically. The importance of adhering to ethical principles and guidelines is emphasized, particularly in applications related to computational pathology.

Fig. 2 depicts the article's structure. To the best of our knowledge, this article represents the first comprehensive examination offering contemporary insights and analyses, along with proposed solutions to the ethical challenges of applying foundation models in computational pathology.

II. Related Works

This article builds upon several key studies in the field of AI ethics and foundation models in computational pathology. McKay et al. [19] discusses the ethical challenges of AI-driven digital pathology, focusing on data privacy, bias, and algorithmic fairness. However, our work expands on these concerns by addressing additional issues specific to foundation models, such as emergent behaviors and compositionality.

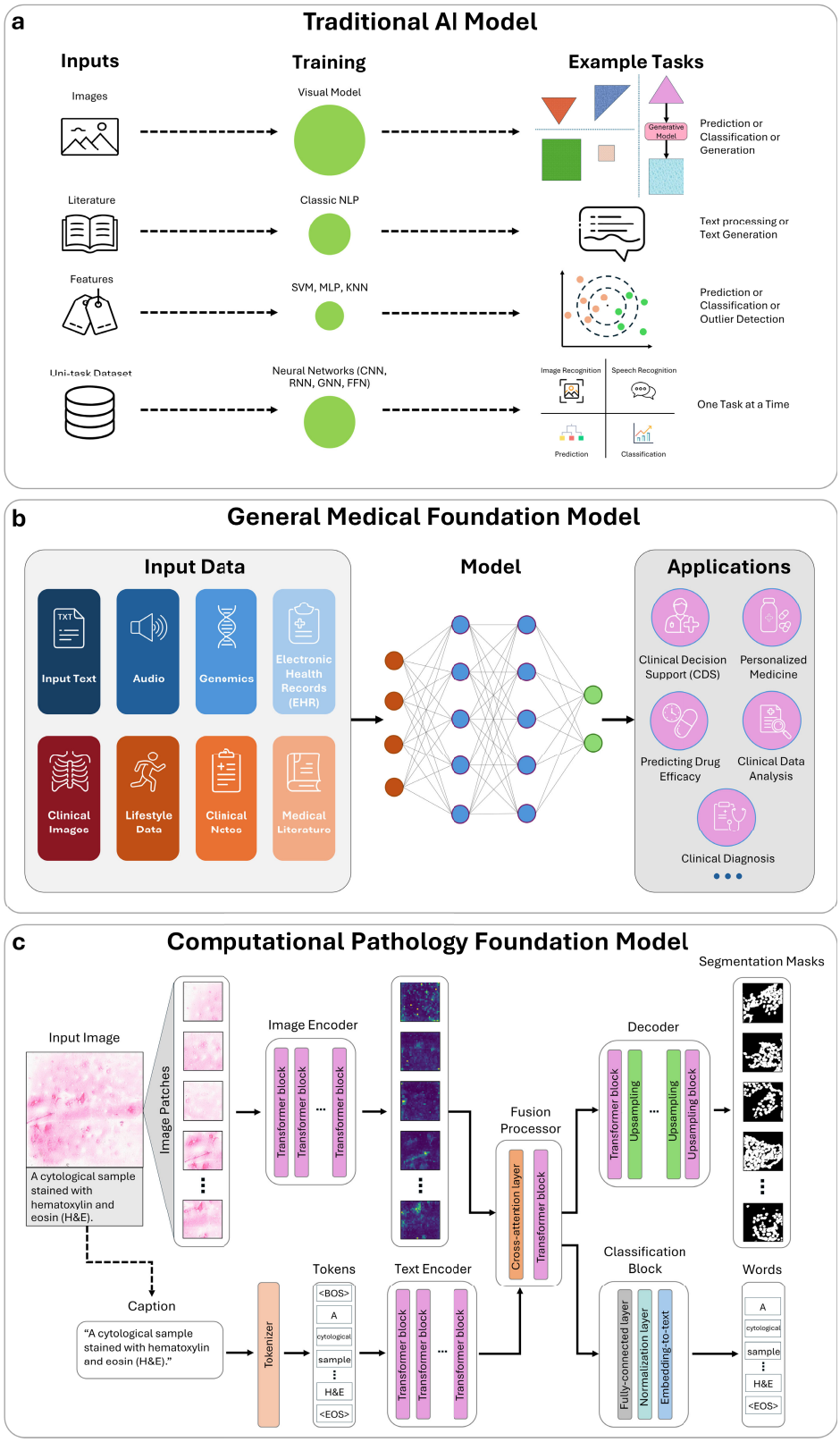

Fig. 1. A comparison in functional applications between different AI model architectures. **a**, The architecture of a traditional AI model consists of various types of inputs, the type of model itself, and the specific task it performs. **b**, The architecture of a general medical foundation model. **c**, The architecture of a typical computational pathology foundation model.

Waqas et al. [24] explores the promising potential of generative AI in digital pathology and its application to cancer diagnosis. While it focuses on the transformative capabilities and inherent characteristics of foundation models, our paper complements this by offering a more detailed analysis of how these characteristics relate to the ethical issues arising from their application in computational pathology. Additionally, we examine how these ethical issues impact medical

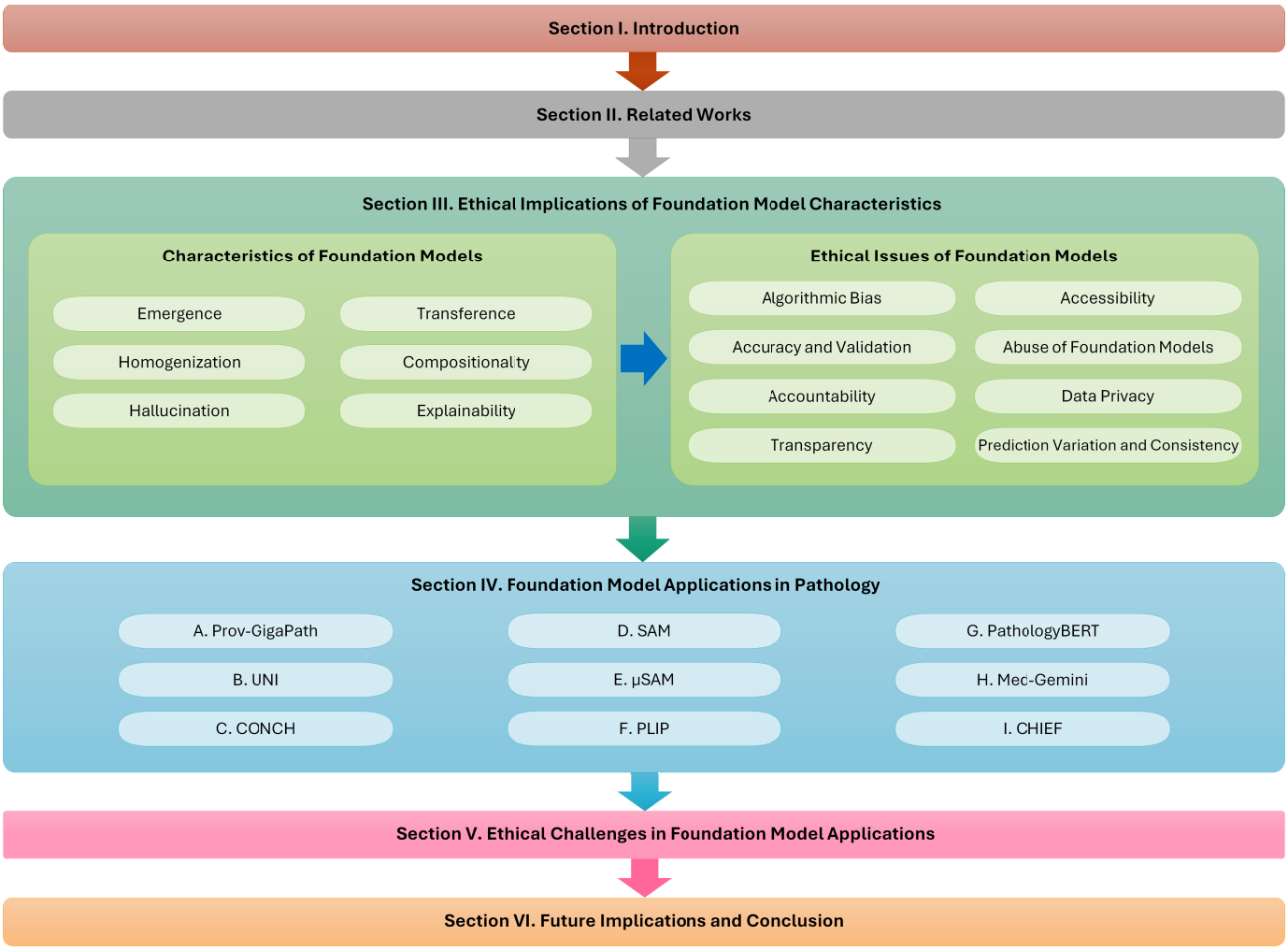

Fig. 2. Organization of this paper, which is divided into six sections: I Introduction, II Related Works, III Ethical Implications of Foundation Model Characteristics, IV Foundation Model Applications in Pathology, V Ethical Challenges in Foundation Model Applications, VI Future Implications and Conclusion.

decision-making and patient care, providing a comprehensive understanding of the challenges in this domain. Mcdermid et al. [38] focuses on AI explainability and transparency, topics we also explore. However, we extend this discussion by examining the ethical implications of foundation models' generalization capabilities and their application across multiple domains. This broader view allows us to discuss how these models may introduce unpredictable risks in critical healthcare applications. Sorell et al. [39] examines the challenges of AI opacity in computational pathology while we extend this discussion to include issues such as model homogenization and transference. These issues, especially concerning equity and fairness in patient outcomes, are unique to foundation models.

Finally, a comprehensive review of AI ethics by Huang et al. [20] provides valuable insights into the broader ethical considerations of artificial intelligence, but it does not explicitly address the challenges posed by foundation models in pathology. Our paper fills this gap by offering a detailed analysis of the unique ethical issues in this field, such as algorithmic bias, transparency, and the potential for misuse. Additionally, we propose solutions such as differential privacy and federated learning to mitigate these risks, marking a significant contribution to the ongoing discourse on ethical AI in healthcare. This focus on computational pathology and the practical application of foundation models is the core novelty of our work.

III. Ethical Implications of Foundation Model Characteristics

This section begins by outlining the major ethical issues associated with foundation model applications, including algorithmic bias, accuracy and validation, accountability, transparency, accessibility, abuse of AI models, data privacy, and prediction variation and consistency.

1) *Algorithmic bias*: Algorithmic bias refers to the ability of an AI algorithm to discriminate against individuals, groups, or populations, which directly affects the decision-making of the model and thus would have the potential to cause biased or even erroneous results. For example, Optum, a subsidiary of UnitedHealth Group, developed an application to identify high-risk patients with untreated chronic conditions. However, its algorithm has been found to discriminate against Black patients by basing risk on past treatment costs [40]. This risks exacerbating disparities in clinical outcomes, especially in breast cancer, which is 46% more likely

to be fatal for Black women. This example highlights the real-world consequences that biased AI imposes on minority communities [41].

