A Perspective on Trust in Machine Learning and Control for Dynamic Autonomous Systems

AFRL-SUNY Trusted AI Challenge Series

Pramod P. Khargonekar

Department of Electrical Engineering and Computer Science
University of California, Irvine

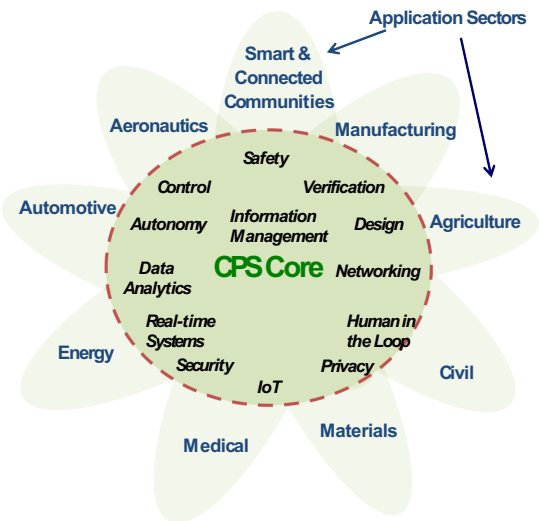
14 October 2020

Outline

1. ML/AI and Cyber-Physical Systems
2. Control Systems Perspective
3. Recap of Recent Machine Learning Breakthroughs
4. What does the Future Hold?

Why ML/AI in Cyber-Physical and Autonomous Systems?

Cyber-Physical Systems



Application Domains

Transportation

- Faster and safer vehicles (airplanes, cars, etc)
- Improved use of airspace and roadways
- Energy efficiency
- Manned and un-manned

Energy and Industrial Automation

- Homes and offices that are more energy efficient and cheaper to operate
- Distributed micro-generation for the grid

Healthcare and Biomedical

- Increased use of effective in-home care
- More capable devices for diagnosis
- New internal and external prosthetics

Critical Infrastructure

- More reliable power grid
- Highways that allow denser traffic with increased safety

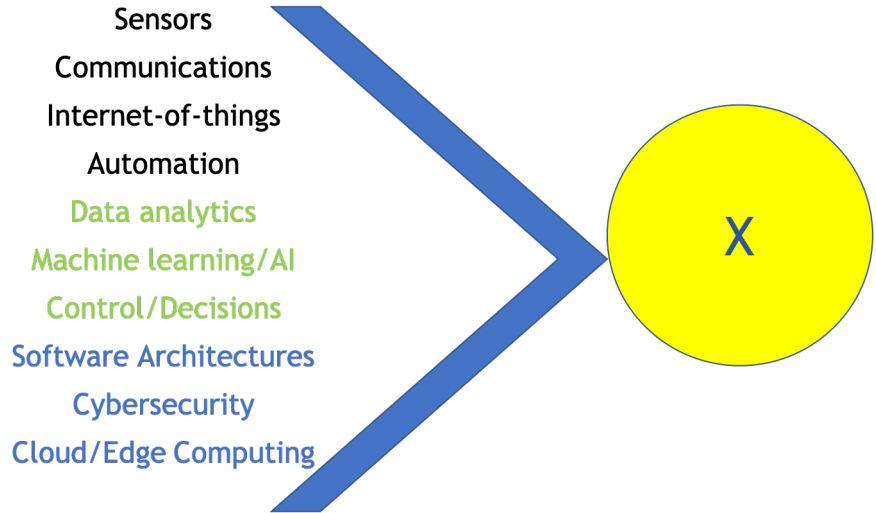
Source: NSF

CPS Properties

- Pervasive computation, sensing, and control integrated into physical systems
- Dynamically reorganizing/reconfiguring
- Networked at multiple scales
- High degrees of automation
- Dependable operation with potential requirements for high assurance of reliability, safety, security and usability
- With or without human interaction/supervision

