Research Statement

Longxi Zhou

I deeply believe that artificial intelligence (AI) can serve as a liberating force for human society. Realizing this potential demands not only technological breakthroughs, but also sustained interdisciplinary efforts to build AI that are trustworthy, accountable, and aligned with human values.

My research philosophy is grounded in the challenges of AI trust and autonomy—two intertwined areas that remain central bottlenecks to the real-world AI deployment. In my view, trust and autonomy are rooted in the same foundation: the capacity of an AI system to act accurately and reliably under uncertainty, while remaining accountable across diverse contexts.

From this perspective, trust and autonomy are not separate goals but co-emergent properties of a well-functioning AI system. *AI Trust* is shaped by performance, robustness, interpretability, and social acceptance. *AI Autonomy*, on the other hand, can be understood along two dimensions: *institutional autonomy*—the extent to which society entrusts or permits AI systems to influence the real world, ranging from recommendation support to authoritative decision-making; and *functional autonomy*—the system's internal capability to reason, adapt, and act without external guidance.

Throughout history, transformative technologies have reshaped society only when supported by robust institutional frameworks. This is particularly evident in healthcare: despite AI systems surpassing expert-level performance in several tasks, their real-world adoption remains limited by persistent challenges in trust and autonomy.

Consequently, my primary research interest lies in AI for healthcare, where I aim to build **autonomous and trustworthy clinical AI systems**. To this end, I focus on three interdependent directions:

1. **Translational Clinical AI** – transforming real-world clinical challenges into deployable AI solutions through close collaboration with hospitals and industry partners.
2. **Interpretable AI** – developing interpretability frameworks that foster *institutional autonomy* and enable safe deployment.
3. **Human-Inspired Autonomy** – designing architectures that learn from human cognitive patterns to support adaptive, autonomous decision-making under clinical uncertainty.

In the following sections, I highlight my major research initiatives and elaborate on my planned future research directions.

Research Direction I: Translational Clinical AI

Rather than optimizing benchmarks, this research direction focus on designing end-to-end systems—from algorithm to deployment—that are clinically aligned, vendor-robust, and validated in hospital settings.

Past Work: Scalable AI Systems for Real-World Clinical Deployment

I led the development of a suite of production-ready AI systems for multi-organ medical imaging, covering lung, heart, airway, and breast. These systems integrated segmentation (e.g., lung lobes, airways, vessels, tumors), classification (e.g., lung and breast lesions), and cross-modal registration across CT and MRI. I also built preprocessing pipelines for denoising, super-resolution, artifact removal, and motion correction—enabling robust performance across scanners and clinical environments.

Many of my models now **commercially deployed across over 100 hospitals**, through a long-term collaboration with Heilongjiang Tuomeng Technology Co. I served as principal architect across algorithm design, clinical integration, and validation, resulting in four patents and scalable commercial adoption. This deployment pipeline laid the foundation for multiple high-impact research efforts, including one of the earliest clinically deployed COVID-19 AI systems, published in *IEEE Transactions on Medical Imaging* (cited 216 times; acceptance rate 7%) [1].

In these projects, clinical and commercial value—not technical novelty alone—defined the research direction. For example, our COVID-19 tool evolved into a longitudinal study of fibrotic lesions in survivors, ultimately leading to a new commercial system for detecting subvisual patterns, later published in *Nature Machine Intelligence* [2]. More recently, I led the development of an AI system for diagnosing pulmonary embolism from non-contrast CT—a modality long believed to be unusable for this task. This AI system have undergone **1,004-case prospective study in emergency care**, currently under review at *Nature Cardiovascular Research* [3].

Future Research: Partnership-Driven Clinical Translation

My future work in translational clinical AI will follow a partnership-driven strategy—one that aligns the scientific goals of academia, the clinical needs of hospitals, the commercial value sought by industry, and the long-term development of students. This approach has already proven effective in my previous projects, where a single AI system yielded peer-reviewed publications, scalable commercial products, and measurable clinical impact.

Instead of committing to a narrow set of technical challenges, I will focus on building adaptable systems and reusable toolkits that can be rapidly customized to new medical scenarios—particularly where unmet clinical needs, technical feasibility, and funding incentives converge. In this way, my work serves as a connective tissue between research, practice, and innovation.