2) *Accuracy and Validation*: When evaluating a foundation model that claims to perform specific tasks and generate results, the criteria for assessment remain an open-ended question. This represents one of the most crucial ethical issues regarding foundation models, particularly in fields like pathology that substantially influence people's well-being. To better illustrate the importance of model trustworthiness, Fig. 3 compares two different user queries fed into two AI models: GPT-4 and a general pathological model. We observed that GPT-4 initially provided an incorrect answer when identifying the animal in the input image. Similarly, we provided a cytology image to the medical pathology model MedGPT [42] and prompted it to identify the image, as shown in Fig. 3(b). To evaluate the model's accuracy, we introduced distracting elements by questioning the validity of its response. Despite the model's ability to maintain consistent predictions even with added noise, it is essential to acknowledge that no model is entirely robust. All foundation models are susceptible to hallucination, and even the most advanced models for a given task have shown instances of such behavior. This highlights the critical need for ongoing vigilance in ensuring robustness and safety when applying foundation models in the medical field.
3) *Accountability*: If an AI model fails in its tasks, causing specific consequences or damages, it becomes crucial to determine accountability and assign responsibilities to ensure fair judgment. Addressing accountability in AI presents a complex and nuanced challenge depending on the circumstances.
4) *Transparency*: In the field of AI ethics and governance, transparency encompasses two key dimensions: algorithmic transparency, which focuses on technical explainability, and information transparency, which addresses the disclosure of relevant information to stakeholders. On the algorithmic side, explaining and understanding the inference processes within machine learning (ML) algorithms—particularly those at the core of current foundation models—remains intrinsically challenging [38]. This obscurity often perplexes both users and developers, raising significant transparency concerns and potentially impeding effective human oversight.

   Equally important is information transparency. In a healthcare context, it is ethically imperative to disclose the use of foundation models to patients as part of their therapy. Patients should be informed about AI's role in diagnosing, prognosticating, and determining treatment processes. Furthermore, collecting or using a patient's data without explicit consent constitutes a breach of data privacy, underscoring the necessity of robust information-sharing practices. Ensuring transparency in algorithmic mechanisms and patient information disclosures allows developers and practitioners to uphold ethical standards while fostering trust.
5) *Accessibility*: It is becoming more evident that the accessibility and availability of emerging AI technologies will directly impact human well-being, and it is no different in a field as crucial as pathology. However, it would be unethical and unfair if only a portion of the population benefits from these technologies. Therefore, there is a legitimate concern for establishing a fair system to distribute AI-related products to the public evenly.
6) *Abuse of foundation models*: AI technology is a double-edged sword, and it is almost unavoidable for any technology to be abused by humans, whether intentionally or unintentionally. The potential consequences of misusing foundation models in pathology are particularly alarming, given their profound connection to our health and well-being.
7) *Data privacy*: With the advancement of big data and AI, the tension between developing AI technology and user privacy protection has intensified [39], [43]. Since the success of most foundation models relies on large-scale real-world input data, there is an increased risk of patient data leakage. Data security incidents have become increasingly common in recent years, making privacy protection a widely recognized and serious ethical issue in AI technology.
8) *Prediction Variation and Consistency*: Human performance is inherently variable, leading to unpredictability in their interactions with AI [44]. For example, we collected lung cancer reports from three pathologists in the same hospital for ground truth, only to find that the diagnostic decisions differed in at least one-third of 120 specimens. In contrast, predictions across different foundation models are often consistent, effectively avoiding biases in pathological scoring, such as visual traps [45]. However, it is important to recognize that progress sometimes comes from diverse interpretations of slightly suspicious tissue, where an emphasis on pure consistency could hinder the development of foundation models.

Next, we provide a detailed analysis of each key characteristic of foundation models, including emergence, homogenization, hallucination, transference, compositionality, and explainability [23], [24], [25] and examine how these characteristics contribute to the ethical issues discussed earlier. Fig. 4 illustrates the relationships between these characteristics and the ethical concerns outlined in this section.

A. Emergence and Ethical Issues

Emergence occurs when foundation models exhibit unexpected capabilities as they scale, leading to new and unforeseen behaviors [23], [24], [47]. While potentially beneficial, these emergent abilities raise significant ethical concerns related to algorithmic bias, as new forms of discrimination may arise without prior intention [48]. For example, as foundation models grow in complexity, their ability to generate biased outcomes may increase, directly impacting accuracy

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3551913

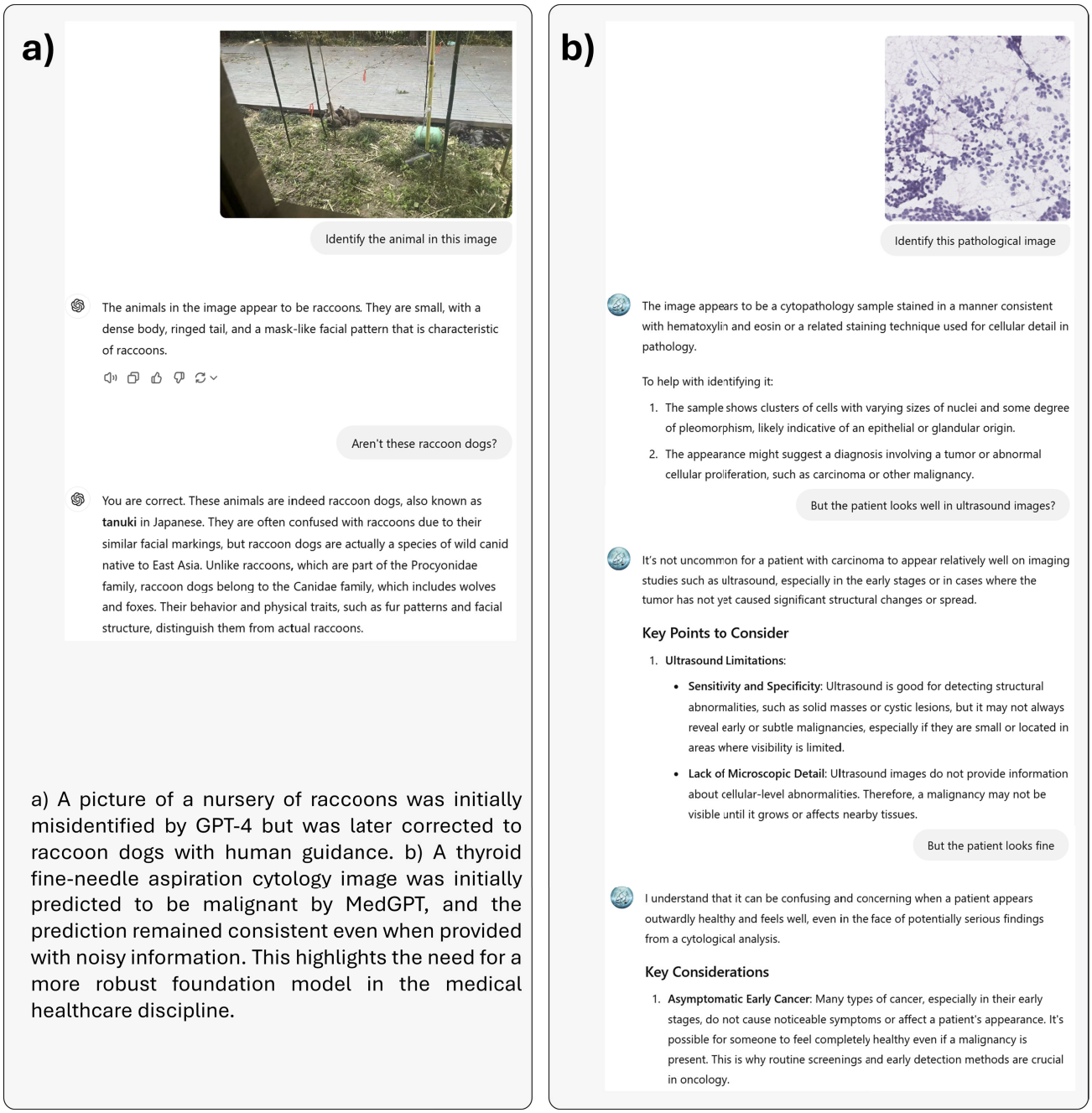

Fig. 3. Comparison of trustworthiness between (a) a general AI model - ChatGPT [46], and (b) a customized medical foundation model - MedGPT [42]. (The test log was captured in November 2024, with some key points or considerations cropped for simplicity in (b)).

and validation. The unpredictability associated with emergent behaviors also raises concerns about accountability when these models cause harm or fail to perform as expected, particularly in critical healthcare applications like pathology. This unpredictability underscores the need for continuous monitoring and rigorous validation of these models to ensure they operate ethically and safely.

B. Homogenization and Ethical Issues

Homogenization refers to the unification of methodologies used to develop foundation models across various machine learning applications, offering significant advantages such as improved efficiency, scalability, and the ability to leverage shared knowledge across domains [23], [24]. However, this approach also amplifies the risk of systemic bias. Models like BERT, GPT-4, and RoBERTa enable broad advancements, yet the biases embedded in these foundational systems tend to propagate across all adapted models, creating widespread and uniform issues [49]. This challenge is further compounded by the emergent qualities of foundation models, which often result in unpredictable behavior and obscure sources of bias.

These biases, inherited without scrutiny, frequently give rise to a loss of diversity in both content and predictions, a phenomenon referred to as outcome homogenization [50], which impacts critical domains such as healthcare, law, and education, where fairness and equity are paramount. For

example, the fact that melanoma occurs less frequently in individuals with black skin introduces a form of algorithmic bias. If diagnostic models are trained predominantly on datasets from populations with lighter skin tones, they may be less accurate in diagnosing melanoma in individuals with darker skin [51]. This bias stems from homogenization, where the model's training on homogeneous data predominantly from one population leads to inaccurate or biased predictions when applied to underrepresented groups. Emergence also plays a role here, as the model may display new and unexpected patterns, such as misdiagnosing melanoma in darker-skinned individuals, due to its initial inability to account for this disparity in training data. Moreover, homogenization challenges the transparency of algorithms, as the lack of diversity in model architectures and training data may obscure the reasons behind specific outputs, complicating the process of ensuring fairness and accountability.

C. Hallucination and Ethical Issues

Hallucination in foundation models, particularly in generative models such as GPT-4, occurs when they produce outputs that appear plausible but are factually incorrect or fabricated, often misaligning with real-world knowledge or data [24], [52]. This issue raises significant ethical challenges, including concerns about accuracy, prediction variation and consistency, the abuse of foundation models, and accountability. Hallucination is driven by factors such as limited datasets, overgeneralization, and the lack of real-time data [53]. In high-stakes applications like computational pathology and medical diagnosis, these seemingly confident yet inaccurate responses are particularly harmful [54]. This characteristic challenges accountability, as errors arising from hallucinations may be mistaken for valid results, undermining the credibility and reliability of AI systems. Hallucinations also compromise the transparency of algorithms, as it becomes more difficult for users to understand why certain predictions were made, particularly in high-stakes medical contexts.