Source: NSF

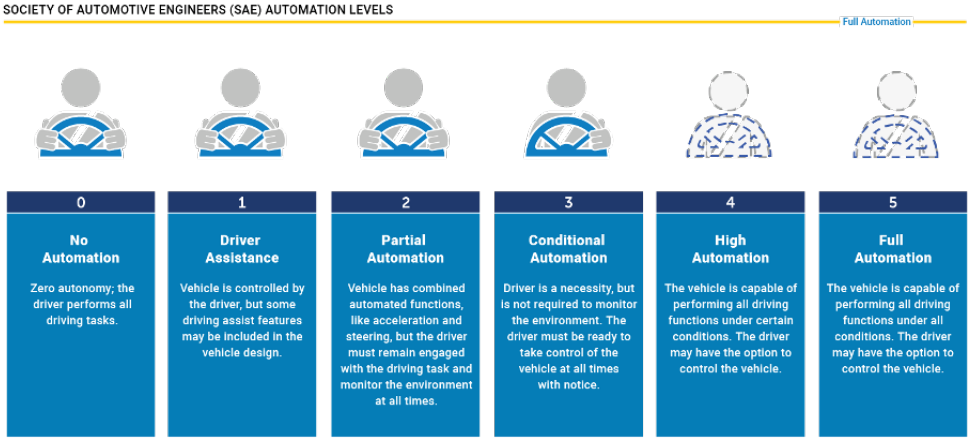
Smart-X: ML/AI into CPS



Smart and Autonomous CPS: Examples

- Smart-X
  1. Smart manufacturing
  2. Smart electric grid
  3. Smart transportation
  4. Smart cities
  5. Smart health
- Autonomous systems
  1. Unmanned air vehicles
  2. Self-driving cars
  3. Autonomous robots
- **ML/AI will be gradually and increasingly integrated into CPS to create smart-X with increasing levels of autonomy.**

Example: Level 5 Autonomy in Self-Driving Cars



Smart Autonomous Systems

How can we Ensure Trust in such Systems?

Perspective from Control Systems

Control Systems : Strong Theoretical Foundations

- Stability theory
- Optimal control
- Linear multivariable control
- Robust control
- Nonlinear control
- Adaptive Control
- Stochastic control
- Distributed control

Adaptive Control

- Control systems that can learn about and adapt to changes in the system/environment
- A long-standing goal in control theory
- Deterministic and stochastic models
- Parameter learning, neural network learning, ...
- Mathematical results on closed loop stability, convergence of parameters, ...
- Transient performance remains a hard problem
- Recent results on flight tests of adaptive control

Formal Methods

- Specify desired behavior from the controlled system in terms of logical statements
- Temporal logics: linear temporal logic (LTL), metric temporal logic (MTL), signal temporal logic (STL), ...
- Model checking, theorem proving, etc. for verification
- Alternative lightweight formal methods
- Investigations of robustness
- Concerns about specifying the desired behaviors
- Concerns about scaling to complicated systems
- Concerns about overall time and effort needed

Key Ideas

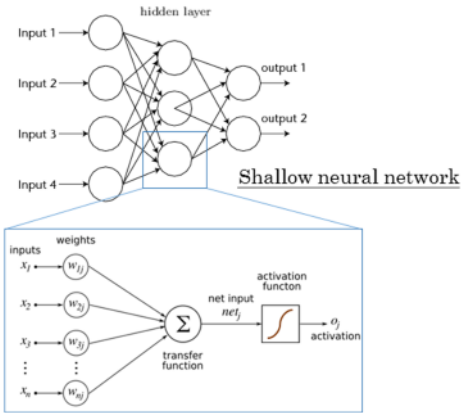
- Heavy use of mathematical models and techniques
  1. Differential and difference equations: ODEs, PDEs, ...
  2. Discrete-event models
  3. Hybrid models
  4. Deterministic and stochastic
  5. Distributed, networked, hierarchical, ...
  6. Toolsets for analysis and design: time-domain, frequency domain, simulations, optimization, numerical computations, ...
- Explicit accounting of modeling errors and robustness guarantees
- Mathematically provable properties
- Bottomline: so long as mathematical assumptions hold, conclusions are guaranteed.