Research Direction II: Human-Centered AI Interpretability

This research direction aims to advance the *institutional autonomy* of AI systems by enabling human-AI collaboration through interpretable design. I treat "interpretability" not as a static attribute of models or methods, but as a structured interface that supports trust, accountability, and decision-making between humans and machines.

Past Work: AI Interpretability Beyond Post Hoc and Human Perceptual Limits

In my study published in *Nature Machine Intelligence* [2], I developed an interpretable AI system that enabled radiologists to detect previously invisible post-COVID fibrotic lesions. This work exemplified a **collaboration-centered interpretability approach**: rather than providing post hoc justifications after AI prediction, radiologists interact with our AI system (called DLPE) to suppressing irrelevant structures and optimizing visualization. DLPE improves the visibility of subtle structures by dozens of times, serving like a "CT microscope" to clinicians. It is now being used in clinical workflows, reinforcing the idea that interpretability is most effective when embedded into expert reasoning and visual practice—not as a post hoc justification. Project link.

As AI systems begin to make predictions that exceed human perceptual capabilities, a new paradigm of medical trust becomes necessary. In my recent work, I developed the first AI model—called SPEA—capable of diagnosing pulmonary embolism (PE) from non-contrast CT scans, a modality in which PE lesions are invisible to the human eye.

This breakthrough marks a shift in medical AI from a clinician-dependent decision support tool to an **independent diagnostic agent**, whose outputs cannot be directly verified by human observers. Thus, the traditional Clinical Decision Support System (CDSS) paradigm, which assumes post hoc human validation, becomes infeasible. As a result, ensuring **AI verifiability without human verification** becomes both a technical and ethical imperative.

To this end, I proposed the *Hide-and-Seek Self-Verification* (HSS) strategy, which embeds verifiability directly into AI model design, providing an **autonomous verifiability framework**. Unlike conventional post hoc explainability methods, HSS generates clinician-interpretable predictive indicators based on signal-noise spatial discrepancies rooted in domain knowledge, ensuring verifiability when medical AI exceeds human perceptual limits (see Figure 1 for its intuition).

To support the validity and potential generalizability of HSS, I developed a theoretical framework under reasonable assumptions, ensuring that HSS is statistically well-founded. Notably, SPEA and HSS were

prospectively validated in emergency care across 1,004 patients, offering a real-world, high-stakes demonstration of structured, quantifiable explainability. This project is currently under review at *Nature Cardiovascular Research* [3] (project link).

Future Research: Interpretability as a Trust Interface

1) Extending the HSS framework to other tasks. In our pulmonary embolism project, the HSS framework was well received by clinicians during prospective evaluations in emergency settings. Building on this success, I aim to further strengthen its theoretical foundations and generalize the framework to other high-stakes domains where verifiability and trust are critical—for example, oncology, intensive care, or multi-modal diagnostic decision-making

2) **Foundation models for trust-centered interpretability.** Different stakeholders—such as clinicians, lawyers, and computer scientists—hold fundamentally different views on what constitutes "AI interpretability." For example, doctors often find algorithmic explanations like Grad-CAM unintelligible; they instead prefer outputs that resemble familiar clinical reasoning patterns, such as biomarker-like indicators. Therefore, when computer scientists design interpretable AI systems, it is critical to first understand how interpretability is defined by the users themselves. In future work, I aim to develop a foundation model that systematically learns and distills domain-specific expectations of interpretability. This model would serve as a bridge between technical explanations and real-world user needs—ultimately helping computer scientists understand what counts as meaningful explanation across disciplines. The long-term goal is to synthesize these learned representations into unified, verifiable standards for AI safety and accountability.

Research Direction III: Human-Inspired Autonomy and Multimodal Learning

This research direction explores a new methodology towards human-level autonomy (autonomy refers to *functional autonomy* in this section) and its near-term validations in medical AI. I argue that human-level autonomy is an emergent property of complex systems, not something that can be reduced to finite rules or handcrafted templates, such as reinforcement learning reward functions or pre-scripted agent behaviors. The success of recent multimodal large models suggests that advanced cognitive capacities can emerge from self-supervised learning on large-scale data. However, current efforts to build autonomous AI have largely overlook this insight and based on reductionist approaches. They rarely leverage *explicit human data* associated with autonomy—such as data from eye-tracking, fMRI, EEG, or brain-computer interfaces.