The abuse of foundation models becomes an even greater risk when users trust these models' outputs without critically evaluating them, intentionally or unintentionally. In medical settings, where life-and-death decisions are often made, overreliance on hallucinated information leads to serious consequences, such as misdiagnoses, incorrect treatment recommendations, or patient harm. This highlights the need for robust human oversight in AI applications, ensuring that professionals remain responsible for decision-making rather than uncritically relying on model outputs.

D. Transference and Ethical Issues

Transference refers to the capability of foundation models to apply knowledge from one domain or task to a related one, enabling efficient adaptation to new tasks through transfer learning [24], [55]. While this facilitates efficient adaptation, it also carries risks, including transferring algorithmic bias from one context to another [23], [56]. For example, the underrepresentation of African hospitals in international medical datasets exacerbates biases, such as racial and demographic biases, limiting the models' ability to generalize effectively across diverse populations and potentially hindering equitable healthcare outcomes. This under-representation highlights the issue of transference, where models trained on datasets lacking diversity may struggle to generalize to underrepresented populations. As a result, these models may produce inaccurate predictions for these groups, as the knowledge transferred from one domain, such as a population of predominantly lighter-skinned individuals, to another, such as African populations, fails to account for critical contextual differences. Furthermore, patients' data privacy is at risk when foundation models trained on specific populations are applied to diverse healthcare settings without proper consideration of individual patient circumstances or obtaining their consent.

E. Compositionality and Ethical Issues

The compositionality property of foundation models refers to their ability to flexibly integrate and reconfigure learned components or patterns, enabling them to generalize across new tasks, domains, or contexts. This ability allows the model to adapt its knowledge to novel situations, even without direct exposure to those scenarios during training [24], [57]. This capability enhances the model's ability to tackle new tasks with minimal task-specific data, contributing to zero-shot and few-shot learning. However, compositionality potentially leads to misinterpretation of complex scenarios and overconfidence in the model's capabilities, resulting in incorrect outcomes and posing ethical risks in life-critical applications like healthcare, where context and nuance are essential. This occurs because compositionality limits the expressivity of the representation, preventing it from accounting for unique semantics, exceptions, and context-driven correlations [23], [58].

F. Explainability and Ethical Issues

The explainability characteristic of foundation models refers to their ability to provide transparent and interpretable reasoning behind their decisions, enabling users to understand how the model arrives at its outputs [23], [24]. This characteristic is crucial in fields like computational pathology, where clear explanations of model predictions are necessary for ensuring trust and informed decision-making [38]. However, the complexity of foundation models, particularly in high-dimensional tasks like medical diagnosis, often makes explainability difficult, undermines trust in the model's predictions, and raises concerns regarding the transparency of the algorithm as well as accountability [59]. This lack of clarity directly impacts information transparency, as patients should be informed about how AI contributes to their diagnosis and treatment decisions. Without clear explanations, patients may lose trust in the AI-driven processes involved in their care, potentially leading to harmful consequences such as misinformed decisions or delays in treatment.

IV. Foundation Model Applications in Pathology

This section explores the various applications of foundation models in computational pathology. While the use of these models in this field is still evolving, significant advancements

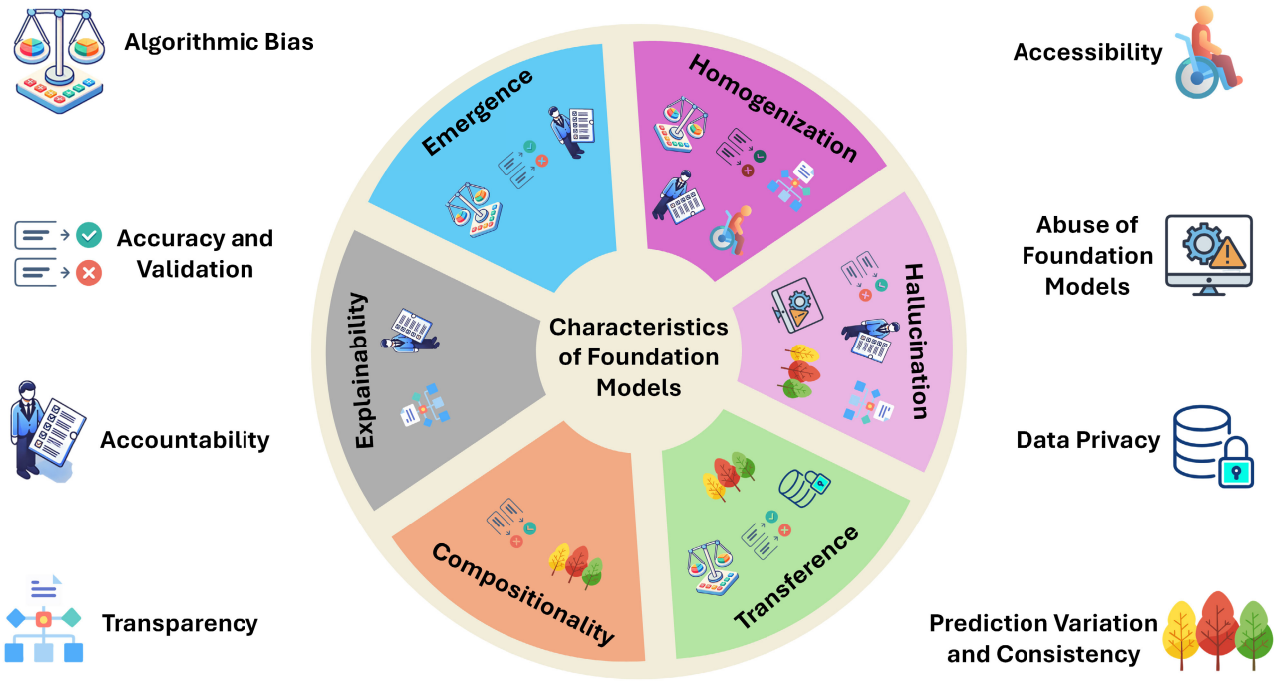

Fig. 4. Ethical issues relating to the key characteristics of foundation models. The ethical issues covered include algorithmic bias, accuracy and validation, accountability, transparency, accessibility, abuse of AI models, data privacy, and prediction variation and consistency.

have been made in developing models tailored to address specific challenges. These developments span a wide range of tasks, ranging from disease identification [60], [61], cancer sub-type prediction tasks [62], [63], [64], semantic segmentation [65], [66], information extraction [67], and many more.

A. Prov-GigaPath

Prov-GigaPath, a recently published foundation model by Xu et al. [60], is designed to address two major challenges that hinder the development and implementation of pathology foundation models in real-world clinical applications. Firstly, designing a model architecture that effectively captures both local patterns in individual tiles and global patterns across WSIs remains challenging. Existing models often treat each image tile as an independent sample and approach slide-level modeling as multiple-instance learning. This method restricts the model's ability to capture the overall more complicated global patterns in gigapixel WSI. Secondly, in cases where pretraining has been performed on extensive patient data from real-world settings, the resulting foundation models are typically not publicly accessible. This limitation restricts their broader applicability in clinical research and applications.

To tackle these issues, Prov-GigaPath constitutes a state-of-the-art vision transformer (ViT) [68] named DINOv2 [69] for pretraining large pathology foundation models on gigapixel pathology slides. DINOv2 enables embedding image tiles as visual tokens, effectively transforming a slide into an extended sequence of tokens. In this way, the model is pretrained at the image level using DINOv2 self-supervised learning. In contrast, at the whole-slide level, it employs a self-supervised learning method via a masked auto-encoder that learns from a sequence of tokens, thereby improving its effectiveness in capturing complex global patterns. Finally, since Prov-GigaPath is fully open-weight and publicly available, the second challenge of model accessibility is effectively resolved. By doing so, researchers and practitioners across the community will have equal access to the model, enabling collaborative advancements and enhancing its broader applicability in clinical research and applications.

Prov-GigaPath shows remarkable potential in improving tumor mutation prediction by leveraging this task as an image-classification task. In predicting 18 biomarkers that have the highest mutation occurrence for a pan-cancer setting, Prov-GigaPath achieved a 3.3% improvement in macro-area under the receiver operator characteristic (AUROC) and an 8.9% improvement in macro-area under the precision-recall curve (AUPRC) when compared to other best methods [60]. For the task of identifying nine major cancer subtypes, Prov-GigaPath also demonstrated better performance than other competitive models, suggesting that the integration between DINOv2 and auto-encoders improves the extraction of meaningful features at both the image-level and whole-slide-level.

B. UNI

Despite the advent of various foundation models designed to address challenges in computational pathology, only a few are capable of generalizing tasks across different domains [70]. The development of UNI by Chen et al. [62] introduces a general-purpose, versatile ViT model that utilizes *transfer learning* [71] to combine multiple tasks, including ROI-level classification, segmentation, image retrieval, and slide-level weakly supervised learning. Transfer learning is an ML technique in which a model's knowledge learned from one task is reused to improve its performance on another related task. Using transfer learning, UNI shows substantial performance

uplift in various diagnostic tasks, such as cancer detection, cancer grading and subtyping, nuclear segmentation, organ transplant assessment, and several pan-cancer classification tasks. The UNI model is pretrained on Mass-100K, a dataset of over 100 million tissue patches from 100,426 H&E whole-slide images (WSIs) spanning 20 tissue types. It is evaluated on 34 computational pathology tasks, including cancer subtyping, biomarker screening, and segmentation, using diverse datasets like the OncoTree cancer classification system and curated slides from Brigham and Women's Hospital, ensuring robust testing of its generalization capabilities.

C. CONtrastive Learning From Captions for Histopathology (CONCH)

After discussing UNI, we must also highlight another foundation model known as CONCH, as both models were developed by researchers at Harvard Medical School's Brigham and Women's Hospital and presented in two companion papers published in Nature Medicine [72]. Although both models share similar goals in attempting to overcome the limitations of usage scenarios common in many current AI systems, Lu et al. [63] developed and trained CONCH to understand both pathology images and language. CONCH is trained on a database comprising more than 1.17 million image-text pairs, enabling pathologists to query tissue sample images according to certain features of interest. In their study, Lu et al. analyzed the accuracy of CONCH in recognizing up to 30 categories of brain tumors, which are all classified as rare cancer types according to the RARECARE project's definition [73]. The results show that CONCH is able to produce strong-performing classification accuracy when combined with weakly supervised learning, achieving a balanced accuracy of 68.2% that surpasses the vision-only self-supervised learning CTransPath model as well as other visual-language pretrained models, including PLIP [74], OpenAICLIP [75], and BiomedCLIP [63], [76].