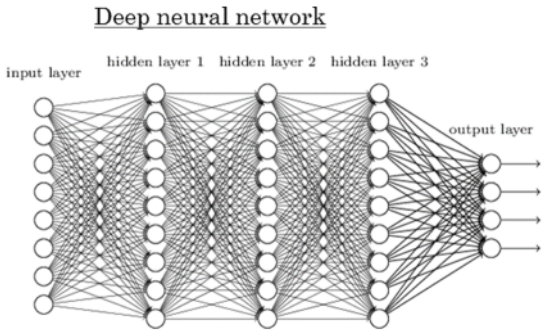
Recap of Recent Machine Learning Breakthroughs

Computational Intelligence: Pattern Recognition or Model Building

- Two fundamentally different perspectives on learning from data:
  1. Statistical pattern recognition from data for prediction and control
  2. Using data to build causal models to understand, predict and control
- Possible to combine these two approaches
- Causality is a critical issue

Deep vs Shallow Neural Networks





Source: github

Key Advantage of Deep Networks

"... shallow classifiers require a good feature extractor ... one that produces representations that are selective to the aspects of the image that are important for discrimination ... The conventional option is to hand design good feature extractors, which requires a considerable amount of engineering skill and domain expertise. But this can all be avoided if **good features can be learned automatically ... This is the key advantage of deep learning.**"

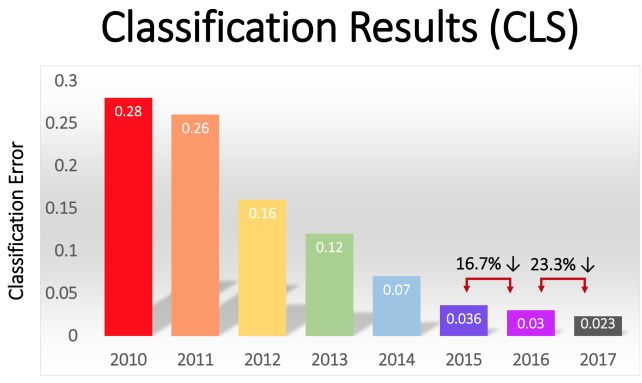
Deep Learning, LeCun, Bengio, and Hinton, Nature, 2015.

Major DL Innovations

- Convolutional neural networks
- Long Short Term Memory (LSTM) for sequential data
- Numerous algorithmic and architectural innovations
- Training and optimization of extremely large networks
- Use of graphics processors for computation
- Leveraging large volumes of training data

Breakthrough in Vision: ImageNet Competition

ImageNet Classification with Deep Convolutional Neural Networks, Krizhevsky, Sutskever, and Hinton, 2012



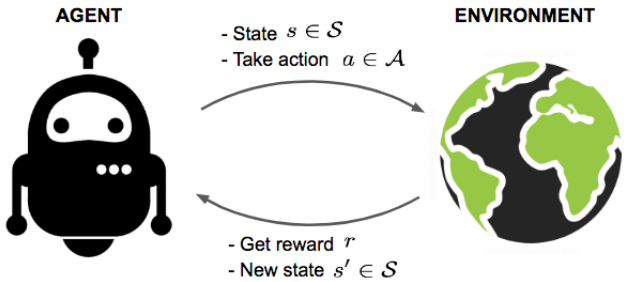
Source: image-net.org

Are we done with ImageNet?, Beyer et al. 2020

Recurrent Neural Networks

- Recurrent neural networks (RNNs): neural network models with the ability to pass information across time steps
- Suitable for modeling data that are
  - Sequential and dependent
  - Of varying input lengths
- RNNs: natural choice for time series and other sequential applications
- Long Short Term Memory (LSTM) Networks: the state-of-the-art RNNs
- Speech processing, machine translation, ...

RL Framework



The "agent" is the controller and the "environment" includes the plant, uncertainty, disturbances, noise, etc.