Thus, this direction embraces a complexity-oriented paradigm, and seeking methods for emergent autonomy from complex systems.

Past Work: Learning Latent Regularities for CT Degradations

In recent work, I developed a self-supervised learning framework to model complex degradation patterns in CT scans. These degradations result from intricate physical processes such as photon decay and beam hardening, which traditional reductionist models attempt to explain individually. However, such models account for only 30%–70% of the observed degradation energy. By leveraging inter-slice continuity differences, we trained a foundation model—HorusEye—that restores X-ray images across seven modalities and five tasks, achieving diagnostic quality at ~4% radiation dose (under review at *Nature Computational Science* [4], project link).

This work illustrates the power of data-driven learning in modeling systems too complex for reductionist approaches. It also informs my future goal: developing AI agents with emergent human-level autonomy through multimodal human data such as eye-tracking, EEG and fMRI.

Future Research: Emergent Autonomy from Human-Cognitive Signals

Building on this foundation, I plan to pursue three lines of work.

1) Multimodal cognitive data collection. I aim to collect a multimodal dataset that captures manifestations of human autonomy. For instance, eye-tracking can reveal shifts in attention and task focus; fMRI or EEG can provide signatures of mind-wandering, conflict monitoring, or spontaneous reengagement; and self-narrated explanations—spoken or written—can contextualize internal intentions, hesitations, or shifts in motivation.

2) Near-term validations through clinician modeling. After the dataset of clinician eye-tracking and verbal explanations is collected, I plan to model clinicians' cognitive strategies in high-dimensional tasks such as whole-slide pathology, volumetric CT/MRI interpretation, or video-based diagnostics. These cognitive traces may help AI agents prioritize salient features, generalize to rare and complex conditions, and make decisions under uncertainty.

3) Self-supervised learning frameworks. I will develop self-supervised learning frameworks tailored to this rich multimodal dataset, enabling models to learn latent regularities in human cognition. Such frameworks are essential for training agents capable of internal modeling, adaptive focus, and autonomous behavior beyond reductionist approaches.

Unifying Vision: Towards Clinical Autonomous Agents

By integrating the above three research directions, I envision a revolutionary clinical AI that are not only accurate and robust, but also both institutionally and functionally autonomous. This unified vision is structured across three progressive stages:

Short-Term Vision (5–10 years): Multimodal AI and Clinical Trials

Enhance the multimodal perception and interaction capabilities of clinical AI systems, enabling them to interpret rich non-verbal cues such as facial expressions, gestures, tone, and emotional states—information that is vital in clinical decision-making.

Further, AI should learn from physicians' reasoning processes and value systems, thereby improving functional autonomy in ways that are aligned with human judgment and ethics.

This stage also involves increasing human-AI trust and conducting pilot deployments and prospective clinical trials to evaluate the safety and efficacy of AI systems in real-world settings.

Mid-Term Vision (10-20 years): Guideline Integration and Personalized Care

Promote the formal integration of AI into clinical diagnostic guidelines, where AI may be authorized to generate definitive diagnostic reports under specific conditions.

Advance the AI's decision-making capacity by incorporating a wide range of multimodal data—patient history, genomics, socioeconomic factors, social support, and individual values—to deliver cost-effective, personalized treatment plans.

Simultaneously, foster AI's sense of humanistic care and social responsibility, paving the way for the emergence of general-purpose, multimodal clinical AI agents.

Long-Term Vision (20+ years): AI Revolution in Medicine

Establish comprehensive technical standards for clinical AI agents and facilitate their widespread, regulated deployment.

This includes the development of ethical, legal, and institutional frameworks governing the use of autonomous AI in medicine—for example, whether AI-generated prescriptions and diagnoses are eligible for insurance reimbursement, how responsibility is assigned in cases of misdiagnosis, and how AI behavior is to be audited and regulated.

Ultimately, this vision aims to enable the medical profession to smoothly embrace the transformation of both productivity and social structure led by the AI revolution.

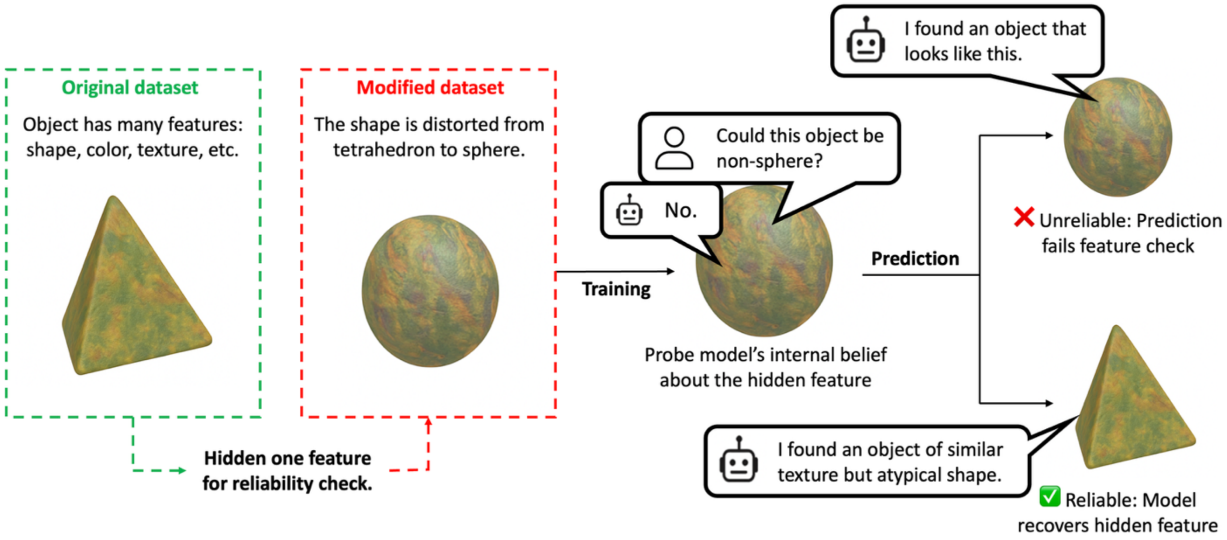
Original dataset
Object has many features: shape, color, texture, etc.
Modified dataset
The shape is distorted from tetrahedron to sphere.
Hidden one feature for reliability check.
Training
Could this object be non-sphere?
No.
Probe model's internal belief about the hidden feature
Prediction
I found an object that looks like this.
✗ Unreliable: Prediction fails feature check
I found an object of similar texture but atypical shape.
✓ Reliable: Model recovers hidden feature

Figure 1. Intuition behind the Hide-and-Seek Self-Verification (HSS) strategy. The true object is a textured, colored tetrahedron. During training, one key feature—shape—is hidden by distorting it into a sphere, forcing the model to learn under partial information. At inference, we test whether the model can recover the hidden feature from remaining cues (e.g., texture). Successful recovery suggests reliability; failure suggests hallucination. This mechanism defines predictive indicators and a reliability score, enabling trust assessment even without human verification.

Reference:

[1] **Longxi Zhou** *et al*., "A Rapid, Accurate and Machine-Agnostic Segmentation and Quantification Method for CT-Based COVID-19 Diagnosis". ***IEEE Transactions on Medical Imaging***, 2020.

[2] **Longxi Zhou** *et al.,* "An interpretable deep learning workflow for discovering subvisual abnormalities in CT scans of COVID-19 inpatients and survivors". ***Nature Machine Intelligence***, 2022.

[3] **Longxi Zhou** *et al.,* "AI trust beyond perceptual limits: pulmonary embolism diagnosis on non-contrast CT". ***Nature Cardiovascular Research*** (Under Review, GitHub Link).

[4] Yuetan Chu[#], **Longxi Zhou**[#] *et al.,* "HorusEye: A self-supervised foundation model for generalizable X-ray tomography restoration". ***Nature Computational Science*** (Under Review, GitHub Link). [#]Co-first authorship.