D. Segment Anything Model (SAM)

Semantic segmentation of pathological entities holds significant clinical value in computational pathology workflows. Semantic segmentation involves dividing sample images into discrete regions corresponding to various tissue structures, cell types, or sub-cellular components. Accurate and efficient semantic segmentation is critical for multiple pathological applications, including tumor detection, grading, prognosis, and examining tissue architecture and cellular interactions.

The Segment Anything Model (SAM) is a recently developed foundation model by Kirillov et al. [65] from Meta AI, designed for universal application in segmentation tasks. SAM is inspired by Natural Language Processing (NLP) models with a unique characteristic of utilizing prompt engineering. Hand-crafted text is used to prompt the language model to generate a valid textual response. Similarly, SAM takes in segmentation prompts, such as various sub-cellular structures, and then returns valid segmentation masks correspondingly. SAM aims to develop a promptable model capable of generalizing segmentation tasks. SAM is pretrained on a dataset (SA-1B) comprising over 1 billion masks across 11 million images, enabling it to segment objects based on various user-defined features, including dots, bounding boxes, and text. SAM's evaluation highlights its impressive zero-shot performance (the ability to complete a task without having received any prior training examples), often matching or even exceeding previous fully supervised models across a wide range of tasks. Under such conditions, SAM is fine-tuned to perform specialized semantic segmentation tasks crucial in computational pathology.

E. Segment Anything for Microscopy (μ SAM)

Archit et al. [77] recently introduced μ SAM, a foundation model designed to improve segmentation and tracking in multi-dimensional microscopy data. Leveraging Meta AI's Segment Anything Model (SAM) [65], μ SAM is fine-tuned specifically for microscopy applications, enhancing segmentation quality under various imaging conditions. It supports interactive and automatic segmentation for 2D and 3D data and tracking for time-series data. Additionally, μ SAM demonstrates improved results compared to other proposed models fine-tuned on tasks such as segmenting cells and nuclei in light microscopy and mitochondria in electron microscopy. This advancement represents a significant step forward in utilizing vision foundation models for microscopy, aiming to simplify image analysis in biological research.

F. Pathology Language and Image Pretraining (PLIP)

As mentioned, a significant challenge in training AI models in pathology is the lack of large-scale annotated publicly accessible medical images. To address this issue, Huang et al. [74] sought opportunities on public forums and crowd platforms, such as medical Twitter, to collect pathology images. As a result, they created OpenPath, the largest publicly available dataset of pathology images annotated with natural text, comprising over 208,414 images. To demonstrate the utility of OpenPath, Huang et al. [74] designed and trained a visual-language foundation model called PLIP, utilizing this dataset for its training.

Unlike other supervised learning and segmentation pathology models trained solely on categorical labels, visual-language models utilize both image data and semantic knowledge from corresponding natural text, making PLIP perform exceptionally well in zero-shot image classification tasks. During the training phase, the PLIP model generates two embedding vectors using both the image and text encoders. These are then optimized through contrastive learning to be similar for each paired image and text vector and dissimilar for non-paired images and texts. By leveraging the benefits of contrastive learning and semantic descriptions from the OpenPath database, PLIP is capable of handling a wide range of inferences across various medical applications. As an example, when given an image and multiple disease descriptions, PLIP identifies which description best matches the image, making PLIP a powerful tool for computational pathology. This functionality does not require explicit training and differentiates PLIP from other supervised foundation models.

G. PathologyBERT

Another challenging problem in computational pathology is text mining, which refers to the process of transforming unstructured text into a structured format in order to identify meaningful patterns effectively. Text mining is a difficult task due to the variability in the structure and format of reports and the frequent introduction of new cancer subtype definitions in pathology. The development of more advanced NLP models promotes a better understanding of contextual relationships in pathology text mining by utilizing attention-based Encoder-Decoder architectures. One of the most popular modern NLP models is the Bidirectional Encoder Representations from Transformers (BERT) [78], a contextualized language representation model employing a multi-layer bidirectional encoder. The BERT model constitutes a transformer neural network that uses parallel attention layers instead of sequential recurrence, which enables BERT to be capable of representing words or sequences in ways that capture contextual information, allowing the same sequence of words to have different representations and meanings depending on the context in which they appear.

PathologyBERT, developed by Santos et al. [67], is a specialized adaptation of the BERT model designed to address the shortcomings of general language models in pathology by incorporating domain-specific knowledge. Unlike standard transformer models that rely on generic medical vocabulary, PathologyBERT is pretrained on pathology-specific texts, allowing it to understand specialized terminology and contextual nuances better. By leveraging a domain-adapted vocabulary and fine-tuned tokenization strategies, PathologyBERT improves performance on pathology-related NLP tasks, such as masked language prediction, information extraction, and report classification. This specialization allows it to overcome the limitations posed by traditional WordPiece [79] tokenization, making it a more effective tool for pathology-related text analysis.

H. Med-Gemini

In this section, we will focus on the pragmatic aspects of a particular foundation model in helping pathologists perform their tasks more effectively. One of the most well-known AI models recently is ChatGPT, developed by OpenAI, with its latest and most powerful iteration named GPT-4 [46]. GPT-4 has significantly advanced natural language understanding and generation, featuring improved performance, accuracy, and contextual comprehension. As a result, it has inspired the development of numerous fine-tuned models tailored to specific fields of study, such as Med-Gemini [80], the successor to Med-PaLM 2 [81], both developed by Google Research.

Unlike Med-PaLM 2, which is primarily a large language model (LLM) designed for text-based tasks, Med-Gemini is a family of multimodal foundation models that integrate both text and image data. This expanded capability allows Med-Gemini to provide more comprehensive insights for healthcare professionals, including clinical decision support, patient history summaries, and evidence-based treatment recommendations. By incorporating visual data alongside textual information, Med-Gemini offers a more holistic approach to medical diagnostics and decision-making than its predecessor, which relied solely on text-based inputs. It also serves as a first-line source of medical information for patients, answering common health questions and guiding them when to seek professional care. Additionally, Med-Gemini excels at extracting information from clinical notes and reports, an error-prone task for pathologists. For example, determining the presence or absence of cancer in a report is challenging due to context-sensitive terms like "carcinoma," leading to confusion or mistakes [82].

A significant advantage of Med-Gemini is its multimodal capabilities, which combine text and image data to provide more accurate, context-aware insights. Analyzing both clinical notes and medical images allows it to identify patterns that text-only models might overlook. Large multimodal foundation models like Med-Gemini excel in these scenarios [83], [84], integrating complex data to capture subtle nuances, enhancing diagnosis accuracy, and improving decision-making in clinical settings.

I. Clinical Histopathology Imaging Evaluation Foundation (CHIEF) Model

The CHIEF model, developed by Wang et al. [70], is a general-purpose AI framework designed to support cancer diagnosis and prognosis. Unlike traditional models tailored to specific diagnostic tasks, CHIEF utilizes both unsupervised and weakly supervised pretraining methods on a large, diverse dataset of histopathology images, enabling it to recognize a wide range of pathology features. CHIEF demonstrated enhanced performance and generalizability in cancer cell detection, tumor origin identification, molecular characterization, and survival prediction, outperforming existing deep learning models across multiple independent datasets. CHIEF's application extends to survival prediction, where it reliably stratified patients based on prognosis in both training and independent datasets, offering insights into morphological indicators of survival outcomes.

V. Ethical Challenges in Foundation Model Applications

The ethical issues and risks outlined in Section III for each foundation model application are discussed here, along with possible solutions for addressing these issues. Many of these ethical issues are interconnected, and addressing one often contributes to resolving others. Table I illustrates the relationships between each foundation model, its applications, and their correlating ethical challenges. It is important to note that shared ethical issues may exist across different models. Fig. 5 illustrates the relationships between each foundation model discussed, their applications in pathology, and their corresponding ethical issues.

A. Ethical Issues of Prov-GigaPath

Prov-GigaPath, by nature, is a foundation model that integrates the DINOv2 transformer with its masked autoencoder to capture both local patterns in individual pathology

TABLE I
SOME TYPICAL EXAMPLES OF PATHOLOGICAL FOUNDATION MODELS, APPLICATIONS, AND RELATED ETHICAL ISSUES

Foundation Models	Applications	Common Ethical Issues
Prov-GigaPath [60]	Classification tasks, Diagnosis report, Prediction	Accountability
UNI [62]	ROI classification, ROI retrieval, Prediction, Segmentation, Slides classification	
CONCH [63]	Disease classification, Few-shot classification tasks, Prediction, Segmentation, Zero-shot image-to-text, Zero-shot text-to-image	Abuse of Foundation Models
SAM [65]	Zero-shot semantic image segmentation	Data Privacy
µSAM [77]	Zero-shot semantic image segmentation	
PLIP [74]	Prediction, Text-to-image, Zero-shot semantic image classification	Transparency
PathologyBERT [67]	Classification tasks, Language prediction, Text mining	
Med-Gemini [80]	Classification tasks, Diagnosis report, Image-to-text, Prognosis, Zero-shot capabilities	Accessibility
Med-PaLM 2 [81]	Diagnosis report, Few-shot multi-choice queries, Prognosis	Prediction Variation and Consistency
CHIEF [70]	Biomarker prediction, Classification, Prognosis	

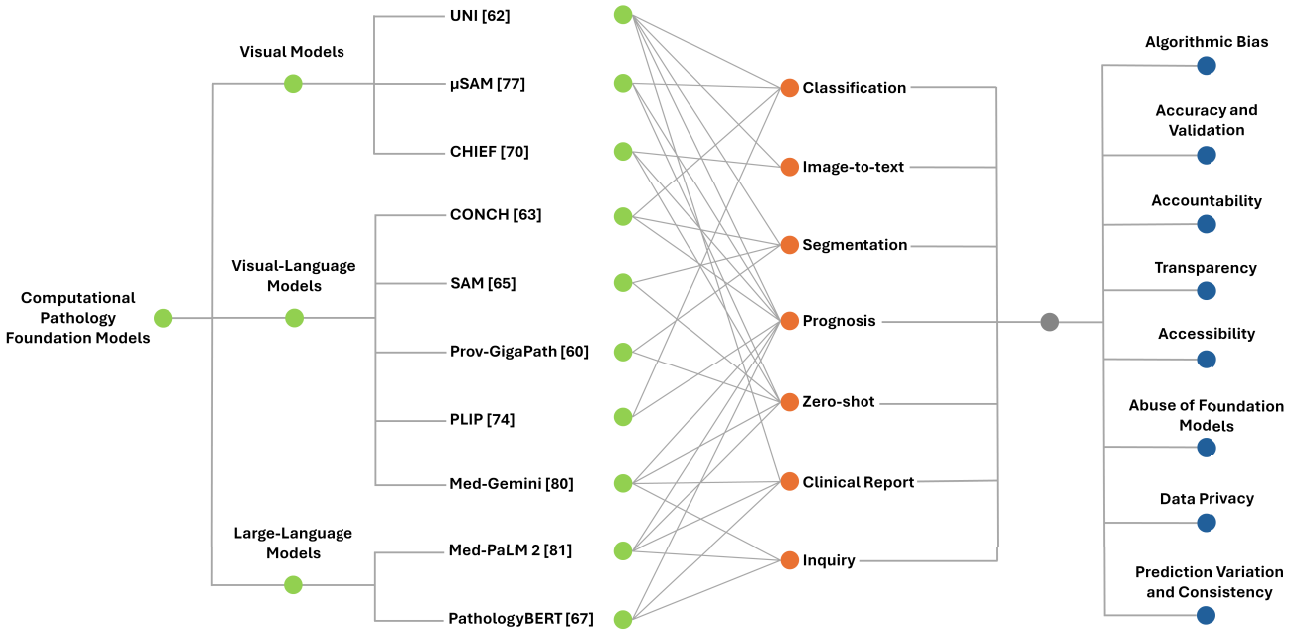

Fig. 5. A breakdown of the list of typical foundation models in this paper and their associated ethical issues. The left side of the figure shows the computational foundation models discussed, divided into three groups according to their types. Various tasks the models perform are shown in the center of the figure. On the right side of the figure, common ethical challenges of the discussed foundation models are shown.

images and global patterns at the whole-slide level. Since Prov-Gigapath is mostly a prediction model for pathological diseases, there are inevitably ethical implications for using it at a clinical level, which will be analyzed in the following subsections.