Source: github

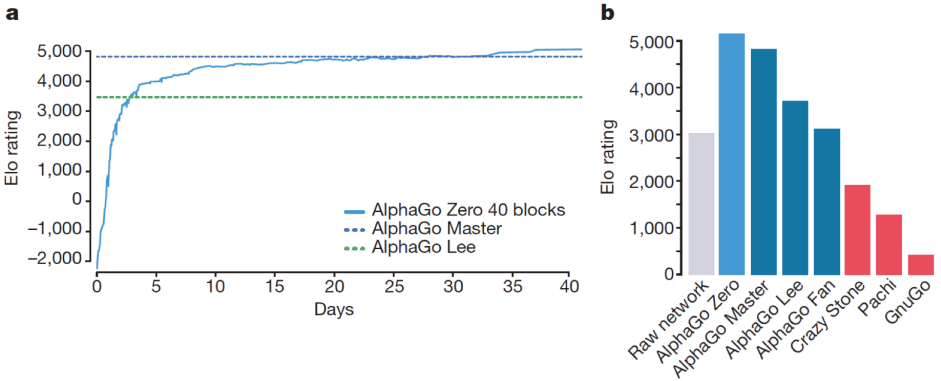
Reinforcement Learning: General Setup

- At each time step, agent observes the state, takes action, and receives a reward
- Goal for the agent: choose actions to maximize total discounted reward
- Optimal action policy is a form of control law
- Can the agent learn the optimal policy by suitable use of state and reward data?
- RL: A general machine learning paradigm to solve problems and attain goals

Key Ideas and Building Blocks

- Bellman's optimality principle: *Tail of an optimal policy must be optimal.*
- Function $Q(a, a)$: optimal policy given by maximizing with respect to a.
- One approach: Learn the Q-function.
- Recent innovations in modern RL
  1. Deep Reinforcement Learning: Use deep neural networks to approximate Q (DQN)
  2. Experience replay to reuse past data
  3. Asynchronous and parallel RL
  4. Rollouts based planning for RL
  5. Self-play for faster learning
  6. Techniques for data efficiency
  7. Techniques for continuous action spaces

AlphaGo Zero Achieves State-of-the-Art Performance



Despite learning by itself from zero prior knowledge,
it learns and outperforms all other algorithms.

Critical Recap of Recent ML Breakthroughs

- DL establishing itself as a major new technology
- Insufficient theory of DL but progress on both approximation and generalization
- Major investments in DL hardware that will make it cheaper to implement
- Deep reinforcement learning — breakthrough performance in board games
- Applications of DRL to physical systems at very early stages
- DL and RL depend on large amounts of data
- DL and (much of) RL are model-free
- Novel research directions: unsupervised learning, federated learning, episodic learning, meta-learning, multi-agent RL, ...
- Enormous global interest in private, academic, government sectors

Several modern ML innovations integrated into smart autonomous systems

What Does the Future Hold?

CPS Autonomy and ML/AI

- Although there has been very impressive progress, we are at early stages in creation of smart autonomous systems
- Further progress in computational cognition, machine learning, and artificial intelligence will be needed for achieving ambitious autonomous systems goals
- There is a great need for conceptual, architectural, and algorithmic advances
- These systems will be designed to learn, adapt, and evolve
- System goal and design specifications will be critical
- Numerous questions about trust will arise in such systems
- Example: Level 5 autonomy in self driving cars

Some Questions

- How should we specify desired properties and behaviors knowing that the smart autonomous system will change as it learns and adapts?
- If we are able to specify desired behaviors, how will we verify that the smart autonomous system will fulfill these specifications?
- Will such systems able to deal with changes in their contexts?
- Will such systems be more vulnerable to malicious and adversarial attacks?
- Will they be able to differentiate between malicious attacks and changes in the environment?
- How will we ensure that smart autonomous systems are subordinate to human control and supervision?
- Could humans lose control over smart autonomous systems?

Potential Directions for the Future

- Increased transparency and comprehensibility in the learning components: explainable AI combined with cyber-physical systems
- Robustness against adversarial attacks
- Greater emphasis on building internal models
- Cognitive CPS: build cognitive properties into future CPS: perception, memory, attention, knowledge representation, problem solving, ...
- Computational self-awareness in smart autonomous systems
- Make explicit the idea of intent in smart autonomous systems

Conclusions

- We are at early stages of building smart autonomous systems
- New challenges and opportunities for ensuring trust
- Many new directions for research and translation
- Ethical issues and societal impacts should be seriously considered

Thank you!

Email: pramod.khargonekar@uci.edu
Website: https://faculty.sites.uci.edu/khargonekar/