# R4 Trustworthy AI

The world is facing unprecedented catastrophic risks, arising from the deadly pandemics and epidemics and intersection of exponential technologies. AI and robotics as two representative technologies of the 4th Industrial Revolution continue to advance rapidly to become increasingly exploitable across domains in multiple ways. While AI and robotics can provide solutions to a wide range of capability gaps and challenges, but the digitization of the world is not intended to replace human involvement completely. Given the limitations of these technologies, the trend raises important questions about the benefits, complications, liabilities, risks, and trust associated with increasing autonomy in safety-critical socio-technical systems.

The use of AI and autonomy involves complex legal, ethical, moral, social, and cultural issues that may impede their development, evaluation, and application by their human partners as a collaborative human-AI symbiosis. However, there currently exists no government regulations in this regard, no coordinated approach, no organized community response, and no international research program seeking for answers to the challenge of understanding and mitigating the risks associated with operating these AI-enabled autonomous systems. Further, there is a lack of guidance to support the design of these safety-critical socio-technical systems while keeping potential benefits, as well as limitations and potential harm, in mind. It is imperative that the appropriate and validated processes to ensure system reliability, robustness, resilience, responsibility (R4), and trustworthiness before they are integrated more widely into our organizations, systems, operations, and society.

This lecture aims to address the needs for system designers, developers, project manager, researchers, and all practitioners who are interested in building and using 21st century human-AI symbiosis technologies. The lecture will discuss technical challenges AI is facing, risks associated with the interactions between human and AI partners, and a potential solution from systems design perspective by introducing an intelligent adaptive system (IAS) framework and associated methodology to address these challenges and risks. IASs are human-machine symbiosis technologies that exhibit collective intelligence enabled by optimized human-machine interactions based on their joint capabilities, strengths, and limitations to achieve shared goals.

The lecture will provide guidance for understanding the requirements of R4 Trustworthy AI and mitigating the potential risks associated with employing AI-enabled socio-technical systems. The evolutional nature of systems design strategy and methodology in the past 7 decades since the 1950' for human-machine systems (HMS) technologies will be reviewed. The philosophies and principles of technology-centred design (TCD), human-centred -design (HCD), and interaction-centred design (ICD) paradigms will be discussed in detail. Analytical methodologies for functional requirements of the IASs, design methodologies, implementation strategies, and evaluation approaches will be elaborated with real-world examples of designing and developing AI-enabled autonomous systems when considering context constraints of technology, human capability and limitations, and functionalities that the system should achieve.