1) Algorithmic Bias: Firstly, Prov-GigaPath is subjected to algorithmic bias since Prov-Path, the database it pretrained on, may contain degrees of bias due to data gathered from more than 28 different cancer centers. In particular, the patient data extracted from the 28 cancer centers may include varying degrees of bias related to race, ethnicity, age, gender, health conditions, and other demographic factors of the patients from whom the data is collected. As a result, biases from the data source may propagate to Prov-Path, creating uncertainty about the extent of bias introduced when training Prov-GigaPath on these pathological image tiles [36]. Consequently, while Prov-GigaPath may excel at accurately predicting various diagnostic tasks, there remains a risk that its performance may be suboptimal for specific patient groups due to algorithmic bias [85], [86]. This raises significant ethical concerns regarding accuracy and accountability in patient care. However, the effects of algorithmic bias may be mitigated if strict quality control is implemented on the training data. For instance, experienced pathologists and related professionals play a crucial role in analyzing datasets to ensure high data quality standards before they are trained on models such as Prov-GigaPath.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2025.3551913

2) Accuracy and Validation: Given that Prov-GigaPath may contain algorithmic bias, its results directly impact a patient's ministration. It is crucial that thorough testing is done to validate the accuracy of its predictions. For example, having a competent and recognized third party to test and validate the model is desirable. Such organizations include The Academy of Clinical Laboratory (ACLPS), The American Society for Clinical Pathology (ASCP), and many more.

B. Ethical Issues of UNI

The uniqueness of the UNI model lies in its utilization of transfer learning to generalize a variety of tasks relating to pathological domains. However, the concern arises as to whether UNI's performance in each of these tasks matches or even exceeds that of other models specialized in performing specific tasks. Therefore, the most critical ethical issues that need to be addressed by UNI are accuracy and validation.

1) Accuracy and Validation: Ensuring the quality of UNI's results requires a robust and systematic testing procedure that includes comprehensive evaluation across various tasks and datasets. UNI has already established a thorough benchmarking framework, incorporating several well-regarded benchmarks such as the OT-108, WILDS, and ChampKit. These benchmarks enable the comparison of UNI's performance against other specialized models tailored to specific tasks within computational pathology, including disease classification, segmentation, and diagnosis prediction [87], [88]. The evaluation is based on various metrics, such as precision, recall, F1-score, and area under the curve (AUC), to assess UNI's effectiveness on each task.

Additionally, it is essential to implement cross-validation techniques [66], [89], where UNI is trained on one dataset and tested on multiple unseen datasets from different populations or medical conditions. This would allow a proper assessment of how well UNI generalizes across diverse real-world scenarios. The model's performance on rare or minority cases, such as specific cancer subtypes, should also be closely evaluated to ensure that the model does not introduce bias or underperform in these areas, a common issue in AI models trained on unbalanced datasets.

C. Ethical Issues of CONCH

Since CONCH's objectives are almost identical to UNI's, we will also focus on the accuracy and validation concerns of the CONCH model's philosophy. However, with the additional query functionality that allows pathologists to search for particular pathological images based on features of interest, the accuracy of such queries performed by CONCH must also be examined.

1) Accuracy and Validation: Similar to UNI, to validate the accuracy of CONCH's capability to generalize distinct tasks, it needs to be tested in parallel with supplementary models for each type of task. Testing CONCH's query feature should be a relatively simple task, as pathologists provide CONCH with test query prompts and observe the algorithm's accuracy with the expected images.

D. Ethical Issues of SAM

SAM is a promptable model that utilizes prompt engineering to generalize segmentation tasks, including semantic segmentation, that is commonly used in computational pathology AI applications. In the case of performing semantic segmentation tasks, SAM is challenged to divide sample images into discrete regions according to numerous tissue structures, cell types, or sub-cellular components. Therefore, the accuracy of SAM's semantic segmentation determines its success or failure at performing its job. As SAM is pretrained on the SA-1B dataset constructed by the model's founders, the likelihood of dataset bias cannot be overlooked. Consequently, algorithmic bias is also another potential issue of SAM.

1) Algorithmic Bias: The SA-1B dataset, used to pretrain SAM, raises significant concerns due to its lack of transparency. For instance, the dataset claims to contain over 1 billion masks for 11 million licensed pathological images, yet there is minimal publicly available information regarding its data collection process, curation methodology, or the diversity of its sources. The only known detail about its origin is that the dataset was sourced from a large photo company, as stated on Meta AI's official website: https://ai.meta.com/datasets/segment-anything. This lack of transparency regarding the dataset's origin is a critical issue, as it raises doubts about the dataset's diversity, representativeness, and potential biases. Without clear documentation about the diversity of the data sources and how it was curated, it is difficult to assess the degree of bias present in the dataset, which may adversely affect the performance of SAM, especially when applied to new, diverse, or underrepresented medical datasets.

To mitigate these concerns, it is crucial to establish a rigorous auditing process for the dataset, focusing on identifying and addressing biases across different demographic groups, diseases, and imaging modalities. One potential solution is conducting a thorough bias analysis, including testing the model on data from diverse populations, ensuring that the dataset adequately represents various ethnic groups, age ranges, and medical conditions. Moreover, data augmentation techniques could be employed to artificially balance underrepresented classes and reduce bias in the training data [90], [91].

Additionally, it is essential to implement continuous monitoring of SAM's performance when deployed in real-world healthcare settings, especially in regions or populations not represented in the training data. Regular performance evaluations, along with the integration of feedback loops from clinicians, could help identify any emerging biases or discrepancies in the model's outputs. These feedback mechanisms should be designed to allow for dynamic model updates that adapt to new data, ensuring that the model evolves with changing clinical practices and diverse patient populations.

2) Accuracy and Validation: As SAM is pretrained on a vast dataset, it has the confidence to perform general segmentation tasks such as differentiating pedestrians from vehicles, roads, trees, etc. On the other hand, SAM's effectiveness in executing more specialized semantic segmentations has yet to be proven. As a result, the direct application of SAM on

segmenting pathological images should be exercised with substantial caution due to the need for more training and testing on pathology datasets. Consequently, the application of SAM on semantic segmentation tasks remains a primary ethical concern regarding its segmentation accuracy. To address this, future endeavors to apply SAM to semantic segmentations are encouraged, and with enough testing and fine-tuning, SAM has the potential to become a reliable model for performing tasks within the domain of computational pathology.

E. Ethical Issues of μ SAM

As μ SAM is considered an extension of SAM and is also used for segmentation tasks, it shares the same ethical challenges as SAM, notably algorithmic bias and segmentation accuracy. The fine-tuned nature of μ SAM gives it an advantage in identifying features of objects in microscopy images, such as cells and nuclei in light microscopy (LM) or cells and organelles in electron microscopy (EM) [77]. Although μ SAM is trained on LiveCELL [92], one of the largest publicly available datasets for cell segmentation, there is still potential for sampling biases. The fact that μ SAM is one of the first foundation models to apply segmentation tasks to microscopy images means that its accuracy and validation will be iteratively improved. Therefore, in order to enhance performance, μ SAM should be trained on additional microscopy datasets and thoroughly evaluated for accuracy.

F. Ethical Issues of PLIP

PLIP is a unique foundation model in that it is trained on datasets gathered from crowdsourcing platforms, including medical Twitter. The advantage of PLIP over other competitive models, such as SAM, is that it is a visual-language model that leverages both image data and semantic knowledge from corresponding natural text, enhancing its performance on image classification tasks. However, the crowdsourced database that it is trained on raises concerns for accuracy and validation and algorithmic bias, which is similar to that of Prov-GigaPath.

1) Algorithmic Bias: Pathological data collected from public sources, such as medical Twitter, is inherently more vulnerable to bias and imprecision since there are more inconsistencies with the quality of data from sources that may or may not be credible. The decision to use a crowdsourced database stems from PLIP's aim to address clinical pathology data scarcity. As a result, assessing the extent of bias within PLIP's database remains challenging.

One potential solution to mitigate these issues is implementing bias detection and correction algorithms. These algorithms analyze the dataset to identify patterns of bias, such as underrepresentation of certain groups, and adjust for these biases during the training process. This helps mitigate issues related to biased data in the publicly sourced dataset and enhances the overall fairness and accuracy of the model's predictions.

2) Accuracy and Validation: An immediate disadvantage of using a publicly sourced database, such as medical Twitter, is the difficulty in validating the legitimacy of its data sources. Consequently, PLIP's crowdsourced database may contain fraudulent or inaccurate information, potentially compromising the accuracy of its results. To address these challenges, it is essential to implement robust data validation procedures. These procedures could involve cross-referencing publicly sourced data with verified clinical databases to identify and eliminate inaccuracies or inconsistencies. Additionally, expert review by medical professionals should be incorporated to evaluate the credibility of the data and ensure its alignment with current medical knowledge. Such a validation process will help improve the accuracy and reliability of PLIP, ensuring that the model produces trustworthy results in clinical applications.

G. Ethical Issues of PathologyBERT

PathologyBERT is a robust foundation model that excels in text-mining tasks for computational pathology applications. However, since it exclusively processes textual information, it is susceptible to biases within word embeddings.

1) Algorithmic Bias: Biases in word embeddings are often challenging to detect. For instance, word embeddings were widely used across industries before their hidden stereotypical biases were discovered [93]. Although the existence of these word embedding biases is now recognized, the process of how these biases are learned from training data is not well understood. For this reason, extra caution should be exercised when analyzing results from PathologyBERT to account for potential word embedding bias, especially if the data contains sensitive subjects of matter.

