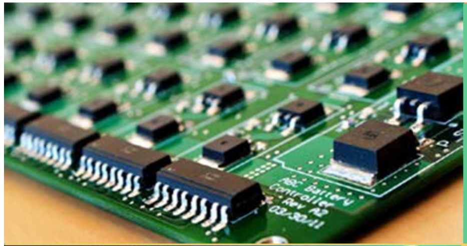
Safe Learning-Enabled Systems
ICCPS Panel

NSF Team

- Jie Yang, CISE/IIS, jyang@nsf.gov
- Anindya Banerjee, CISE/CCF, abanerje@nsf.gov
- David Corman, CISE/CNS, dcorman@nsf.gov
- Pavithra Prabhakar, CISE/CCF, pprabhak@nsf.gov

Participating CISE Divisions



Computing and Communication Foundations advances computing and communication theory, algorithms for computer and computational sciences and architecture and design of computers and software.

Computer and Network Systems invent new computing and networking technologies and finds new ways to make use of current technologies.

Information and Intelligent Systems studies the interrelated roles of people, computers, and information to increase our ability to understand data, as well as to mimic the hallmarks of intelligence in computational systems.

Partners

- Open Philanthropy Project LLC
- Good Ventures Foundation

AI Systems are Growing Rapidly

- As artificial intelligence (AI) systems rapidly increase in size, acquire new capabilities, and are deployed in high-stakes settings, their safety becomes extremely important.
- Ensuring system safety requires more than improving accuracy, efficiency, and scalability: it requires ensuring that systems are robust to extreme events and monitoring them for anomalous and unsafe behavior.

Undesirable Behaviors

- Developers must ensure that when deployed, undesirable system behaviors do not arise. Undesirable system behaviors encompass:
  - overt blunders like prediction errors and system crashes,
  - silent failures, like reporting unjustified confidence levels out-of-distribution, and competently achieving unintended objectives.

Learning-Enabled Systems

- Learning-enabled systems are systems with learning-based components that include, but are not limited to, deployed systems in healthcare and medicine, criminal justice, autonomous and cyber-physical systems, and finance.
- Learning-enabled systems also include foundational learning-based systems that may be subsequently applied in many downstream domains.

Goals of the Program

- Solicit foundational research that leads to the design and implementation of learning-enabled systems in which safety is ensured with high levels of confidence.
- The program will be considered a success if developers of future learning-enabled systems can (i) informally explain why the systems can be deployed safely in unpredictable environments and (ii) back these informal explanations with rigorous evidence that the system satisfies precise safety specifications.

Goals of the program (Con’t)

- The program solicits proposals that advance general theories, principles, and methodologies for the design of safe learning-enabled systems, that go beyond specific problem instances, and that are applicable to state-of-the-art learning systems, including considerations for scalability and deployability.

Ideals for the Program

- Proposals that have the potential to make strong advances in the design and implementation of safe learning-enabled systems as well as advancing methods for reasoning about the safety of those systems when they are deployed in unpredictable environments.
- An ideal proposal will demonstrate how these two objectives will be achieved, provide evidence that its proposed approach will improve notions of safety, and argue the potential for lasting impact both on rigorous safety evaluation methods and on the design and implementation of safe learning-enabled systems.

Safety Guarantees

- Verifying that learning systems achieve safety guarantees for all possible inputs may be difficult.
- Considerations for establishing safety guarantees :
  - systematic generation of data from realistic (yet appropriately pessimistic) operating environments.
  - resilience to “unknown unknowns”, which necessitates improved methods for monitoring hazards or behaviors.
  - new methods for reverse-engineering, inspecting, and interpreting the internal logic of learned models.
  - methods for improving the performance by directly adapting the systems’ internal logic.

Safety Requirements

- Any system claiming to satisfy a safety specification must provide rigorous evidence through analysis corroborated empirically and/or with mathematical proof.
- Proposals that increase safety primarily as a downstream effect of improving standard system performance metrics unrelated to safety (e.g., accuracy on standard tasks) are not in scope.

All Proposals Must

- State the notion of end-to-end mathematically- (i.e., precisely and without ambiguity, and accounting for error bounds) or empirically-based safety, in plain English.
- Justify why the end-to-end safety properties are critical to the system.
- Identify environmental assumptions for the safety properties.
- Provide automated/semi-automated/interactive techniques for establishing the degree to which the safety properties are satisfied.
- Demonstrate that these techniques achieve safety, mathematically or empirically, through rigorous simulation, prototyping, and integration with actual learning-enabled systems.

Notions of safety include, but are not limited to

- Robustness and resilience to tail-risks,
- Monitoring systems for anomalous and unsafe behavior,
- Interpreting, reverse-engineering, and inspecting a learned system’s internal logic,
- Reliability under human error.

Project Classes and Deadlines

- Foundation Projects
  - up to $800,000 total budget with durations up to three years.
- Synergy Projects
  - $800,001 to $1,500,000 total budget with durations up to four years

Full Proposal Deadline(s) (due by 5 p.m. submitter's local time):

- May 26, 2023
- January 16, 2024

NSF Review Criteria

- ❑ Intellectual Merit/Broader Impacts:

1. What is the potential for the proposed activity to
    1. Advance knowledge and understanding within its own field or across different fields (Intellectual Merit); and
    2. Benefit society or advance desired societal outcomes (Broader Impacts)?
2. To what extent do the proposed activities suggest and explore creative, original, or potentially transformative concepts?
3. Is the plan for carrying out the proposed activities well-reasoned, well-organized, and based on a sound rationale? Does the plan incorporate a mechanism to assess success?
4. How well qualified is the individual, team, or organization?
5. Are there adequate resources available to the PI for the proposed activities?

Additional Solicitation Specific Review Criteria

The proposals will also be evaluated based on:

- **Components:** The discussion of the learning-enabled components. The proposal should describe the components and provide reasons why they are appropriate for the system being studied.
- **Rationale:** A rationale in plain language why the end-to-end safety properties are critical to the learning-enabled system.
- **Safety Plan:** Description of the environmental assumptions under which the safety properties are ensured, as well as inclusion of an automated/semi-automated/interactive techniques for establishing the degree to which the safety properties are present in the learning-enabled system.
- **Validation:** A plan to validate these techniques to demonstrate that they can achieve the mathematically or empirically-specified safety guarantees through rigorous (as opposed to ad-hoc) simulation, prototyping, and integration with actual (including sub-scale) learning-enabled systems.
- **For Synergy projects, the validation plan must include experimentation on an actual learning-enabled system**