H. Ethical Issues of Med-Gemini

Med-Gemini is a compelling model that is excellent at natural language understanding and is capable of providing relatively intelligent answers to prompts and questions. However, models as powerful as Med-Gemini are more easily abused by users.

1) Abuse of Foundation Models: The generative capability of Med-Gemini is particularly prone to human abuse. For instance, excessively or carelessly relying on Med-Gemini's responses in the field of pathology could result in unintentional misuse or abuse of its algorithm, potentially leading to severe consequences such as incorrect diagnoses and the oversimplification of complex medical conditions. In cases where multimodal models like Med-Gemini produce seemingly sound answers to prompts, they give users a false sense of trust when, in reality, they provide answers that deviate significantly from the truth [94], which relates to the hallucination issue of foundation models.

Recent research has shown that despite advancements, multimodal large language models often produce outputs inconsistent with input data, especially in high-accuracy fields like healthcare [95]. These models sometimes generate information that seems plausible but is ultimately misleading. Potential solutions include improving cross-validation across modalities, enhancing multimodal training with integrated data, and performing regular audits in real-world settings to ensure consistent and reliable outputs. These measures help reduce hallucinations and improve the trustworthiness of multimodal foundation models.

I. Ethical Issues of CHIEF

The nature of CHIEF as a pure classification model and its reliance on weakly supervised learning makes its predictions challenging for clinicians to interpret. This lack of interpretability leads to potential accountability concerns.

1) Accountability: Since CHIEF's decision-making process is not fully explainable, it becomes difficult to pinpoint the source of errors or misdiagnoses. In medical applications, where AI systems influence critical decisions, it is essential to establish clear accountability. If CHIEF produces an incorrect diagnosis or recommendation, determining whether the error lies with the model, the dataset, or the human user interpreting the output is challenging. This ambiguity regarding responsibility complicates legal and ethical considerations, particularly in cases where inaccurate AI-driven decisions cause harm to patients. Clear guidelines must be established to delineate the roles of both developers and clinicians in overseeing and validating model outputs. Additionally, incorporating human-in-the-loop systems, where clinicians review AI suggestions, ensures that accountability for patient care remains with healthcare providers.

J. Common Ethical Issues of Pathological Foundation Models

The following ethical issues apply to all foundation models introduced in Section IV, as they present common challenges in computational pathology. This subsection examines these concerns in greater detail to highlight their impact and potential solutions.

1) Accountability: In unforeseen and unfortunate scenarios where an AI model fails at its designed tasks and causes adverse consequences, the question of responsibility arises. This is one of the most complicated yet crucial ethical issues involving pathological foundation models, as the lives of patients seeking treatments may be at the mercy of some of these models in the most critical cases. Several factors contribute to the potential failure of foundation models, including defect algorithms, biased input data, improper operation or application, or other elements, including human errors.

Since foundation models cannot be held responsible directly in cases of failure, accountability falls upon the human factors in the design, implementation, deployment, and use of these models. Accountability ensures that if an AI model makes a mistake or causes harm, a responsible party is identified, whether it be the designer, developer, the organization using the model, or a combination of these. In cases where damages have already occurred, accountability is crucial to ensuring that the affected victims and their families receive adequate compensation for their loss. Simultaneously, there is also a dire need for appropriate laws and judicial rules to be reviewed and updated frequently to uphold justice and fairness in AI systems.

2) Abuse of Foundation Models: Although recorded cases of abuse of AI algorithms relating to pathology are rare, their potential impact on patients and society should not be underestimated. Human abuse of AI models falls into two categories: intentional and unintentional abuse, both of which have the potential to result in adverse consequences. In regards to pathology, an example of deliberate abuse of AI technology might be the unethical and unauthorized use of pathological data for malevolent intentions, such as developing biological weapons.

As an example of an unintentional abuse of AI models, an incorrect application of a foundation model caused by human error may be life-threatening to patients who have placed their trust in the therapy. Unfortunately, there are countless examples of human abuses of AI technology, even within the domain of pathology. Therefore, society must be prepared to address AI applications' potential risks and consequences. Providing guidelines, strict management, and protection of clinical data and AI models is a crucial first step in minimizing the risk of intentional misuse of these models.

Due to its subtle nature, unintentional abuse of AI models caused by human error is more difficult to resolve. Establishing appropriate testing and validation principles and procedures should decrease the chance of algorithmic errors. Providing adequate and abundant training to pathologists and other professionals who utilize foundation models is an excellent approach to reducing the possibility of human errors.

3) Data Privacy: In computational pathology, the scarcity of real-world clinical data creates a dilemma: On the one hand, there is an urgent need for clinical data, but on the other hand, the more data gathered, the more likely that patients' privacy is violated. To address this dilemma, pathologists must ensure respect for privacy and data protection when applying AI systems throughout their lifecycles. This involves implementing effective administration and management for all data used and generated by the AI systems. In particular, the collection, usage, and storage of all sensitive data must meet compliance with relevant data privacy laws and regulations.

Additionally, data and algorithms must be safeguarded against theft. In the event of data leakage or loss, the responsible party must promptly inform the affected individuals to minimize the loss or impact. To address data privacy concerns, one of the main approaches to privacy-preserving machine learning algorithms and data analysis, called differential privacy, is introduced [96]. Differential privacy is a mathematical framework designed to protect individuals when their data is being used in data sets.

Differential privacy ensures that a data analyst cannot obtain additional information about any individual after analyzing the data. It also ensures that other analysts will not form significantly different perceptions of an individual after accessing the database. Differential privacy operates by introducing randomness into the dataset, which does not impact the overall analysis. Furthermore, current and future sources of auxiliary information from other datasets must not compromise individual privacy.

Furthermore, a new ML paradigm called federated learning has been proposed to mitigate the risk of privacy leakage in ML processes [97]. Federated learning focuses on settings where a single ML model is collaboratively trained by different clients using decentralized, heterogeneous data from each client, and the model is iteratively improved until it is fully

trained. This way, there is no data exchange between clients, so the chance of leaking more sensitive data is minimized.

4) Transparency: Transparency, understood as both algorithmic and informational, is critical to the responsible use of machine learning (ML) and foundation models. Algorithmic transparency concerns the technical explainability of these models, focusing on how and why an algorithm produces particular outputs. One of the core challenges in achieving algorithmic transparency is the "transparency problem" [98], which refers to the inherent difficulty in understanding and reasoning about the inference processes of complex ML systems. This lack of clarity undermines end-user trust in the produced outputs. To overcome this issue, proposals have centered on three key techniques: model approximation techniques, visualization techniques, and intrinsic explanation methods.

- Model approximation techniques: Involve using simpler, more interpretable models, such as decision trees or linear regression, to approximate the behavior of complex models. Examples of popular model approximation techniques include LIME (Local Interpretable Model-Agnostic Explanations) [99], Anchors [100], and the concept of knowledge distillation through the teacher-student model framework [101].
- Visualization techniques: Aim to help users better understand AI models visually. Feature importance visualizations constitute bar charts or heatmaps that help users identify which variables most influence a model's decisions, such as the Grad-CAM (Gradient-weighted Class Activation Mapping) method [102]. Meanwhile, DeepLIFT is a method for visualizing and understanding the contributions of individual neurons to a model's output through backpropagation, offering a way to interpret complex neural networks [103]. Other visualization techniques, such as t-SNE, are used to visualize high-dimensional data, providing insights into how models process and interpret the data [104].
- Intrinsic explanation methods: Design AI models with architectures that incorporate interpretability into their structure, primarily using attention mechanisms that highlight the most relevant components of the input for a prediction. ASDNet [105], MedSkip [106], and DeepCIN [107] are examples of such models that utilize attention mechanisms to generate predictions.

In addition to these proposed solutions, there are auxiliary measures that further mitigate the transparency problem. For instance, developers are encouraged to provide publicly accessible documentation and guides that help end-users better understand their models. At the same time, users should take responsibility for thoroughly learning about the model and recognizing the associated risks and potential consequences.

Information transparency, particularly relevant in healthcare, is essential to ensuring that stakeholders are fully informed about the role of AI in clinical decision-making. Patients, for instance, must be explicitly informed if AI technologies have been utilized to diagnose or treat their conditions. Such disclosures are ethically imperative as they uphold patient autonomy and reinforce trust between patients and medical professionals.

Equally important is the transparent handling of patient data. The collection and use of personal information without informed consent violate privacy and erode confidence in AI-driven healthcare systems. To address this, organizations should adhere to established data privacy frameworks such as the General Data Protection Regulation (GDPR) [108] or the Health Insurance Portability and Accountability Act (HIPAA) [109] in their respective jurisdictions. These frameworks provide guidelines for obtaining informed consent, anonymizing sensitive data, and ensuring data usage aligns with the patient's consented purposes. Mitigating these risks further requires the implementation of clear regulatory guidelines and ethical data management practices. This involves adopting methodologies like the Privacy Impact Assessment (PIA) to evaluate the potential risks associated with data collection and usage.

By integrating algorithmic and information transparency, practitioners create an environment of trust, facilitate informed participation from all stakeholders, and promote responsible innovation. Specifically, methodologies such as the Explainable AI (XAI) program are applied to enhance algorithmic transparency, ensuring that AI decision-making processes are interpretable and accessible to clinicians and patients alike [110]. Together, these measures provide a robust foundation for the ethical and transparent use of machine learning and foundation models in healthcare.

5) Accessibility: AI foundation models have the potential to transform healthcare by improving accessibility and efficiency, but they may also deepen global disparities. Advanced AI models enhance diagnostics and personalize treatment in well-resourced areas, yet they often require significant investment and infrastructure lacking in low-resource settings. Addressing this disparity requires a commitment to developing AI tools that are adaptable, affordable, and accessible globally. This involves creating technology in diverse and resource-limited environments while fostering collaboration, training, and policy support to ensure equitable distribution. The responsibility of evaluating AI models falls under several parties, such as pathology associations, research organizations, governments, and public safety agencies. In doing so, AI models are leveraged to improve healthcare outcomes across all populations, helping to close existing gaps and contribute to a more inclusive, globally equitable healthcare system.

6) Prediction Variation and Consistency: Given the current experimental results, combining AI and human expertise offers a viable approach to achieving diagnostic accuracy while maintaining a safety net in clinical imaging. Incorporating different experiences from various pathologists may add to the diversity of predicted results.

VI. Future Implications and Conclusion

In prediction tasks involving pathological image foundation models, it is evident that numerous ethical issues remain and require solutions based on the discussions above. It is crucial to mitigate, prevent, and prepare for the potential consequences that may impact our well-being in the present and future. Given the unpredictable and unforeseeable nature of AI advancement, it is impossible to provide an exhaustive list of ethical challenges, encompassing those currently affecting

society and new challenges that may emerge in the future. Nevertheless, various perspectives exist on how AI and foundation models evolve to align more closely with established ethical principles. We have summarized these perspectives into four key areas: multidisciplinary collaboration, technological approaches, societal engagements, and continuous monitoring and evaluation.

AI ethical issues are complex and multifaceted, involving not only technical challenges but also social, legal, and philosophical considerations. Undoubtedly, the discipline of AI ethics, both within and beyond the field of computational pathology, requires collaboration among multiple parties, including AI scientists, engineers, ethicists, governments, and the general public. Integrating expert knowledge will achieve a more holistic understanding and evaluation of AI ethics. For example, ethicists highlight the moral implications of a new foundation model deployment, while social scientists examine its impact on various communities.

Regarding technological approaches, AI scientists and developers should embrace ethical AI design and development methodologies, such as WHO guidance on "Ethics and governance of artificial intelligence for health" [111]. This includes adhering to established ethical principles, such as transparency, accountability, and fairness [112]. Organizations should establish ethics committees to review and approve AI projects, ensuring they align with these ethical principles. Additionally, human-in-the-loop (HITL) methods should be implemented to incorporate human oversight into foundation models, particularly in high-stakes applications, to ensure that ethical decisions are made [44], [113], [114].

Addressing AI ethical challenges requires a collective effort from society, particularly through regulation, policy development, and enforcement. Government regulations such as the EU Artificial Intelligence Act [115] are pivotal in setting standards for the ethical development, deployment, and oversight of foundation models. These regulations establish clear transparency, accountability, and fairness guidelines, ensuring that AI technologies are used responsibly.

In computational pathology, such policies could mandate transparency in the use of AI for diagnostic tasks by requiring healthcare providers and developers to disclose how AI systems are integrated into clinical workflows. For example, regulations might compel developers to provide detailed documentation on the datasets used for training models, the algorithms' decision-making processes, and the limitations or uncertainties inherent in AI-generated outputs. This transparency would enable clinicians to understand better and trust the system's recommendations while allowing patients to make informed decisions about their care. Additionally, these policies could enforce the use of explainable AI methodologies to ensure that AI-generated results are interpretable and accessible to medical professionals and patients, fostering greater trust and accountability in healthcare applications.

Continuous monitoring and evaluation are crucial to maintaining the ethical integrity of AI systems throughout their lifecycle. This approach involves conducting regular ethical impact assessments to gauge the ongoing effects of AI on society, ensuring that any unintended consequences are promptly identified and addressed. Feedback loops are essential in this process, providing mechanisms for users, stakeholders, and developers to offer input and continuously refine the system. Post-deployment monitoring enables organizations to observe AI systems in real-world settings, adjust to new ethical challenges, and ensure continued alignment with evolving ethical standards.

We hope this article illuminates the intricate ethical challenges associated with applying foundation models to computational pathology, thereby raising public awareness of these critical and impactful issues. By embracing the perspectives on ethical AI development and regulation presented here, we foster a future for AI that is more transparent, less biased, and increasingly accountable and trustworthy. This is especially vital for foundation model applications in fields as significant and sensitive as computational pathology.

Author Contributions

Rui Fei Du: Writing; Eduard Lloret Carbonell and Jiaxuan Huang: Revision; Sheng Liu and Xiaohang Wang: Perception; Dinggang Shen: Conceptualization; Jing Ke: Conceptualization, re-writing, analysis, perception, supervision, and funding.

References

[1] K. Bera, K. A. Schalper, D. L. Rimm, V. Velcheti, and A. Madabhushi, "Artificial intelligence in digital pathology—New tools diagnosis and precision oncology," *Nature Rev. Clin. Oncol.*, vol. 16, no. 11, pp. 703–715, Nov. 2019.

[2] M. K. K. Niazi, A. V. Parwani, and M. N. Gürcan, "Digital pathology and artificial intelligence," *Lancet Oncol.*, vol. 20, no. 5, pp. e253–e261, Apr. 2019.

[3] F. Sobhani, R. Robinson, A. Hamidinekoo, I. Roxanis, N. Somaiah, and Y. Yuan, "Artificial intelligence and digital pathology: Opportunities and implications for immuno-oncology," *Biochimica et Biophysica Acta (BBA)-Rev. Cancer*, vol. 1875, no. 2, Apr. 2021, Art. no. 188520.

[4] M. Bilal, M. Nimir, D. Snead, G. S. Taylor, and N. Rajpoot, "Role of AI and digital pathology for colorectal immuno-oncology," *Brit. J. Cancer*, vol. 128, no. 1, pp. 3–11, Jan. 2023.

[5] L. Schneider et al., "Integration of deep learning-based image analysis and genomic data in cancer pathology: A systematic review," *Eur. J. Cancer*, vol. 160, pp. 80–91, Jan. 2022.

[6] A. H. Song et al., "Artificial intelligence for digital and computational pathology," *Nature Rev. Bioeng.*, vol. 1, no. 12, pp. 930–949, Oct. 2023.

[7] J. Lipkova et al., "Artificial intelligence for multimodal data integration in oncology," *Cancer Cell*, vol. 40, no. 10, pp. 1095–1110, Oct. 2022.

[8] A. Shmatko, N. Ghaffari Laleh, M. Gerstung, and J. N. Kather, "Artificial intelligence in histopathology: Enhancing cancer research and clinical oncology," *Nature Cancer*, vol. 3, no. 9, pp. 1026–1038, Sep. 2022.

[9] M. S. Hosseini et al., "Computational pathology: A survey review and the way forward," *J. Pathol. Informat.*, vol. 15, Dec. 2024, Art. no. 100357.

[10] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101813.

[11] J. van der Laak, G. Litjens, and F. Ciompi, "Deep learning in histopathology: The path to the clinic," *Nature Med.*, vol. 27, no. 5, pp. 775–784, May 2021.

[12] A. Duggento, A. Conti, A. Mauriello, M. Guerrisi, and N. Toschi, "Deep computational pathology in breast cancer," in *Seminars in Cancer Biology*, vol. 72. Amsterdam, The Netherlands: Elsevier, 2021, pp. 226–237.

[13] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, "The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis," *Comput. Biol. Med.*, vol. 128, Jan. 2021, Art. no. 104129.

[14] H. Go, "Digital pathology and artificial intelligence applications in pathology," *Brain Tumor Res. Treatment*, vol. 10, no. 2, pp. 76–82, 2022.

[15] M. Cui and D. Y. Zhang, "Artificial intelligence and computational pathology," *Lab. Invest.*, vol. 101, no. 4, pp. 412–422, Apr. 2021.

[16] S. Morales, K. Engan, and V. Naranjo, "Artificial intelligence in computational pathology – challenges and future directions," *Digit. Signal Process.*, vol. 119, Dec. 2021, Art. no. 103196.

[17] A. Echle, N. T. Rindtorff, T. J. Brinker, T. Luedde, A. T. Pearson, and J. N. Kather, "Deep learning in cancer pathology: A new generation of clinical biomarkers," *Brit. J. Cancer*, vol. 124, no. 4, pp. 686–696, Feb. 2021.

[18] B. Acs, M. Rantalainen, and J. Hartman, "Artificial intelligence as the next step towards precision pathology," *J. Internal Med.*, vol. 288, no. 1, pp. 62–81, Jul. 2020.

[19] F. McKay, B. J. Williams, G. Prestwich, D. Bansal, N. Hallowell, and D. Treanor, "The ethical challenges of artificial intelligence-driven digital pathology," *J. Pathol., Clin. Res.*, vol. 8, no. 3, pp. 209–216, May 2022.

[20] C. Huang, Z. Zhang, B. Mao, and X. Yao, "An overview of artificial intelligence ethics," *IEEE Trans. Artif. Intell.*, vol. 4, no. 4, pp. 799–819, Apr. 2022.

[21] C. Klein et al., "Artificial intelligence for solid tumour diagnosis in digital pathology," *Brit. J. Pharmacol.*, vol. 178, no. 21, pp. 4291–4315, Nov. 2021.

[22] S. R. Khan, D. Al Rijjal, A. Piro, and M. B. Wheeler, "Integration of AI and traditional medicine in drug discovery," *Drug Discovery Today*, vol. 26, no. 4, pp. 982–992, Apr. 2021.

[23] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.

[24] A. Waqas et al., "Revolutionizing digital pathology with the power of generative artificial intelligence and foundation models," *Lab. Invest.*, vol. 103, no. 11, Nov. 2023, Art. no. 100255.

[25] A. Waqas, J. Naveed, W. Shahnawaz, S. Asghar, M. M. Bui, and G. Rasool, "Digital pathology and multimodal learning on oncology data," *BJR|Artificial Intell.*, vol. 1, no. 1, Mar. 2024, Art. no. ubae014.

[26] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[27] R. Vinuesa et al., "The role of artificial intelligence in achieving the sustainable development goals," *Nature Commun.*, vol. 11, no. 1, pp. 1–10, Jan. 2020.

[28] M. Liu, D. Zhang, E. Adeli, and D. Shen, "Inherent structure-based multiview learning with multitemplate feature representation for Alzheimer's disease diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1473–1482, Jul. 2016.

[29] M. Liu, J. Zhang, P.-T. Yap, and D. Shen, "View-aligned hypergraph learning for Alzheimer's disease diagnosis with incomplete multi-modality data," *Med. Image Anal.*, vol. 36, pp. 123–134, Feb. 2017.

[30] Z. Cui et al., "TSegNet: An efficient and accurate tooth segmentation network on 3D dental model," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101949.

[31] E. Lloret Carbonell, Y. Shen, X. Yang, and J. Ke, "Covid-19 pneumonia classification with transformer from incomplete modalities," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2023, pp. 379–388.

[32] M. A. Wójcik, "Foundation models in healthcare: Opportunities, biases and regulatory prospects in Europe," in *Proc. Int. Conf. Electron. Government Inf. Syst. Perspective*. Cham, Switzerland: Springer, 2022, pp. 32–46.

[33] S. Zhang and D. Metaxas, "On the challenges and perspectives of foundation models for medical image analysis," *Med. Image Anal.*, vol. 91, Jan. 2024, Art. no. 102996.

[34] L. Yuan et al., "Florence: A new foundation model for computer vision," 2021, *arXiv:2111.11432*.

[35] W. Wang et al., "InternImage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14408–14419.

[36] L. A. Celi et al., "Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review," *PLOS Digit. Health*, vol. 1, no. 3, Mar. 2022, Art. no. e0000022.

[37] J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, and L. A. Celi, "The myth of generalisability in clinical research and machine learning in health care," *Lancet Digit. Health*, vol. 2, no. 9, pp. e489–e492, Sep. 2020.

[38] J. A. McDermid, Y. Jia, Z. Porter, and I. Habli, "Artificial intelligence explainability: The technical and ethical dimensions," *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 379, no. 2207, Oct. 2021, Art. no. 20200363.

[39] T. Sorell, N. Rajpoot, and C. Verrill, "Ethical issues in computational pathology," *J. Med. Ethics*, vol. 48, no. 4, pp. 278–284, Apr. 2022.

[40] A. Dahlen and V. Charu, "Analysis of sampling bias in large health care claims databases," *JAMA Netw. Open*, vol. 6, no. 1, Jan. 2023, Art. no. e2249804.

[41] C. Chauhan and R. R. Gullapalli, "Ethics of AI in pathology: Current paradigms and emerging issues," *The Amer. J. Pathol.*, vol. 191, no. 10, pp. 1673–1683, 2021.

[42] M. D. Moor. (2024). *Medgpt*. [Online]. Available: https://chatgpt.com/g/g-jxm5ljJmo-medgpt

[43] S. Dilmaghani, M. R. Brust, G. Danoy, N. Cassagnes, J. Pecero, and P. Bouvry, "Privacy and security of big data in AI systems: A research and standards perspective," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 5737–5743.

[44] F. Gilbert, "Balancing human and AI roles in clinical imaging," *Nature Med.*, vol. 29, no. 7, pp. 1609–1610, Jul. 2023.

[45] F. Aeffner et al., "The gold standard paradox in digital image analysis: Manual versus automated scoring as ground truth," *Arch. Pathol. Lab. Med.*, vol. 141, no. 9, pp. 1267–1275, Sep. 2017.

[46] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[47] A. Srivastava et al., "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," 2022, *arXiv:2206.04615*.

[48] D. Ganguli et al., "Predictability and surprise in large generative models," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Jun. 2022, pp. 1747–1764.

[49] P. Liang et al., "Holistic evaluation of language models," 2022, *arXiv:2211.09110*.

[50] R. Bommasani, K. Creel, A. Kumar, D. Jurafsky, and P. Liang, "Picking on the same person: Does algorithmic monoculture lead to outcome homogenization?" in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 3663–3678.

[51] M. A. Marchetti et al., "Prospective validation of dermoscopy-based open-source artificial intelligence for melanoma diagnosis (PROVE-AI study)," *Npj Digit. Med.*, vol. 6, no. 1, p. 127, Jul. 2023.

[52] H. Alkaissi and S. I. McFarlane, "Artificial hallucinations in ChatGPT: Implications in scientific writing," *Cureus*, vol. 15, no. 2, Feb. 2023, Art. no. e35179, doi: 10.7759/cureus.35179.

[53] S. A. Athaluri, S. V. Manthena, V. S. R. K. M. Kesapragada, V. Yarlagadda, T. Dave, and R. T. S. Duddumpudi, "Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references," *Cureus*, vol. 15, no. 4, Apr. 2023, Art. no. e37432, doi: 10.7759/cureus.e37432.

[54] P. Lee, S. Bubeck, and J. Petro, "Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine," *New England J. Med.*, vol. 388, no. 13, pp. 1233–1239, Mar. 2023.

[55] S. Subramanian et al., "Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 71242–71262.

[56] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jul. 2020.

[57] Z. Ma, J. Hong, M. O. Gul, M. Gandhi, I. Gao, and R. Krishna, "CREPE: Can vision-language foundation models reason compositionally?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10910–10921.

[58] I. Misra, A. Gupta, and M. Hebert, "From red wine to red tomato: Composition with context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1792–1801.

[59] J. Schneider, C. Meske, and P. Kuss, "Foundation models: A new paradigm for artificial intelligence," *Bus. Inf. Syst. Eng.*, vol. 66, pp. 221–231, Jan. 2024.

[60] H. Xu et al., "A whole-slide foundation model for digital pathology from real-world data," *Nature*, vol. 630, no. 8015, pp. 181–188, Jun. 2024.

[61] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomed. Eng.*, vol. 5, no. 6, pp. 555–570, Mar. 2021.

[62] R. J. Chen et al., "Towards a general-purpose foundation model for computational pathology," *Nature Med.*, vol. 30, no. 3, pp. 850–862, Mar. 2024.

[63] M. Y. Lu et al., "A visual-language foundation model for computational pathology," *Nature Med.*, vol. 30, no. 3, pp. 863–874, Mar. 2024.

[64] E. Vorontsov et al., "A foundation model for clinical-grade computational pathology and rare cancers detection," *Nature Med.*, vol. 30, no. 10, pp. 2924–2935, Oct. 2024.

[65] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[66] J. Ke et al., "ClusterSeg: A crowd cluster pinpointed nucleus segmentation framework with cross-modality datasets," *Med. Image Anal.*, vol. 85, Apr. 2023, Art. no. 102758.

[67] T. Santos et al., "PathologyBERT-pre-trained vs. a new transformer language model for pathology domain," in *Proc. AMIA Annu. Symp.*, 2022, p. 962.

[68] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[69] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," 2023, *arXiv:2304.07193*.

[70] X. Wang et al., "A pathology foundation model for cancer diagnosis and prognosis prediction," *Nature*, vol. 634, pp. 970–978, Oct. 2024.

[71] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2009.

[72] *Researchers Design Foundation AI Models for Use in Pathology*, Mass Gen. Brigham Commun., Harvard Medical School News, Boston, MA, USA, 2024.

[73] G. Gatta et al., "Burden and centralised treatment in Europe of rare tumours: Results of RARECAREnet—A population-based study," *Lancet Oncol.*, vol. 18, no. 8, pp. 1022–1039, Jul. 2017.

[74] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, "A visual–language foundation model for pathology image analysis using medical Twitter," *Nature Med.*, vol. 29, no. 9, pp. 2307–2316, Sep. 2023.

[75] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[76] S. Zhang et al., "BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," 2023, *arXiv:2303.00915*.

[77] A. Archit et al., "Segment anything for microscopy," *Nature Methods*, vol. 22, pp. 579–591, Jun. 2025.

[78] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[79] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.

[80] K. Saab et al., "Capabilities of Gemini models in medicine," 2024, *arXiv:2404.18416*.

[81] K. Singhal et al., "Towards expert-level medical question answering with large language models," 2023, *arXiv:2305.09617*.

[82] J. Cheng, "Neural network assisted pathology case identification," *J. Pathol. Informat.*, vol. 13, Jul. 2022, Art. no. 100008.

[83] N. Fei et al., "Towards artificial general intelligence via a multimodal foundation model," *Nature Commun.*, vol. 13, no. 1, p. 3094, Jun. 2022.

[84] D. Truhn, J.-N. Eckardt, D. Ferber, and J. N. Kather, "Large language models and multimodal foundation models for precision oncology," *Npj Precis. Oncol.*, vol. 8, no. 1, p. 72, Mar. 2024.

[85] A. Vaidya et al., "Demographic bias in misdiagnosis by computational pathology models," *Nature Med.*, vol. 30, no. 4, pp. 1174–1190, Apr. 2024.

[86] M. Bernhardt, C. Jones, and B. Glocker, "Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms," *Nature Med.*, vol. 28, no. 6, pp. 1157–1158, Jun. 2022.

[87] W. Zhong et al., "AGIEval: A human-centric benchmark for evaluating foundation models," 2023, *arXiv:2304.06364*.

[88] D. Wang et al., "A real-world dataset and benchmark for foundation model adaptation in medical image classification," *Sci. Data*, vol. 10, no. 1, p. 574, Sep. 2023.

[89] Y. Li et al., "Development and validation of the artificial intelligence (AI)-based diagnostic model for bronchial lumen identification," *Translational Lung Cancer Res.*, vol. 11, no. 11, pp. 2261–2274, Nov. 2022.

[90] C.-C.-M. Yeh et al., "Toward a foundation model for time series data," in *Proc. 32nd ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2023, pp. 4400–4404.

[91] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.

[92] C. Edlund et al., "Livecell—A large-scale dataset for label-free live cell segmentation," *Nature methods*, vol. 18, no. 9, pp. 1038–1045, 2021.

[93] M.-E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. S. Zemel, "Understanding the origins of bias in word embeddings," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2018, pp. 803–811.

[94] R. Azamfirei, S. R. Kudchadkar, and J. Fackler, "Large language models and the perils of their hallucinations," *Critical Care*, vol. 27, no. 1, p. 120, 2023. [Online]. Available: https://link.springer.com/article/10.1186/s13054-023-04393-x

[95] Z. Bai et al., "Hallucination of multimodal large language models: A survey," 2024, *arXiv:2404.18930*.

[96] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Comput. (TAMC)*. Cham, Switzerland: Springer, 2008, pp. 1–19.

[97] Y. Shen, A. Sowmya, Y. Luo, X. Liang, D. Shen, and J. Ke, "A federated learning system for histopathology image analysis with an orchestral stain-normalization GAN," *IEEE Trans. Med. Imag.*, vol. 42, no. 7, pp. 1969–1981, Jul. 2023.

[98] D. Castelvecchi, "Can we open the black box of AI?" *Nature*, vol. 538, no. 7623, pp. 20–23, Oct. 2016.

[99] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[100] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 1527–1535. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11491/11350

[101] N. Frosst and G. Hinton, "Distilling a neural network into a soft decision tree," 2017, *arXiv:1711.09784*.

[102] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[103] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[104] L. Van der Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, Jan. 2008.

[105] D. Nie, Y. Gao, L. Wang, and D. Shen, "ASDNet: Attention based semi-supervised deep networks for medical image segmentation," in *Proc. 21st Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Granada, Spain. Cham, Switzerland: Springer, 2018, pp. 370–378.

[106] E. Pahwa, D. Mehta, S. Kapadia, D. Jain, and A. Luthra, "MedSkip: Medical report generation using skip connections and integrated attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3402–3408.

[107] S. Sornapudi et al., "DeepCIN: Attention-based cervical histology image classification with sequential feature modeling for pathologist-level accuracy," *J. Pathol. Informat.*, vol. 11, no. 1, p. 40, Jan. 2020.

[108] (2016). *Regulation (eu) 2016/679 of the European Parliament and of the Council of 27, Apr. 2016 on the Protection of Natural Persons With Regard To the Processing of Personal Data and on the Free Movement of Such Data (general Data Protection Regulation)*. Accessed: Dec. 21, 2024. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[109] (1996). *Health Insurance Portability and Accountability Act of 1996*. Accessed: Dec. 21, 2024. [Online]. Available: https://www.hhs.gov/hipaa/index.html

[110] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[111] World Health Organization. (2021). *Ethics and Governance of Artificial Intelligence for Health: Who Guidance*. [Online]. Available: https://www.who.int/publications/i/item/9789240029200

[112] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 59–68.

[113] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: A state of the art," *Artif. Intell. Rev.*, vol. 56, no. 4, pp. 3005–3054, Apr. 2023.

[114] Z. Huang et al., "A pathologist—AI collaboration framework for enhancing diagnostic accuracies and efficiencies," *Nature Biomed. Eng.*, vol. 2024, pp. 1–16, Jun. 2024.

[115] Council of European Union. (2021). *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (artificial Intelligence Act)*. Accessed: Dec. 21, 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX