

Adversarial Resilience Learning

Analysis and Resilient Operation of Complex Systems
without Domain Knowledge

Dr.-Ing. Eric MSP Veith <eric.veith@offis.de>



An aerial photograph of a complex, multi-level highway interchange. The roads curve and overlap in various directions, creating a dense network of concrete and asphalt. The scene is set against a dramatic sunset sky with vibrant orange and red hues near the horizon, transitioning to a deep blue at the top. In the background, a city skyline is visible across a body of water, with a prominent cable-stayed bridge on the left. The overall atmosphere is one of modern infrastructure and urban development.

Our Infrastructures become more complex with every day.

A technician wearing a white hard hat and a blue vest is working on a dense array of network cables in a server room. The cables are bundled and organized, with some yellow and blue cables visible. The technician is focused on the task, looking down at the equipment. The background shows server racks and more cables, creating a complex and technical environment.

*We add to this complexity every day:
Through more communication networks and learning systems
— necessarily so.*

*“Machine Learning: The High Interest Credit Card of Technical Debt”
— Sculley, et al. (Google), 2014*

The Power Grid

A Pervasive Example

Power Grids: A Critical Infrastructure

- > Basis for our society!

Spanning Continents

- > From Scandinavia to North Africa, from Ireland to Asia

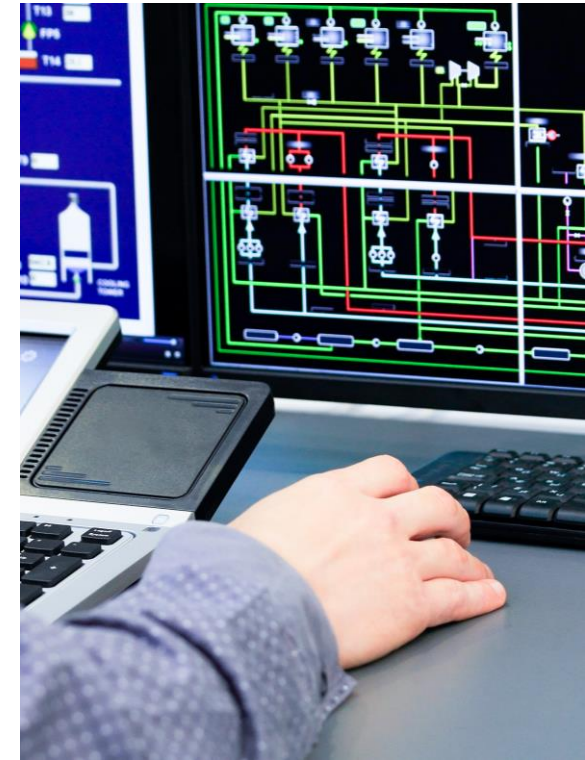
Effects ubiquitously visible

- > Instabilities spread almost instantaneously

Conflicting goals

- > Monetary, technical, and political interests not congruent

Caught in a fast and turbulent change



Major Trends

Influencing the Cyber-Physical System Power Grid

1. Evolution of the Power Grid,

- > Many small generators, critical in sum
- > Competition & business model innovations
- > Growing complexity only manageable through more digitalization – vicious circle!

2. Digitalization

- > IoT trends: Many thousand “intelligent” devices (Nest, baby monitors, Smart TV, etc.) connect to the grid
“In IoT, the ‘S’ stands for ‘Security.’”
- > Buzzword Bingo: Smart Services, Cloud, Outsourcing, AI, Big Data, ...

3. New Threats through Cyber Attacks

- > State-sponsored attacks (*Grey Wars*)
- > Ever more sophisticated tools
- > Strong pressure for backdoors

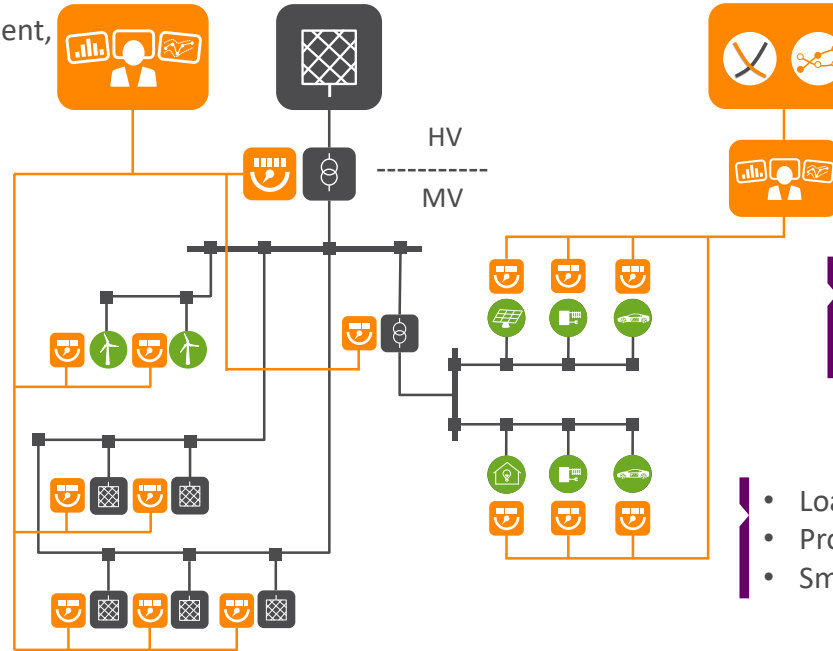


AI in the Power Grid

A Complex System Indeed

- Grid State Prognosis,
- Optimized Reactive Power Management,
- Anomaly Detection in Power Grid and ICT.

- Asset Monitoring,
- Automated Grid Interaction Stage,
- Decentralized Ancillary Services.



- Algorithmic Power Trading,
- Market Integration of Renewables.

- Virtual Power Plants,
- Multi-Modal Optimization,
- Sector Coupling.

- Load & Flexibility Management,
- Prosumer,
- Smart Metering.

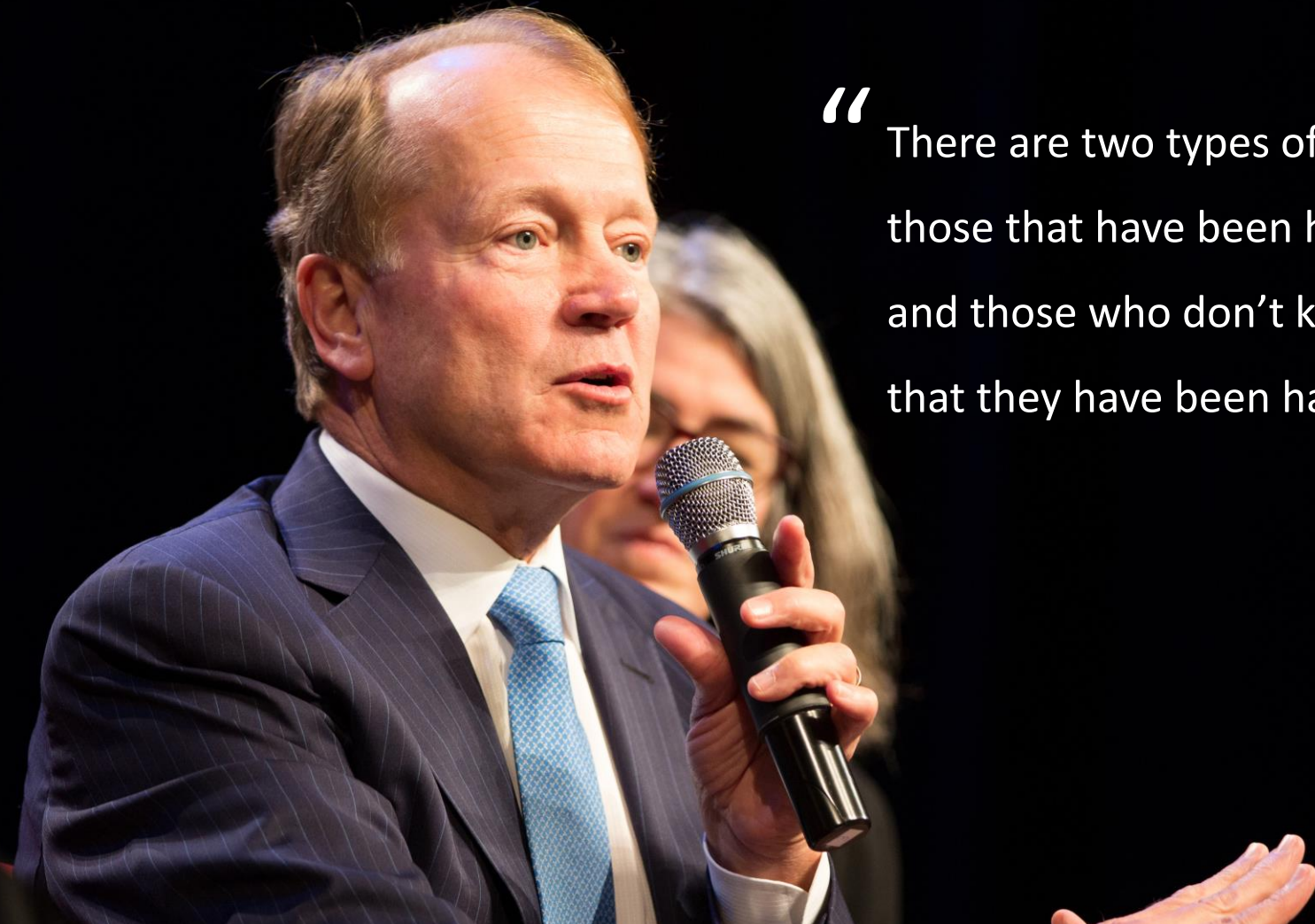
What follows?

Obvious Conclusions Drawn from the Current State of Affairs



1. Trends in digitalization from other areas will flood into our critical infrastructures.
2. Digitalization of our power supply, transportation, etc., will lead to a new threat level – damage potential unfathomable.
3. Digitalization & machine learning are necessary for sustainable, environment-friendly infrastructures: Not only a threat, but also a great potential, even for security!

We don't really understand the interdependent effects between digitalization and critical infrastructures yet.

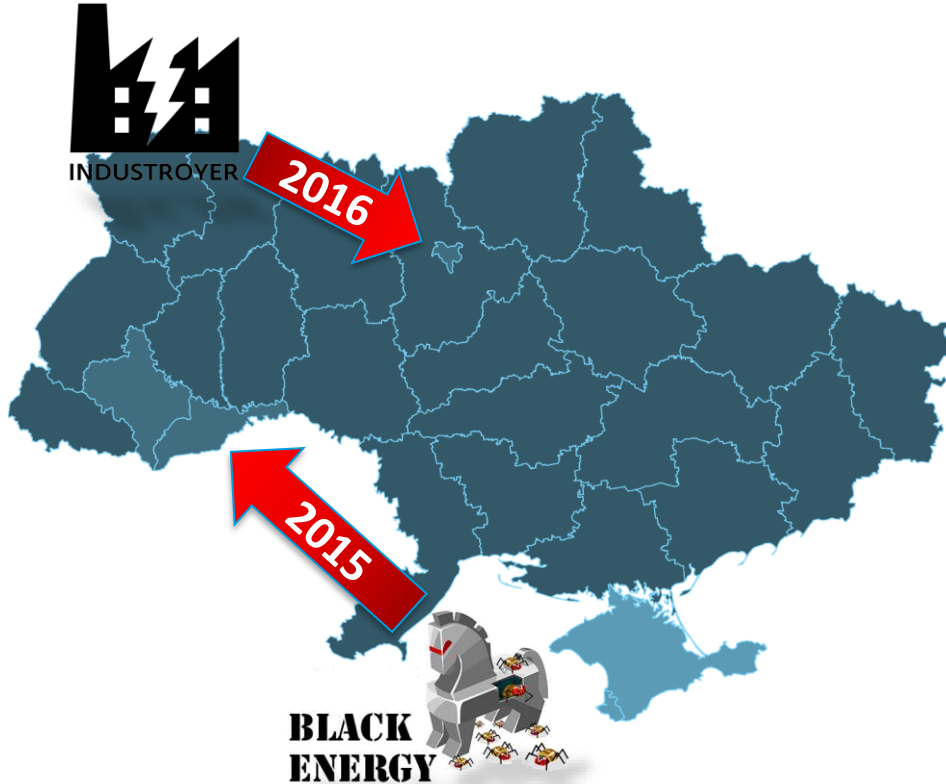


“ There are two types of companies:
those that have been hacked
and those who don't know
that they have been hacked.”

John T. Chambers.

Applies to Critical Infrastructures, too

Attack against the Ukrainian Power Grid



Dec 23rd, 2015

- > Cyber Attack leads to **Blackout**
- > **3 Grid Operators** targeted
- > Operative **Intrusion into Control Systems**
- > Disconnect of **several Transformers**
- > Several Months in Preparation


2016

- > **Highly automated** Variant

Our infrastructures are valuable targets.

Digitalized Critical Infrastructures: A Threat?

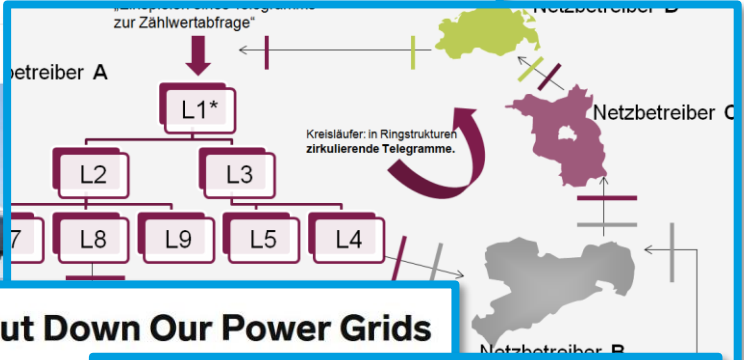

Newspaper Collection



TLP: White
Analysis of the Cyber Attack on the Ukrainian Power Defense Use Case
March 18, 2016

Adversarial Military Information Operations and the New NDAA: The Law of the Gray Zone Evolves

by [Name] Tuesday, December 10, 2019, 5:03 PM



NSA Director: Yes, China Can Shut Down Our Power Grids

AP KEN DILANIAN, Associated Press Nov 20, 2014, 6:28 PM

China and "one or two" other

ANDY GREENBERG SECURITY 07.06.2017 11:36 PM

Hack Brief: Hackers Targeted a US Nuclear Plant (But Don't Panic Yet)

Hackers have reportedly targeted US energy utilities, and may be laying the groundwork for blackouts. But they may yet be a long way from that goal.

Technology | Cybersecurity

Russian hackers hit Ireland's power grid in another cyberattack on UK's critical infrastructure

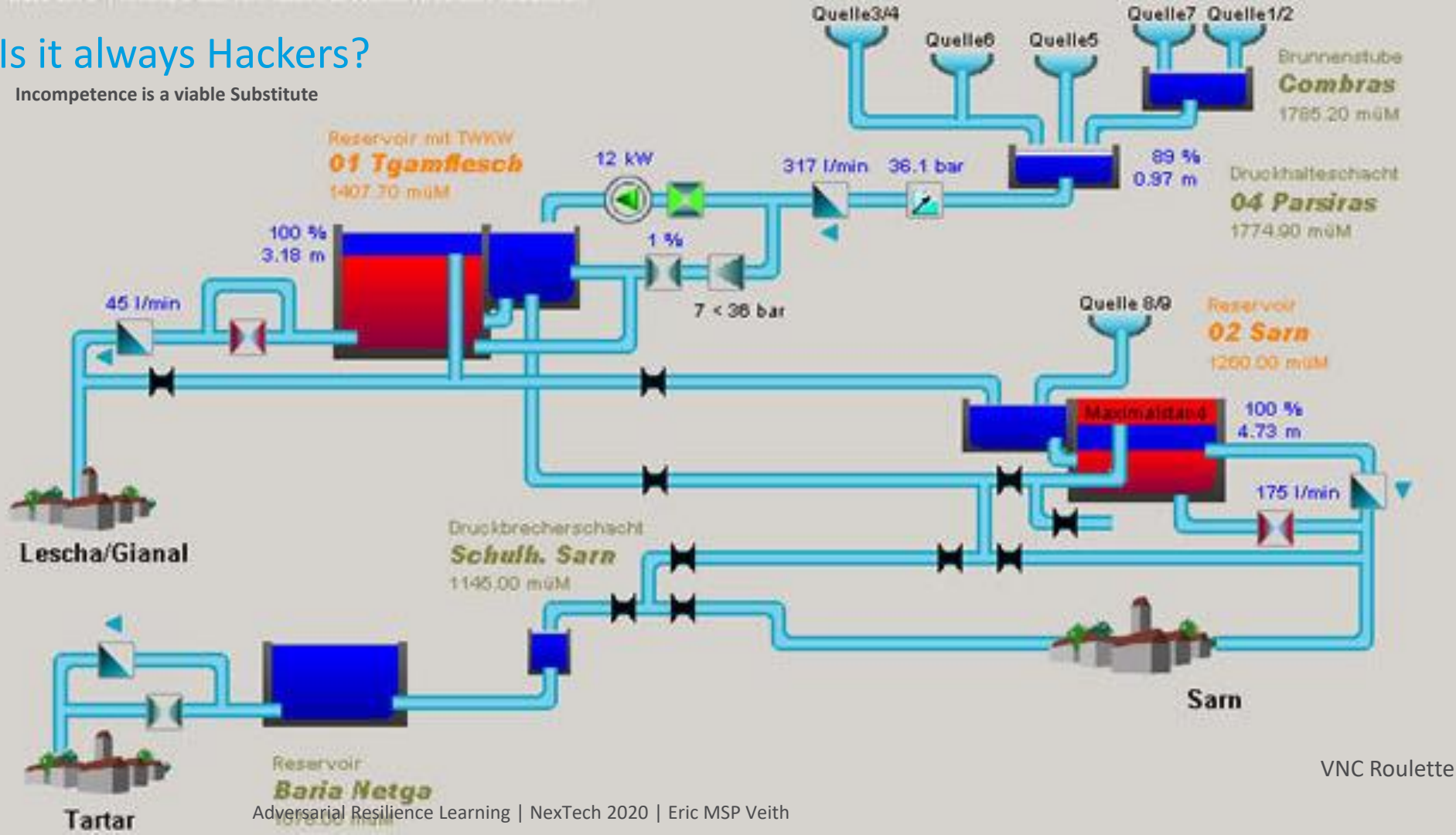
Although there is no evidence to suggest that Irish energy networks were disrupted, security experts reportedly believe hackers have stolen data.

By India Ashok July 17, 2017 09:56 BST



Is it always Hackers?

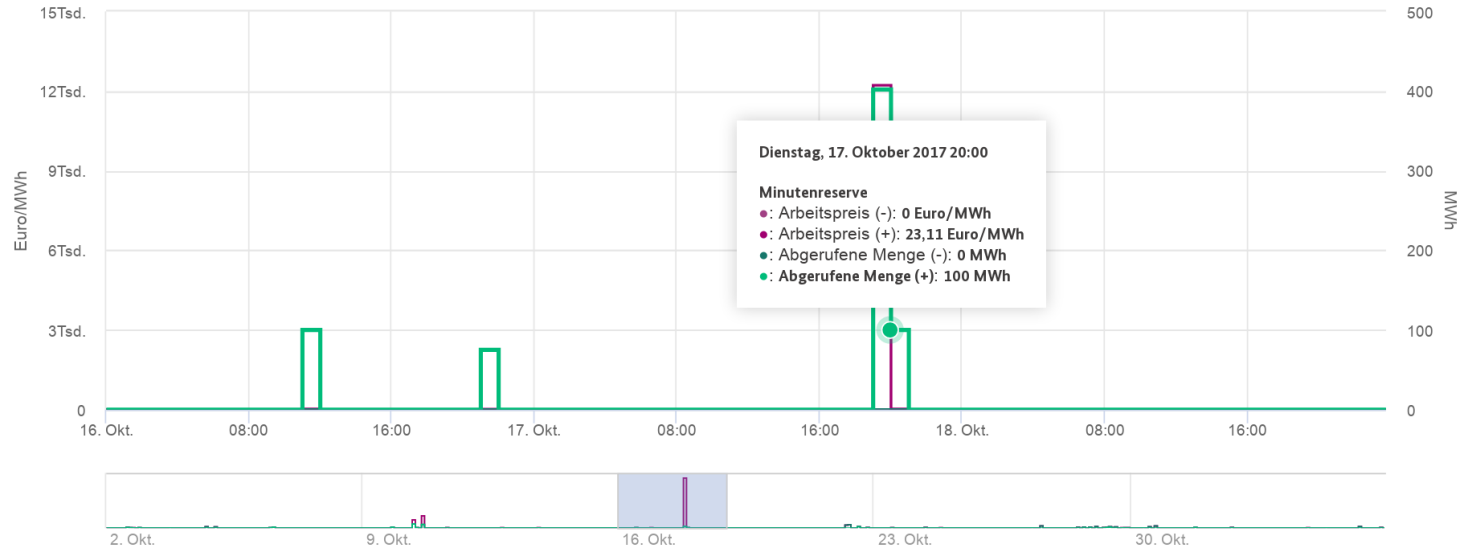
Incompetence is a viable Substitute



VNC Roulette

Market can also be the Culprit

Gaming a Critical Infrastructure?



Systemstabilität - Minutenreserve

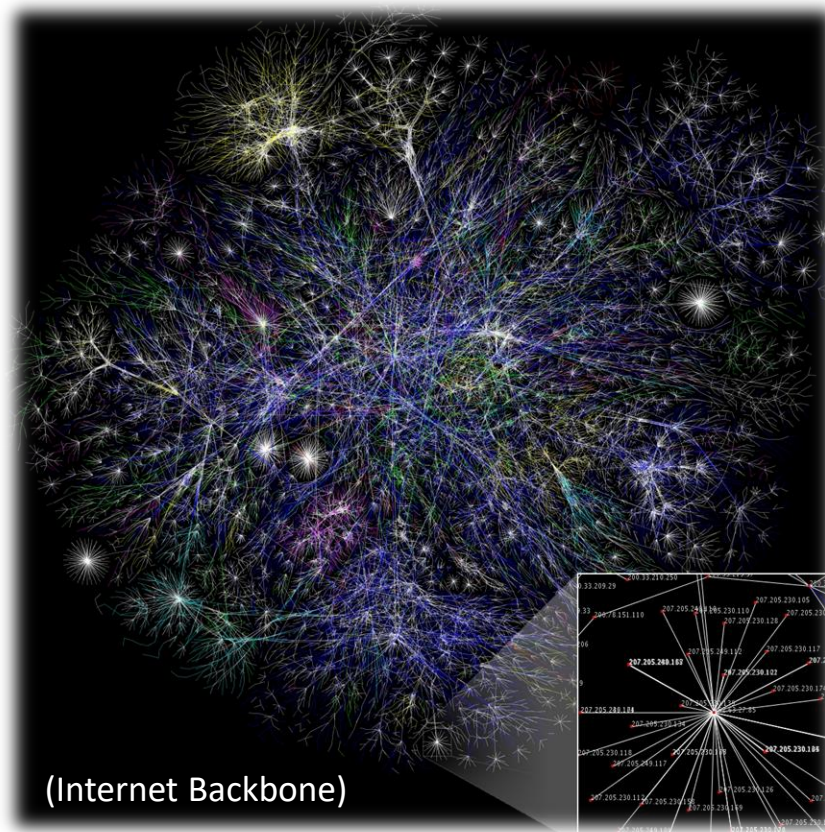
- Abgerufene Menge (+)
- Abgerufene Menge (-)
- Arbeitspreis (+)
- Arbeitspreis (-)
- Vorgehaltene Menge (+)
- Vorgehaltene Menge (-)
- Leistungspreis (+)
- Leistungspreis (-)



Learnings Resilient Control

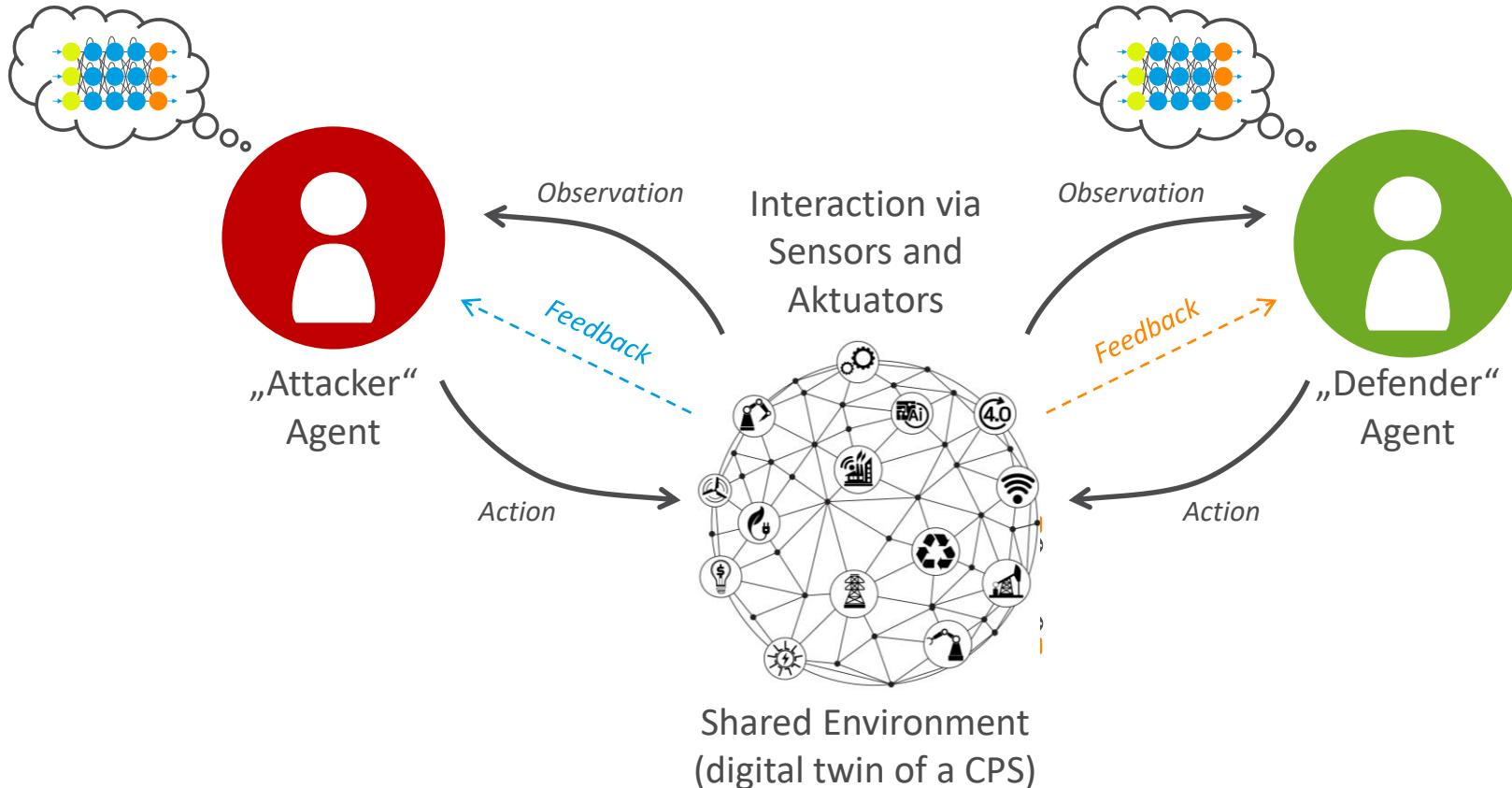
CPS inherently vulnerable

- > Interconnected CPS have always attack surface **due to their inherent complexity**
 - > Low latency of ICT and OT
 - > High interdependence
 - > Complexity in breadth and depth
 - > Critical Services as SPOF (DNS, BGP, SCADA, SDL)
- > Learning Strategies for **automatic issue mangement**



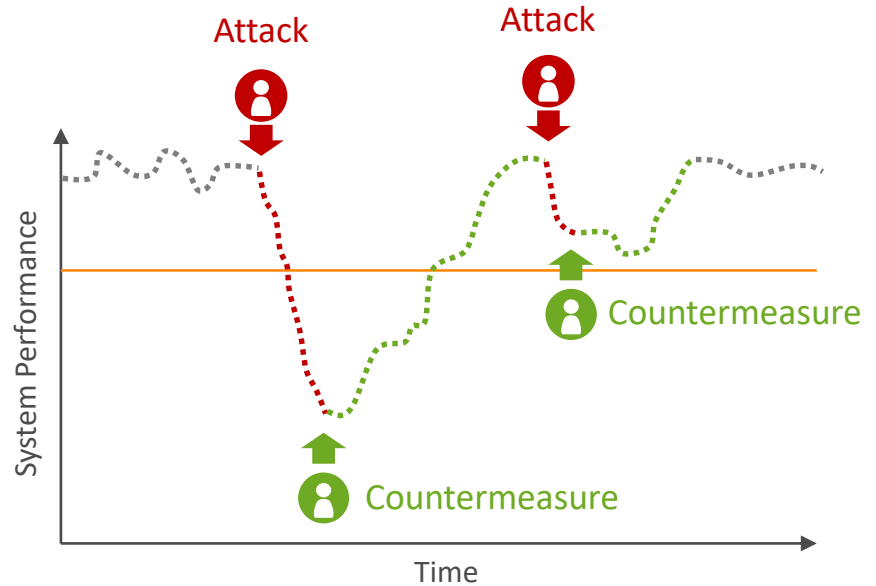
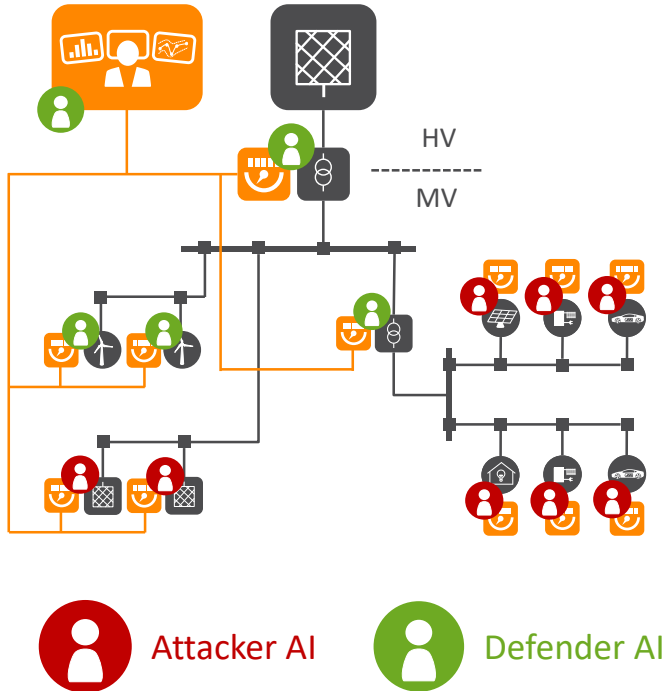
Adversarial Resilience Learning Concept

Competing Agents Learn in a Shared Environment



Demo: Attack on a Power System

Prevention of (sub-)system takeover as a secondary problem



Defender Points

2656

Loads Connected



Generators Connected



Buses Connected



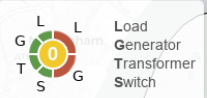
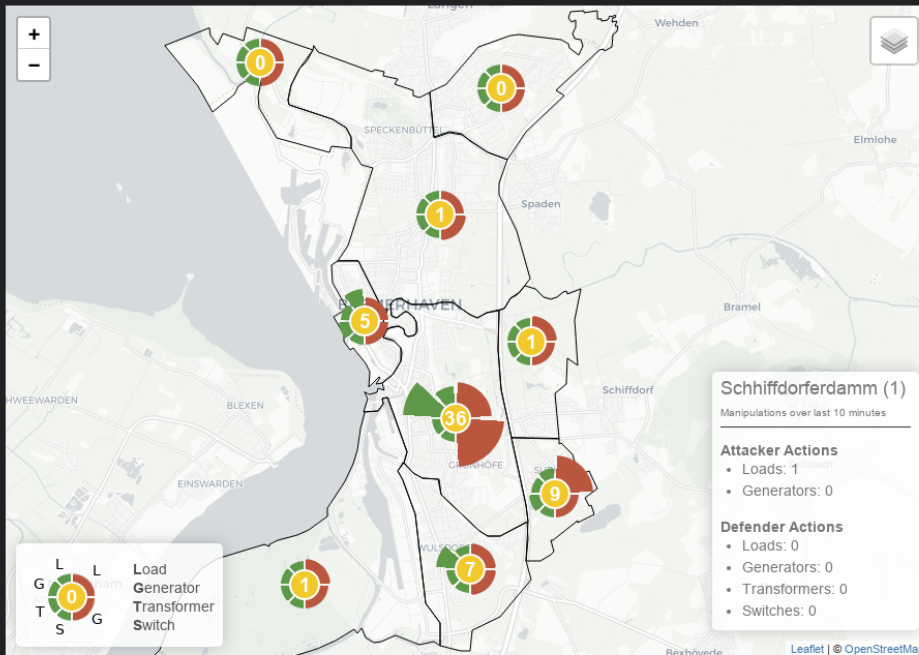
Transformers Connected...



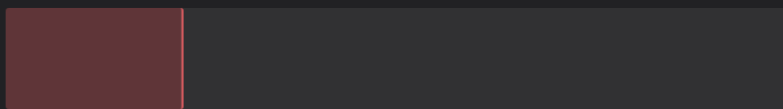
Most Valuable Actions (Defender)

Info	Time	Points
Changed Scaling from Lehe Households - 4 to 0.6667	2018-01-01 02:00:10	-0.25
Changed Scaling from Leherheide Industrielast to 0.8889	2018-01-01 02:26:10	-0.07
Changed Tap_pos from trafo to 1.0000	2018-01-01 01:47:50	-0.07
Changed Scaling from PV Fischereihafen to 0.0000	2018-01-01 02:14:30	-0.07
Changed Tap_pos from trafo to 1.0000	2018-01-01 01:45:30	-0.06
Changed Scaling from MEGS Klinikum	2018-01-01	

Map



Time Left (Coins Left in %)

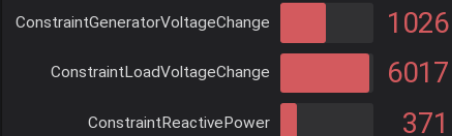


23%

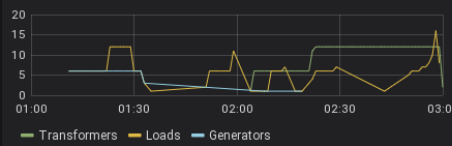
Attacker Points

7344

Constraint Violations



Malfunctions



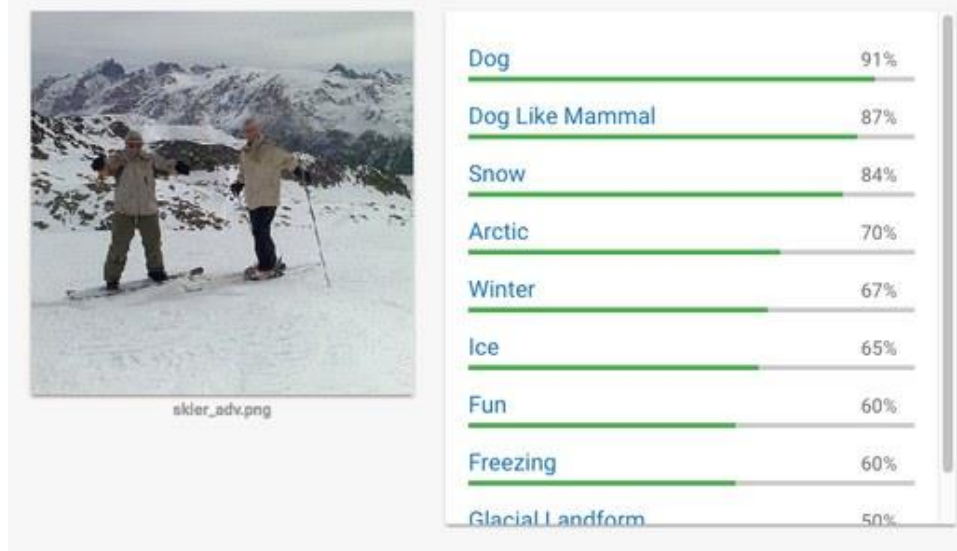
Most Valuable Actions (Attacker)

Info	Time	Points
Changed Scaling from Geestemünde Households - 0 to 0.5000	2018-01-01 01:00:00	0.96
Changed Scaling from Geestemünde Households - 0 to 0.5000	2018-01-01 01:00:00	0.93
Changed Scaling from Geestemünde Households - 0 to 0.5000	2018-01-01 01:00:00	0.91

Adversarial Learning

ARL != AL

- > Attempt to modify input data slightly in order to yield extremely different output from ANN
- > Modification of data not or only slightly visible to humans
- > Ex: RGB noise on a picture, small textual changes in spam messages
- > AL: Finding mechanisms against these attacks



Generative Adversarial Networks

ARL != GAN

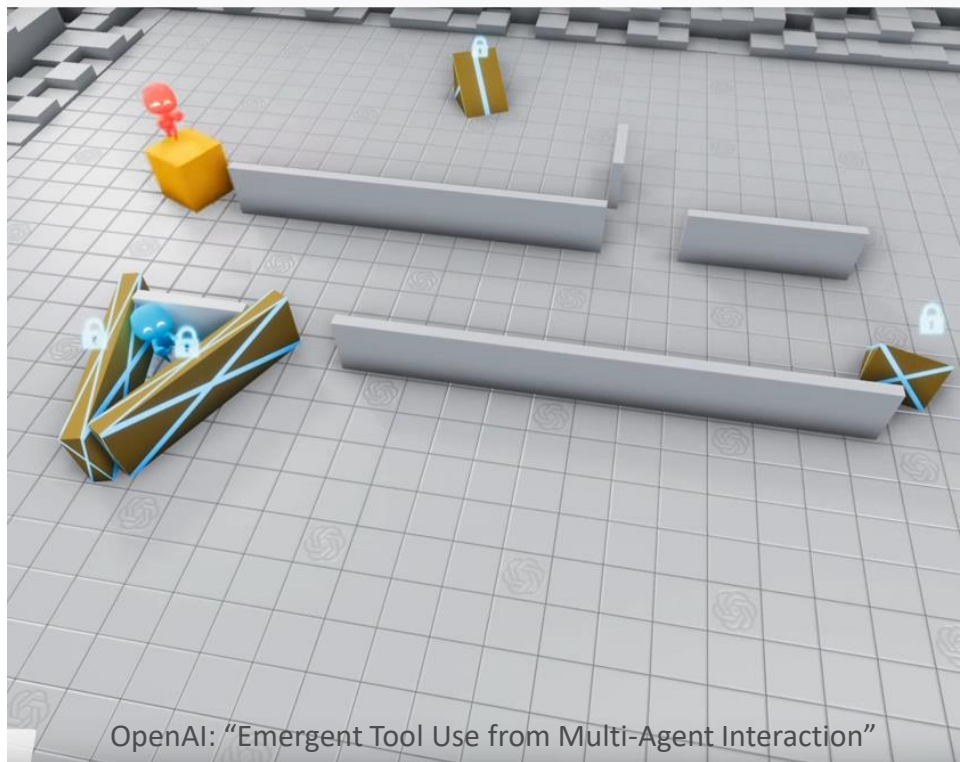
- > Two ANNs with **different roles**:
Generator and Discriminator
- > Zero-sum game
- > Generator maps vector of latent variables to feature space;
Discriminator evaluates the result
- > Error measure: Ability of the Discriminator to differentiate between real and generated data's distribution



Why does it work?

RL Agents discover bugs in the engine

- > Setup: Two groups of agents play *hide and seek*
- > No domain information; agents learn strategies and tool use independently
- > Result: Agents learn to exploit bugs in the underlying game engine
 - > Holes in walls
 - > Sliding boxes
 - > Edge/corner jumps

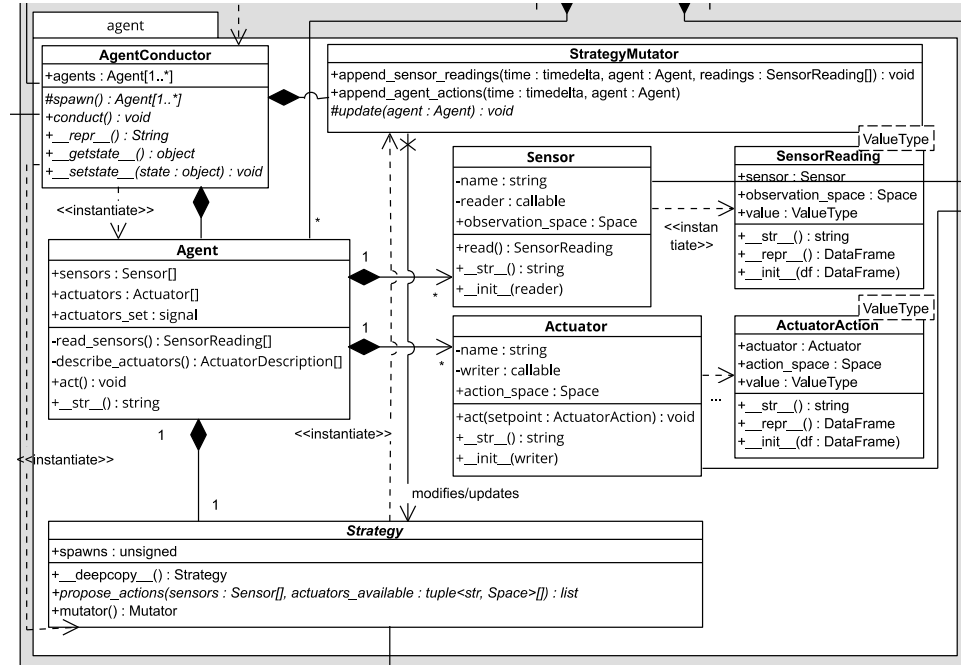


Multi-Algorithm

- > E.g., Q-learning vs. A3C (multiple workers)
- > Conductor orchestrates worker-environment pairs
- > Strategy: The “muscle” in an agent
- > Mutator: The “brain” for several agents

Simple Implementation API

- > Only “brain” and “muscle” need to be implemented for new algorithms

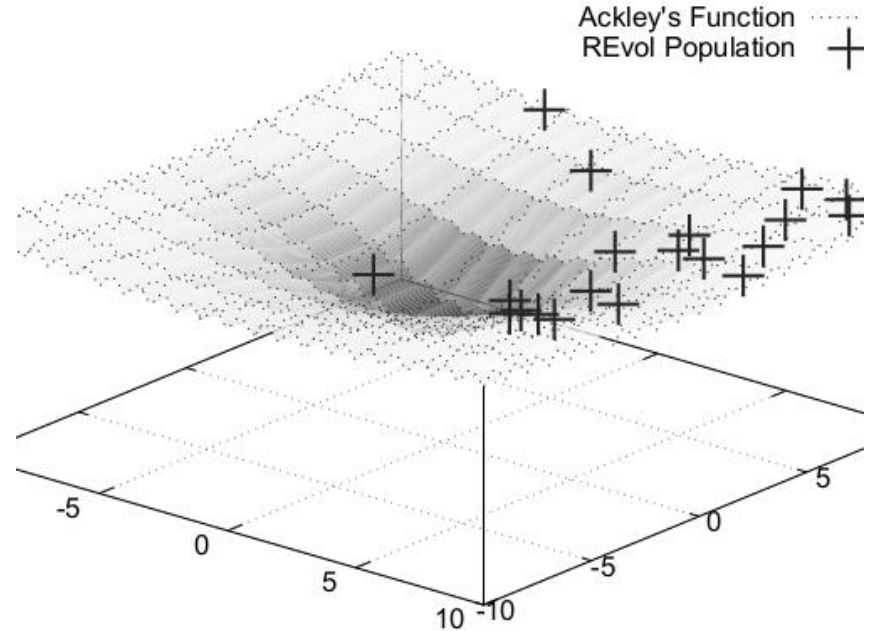


Motivation

- > Pre-defined policy network: Implicit domain knowledge

Combination of Revol and NEAT

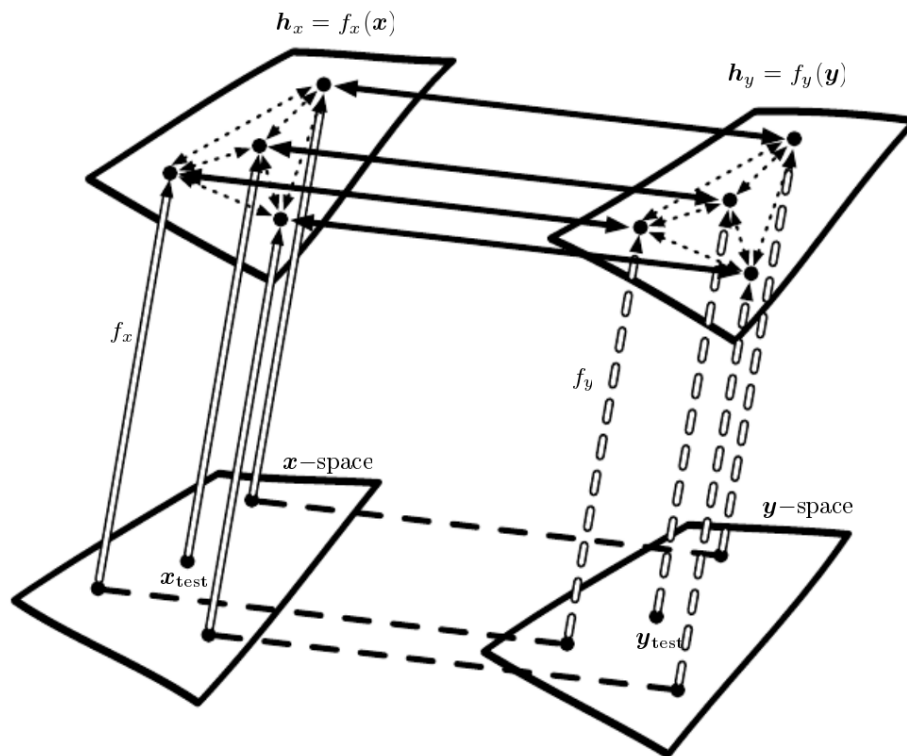
- > Implicit gradient information
- > Dynamic reproduction PDF
- > Speciation
- > Indirect encoding



Transfer Learning for Multi-Agent DRL

Ensure transferability with minimal re-training

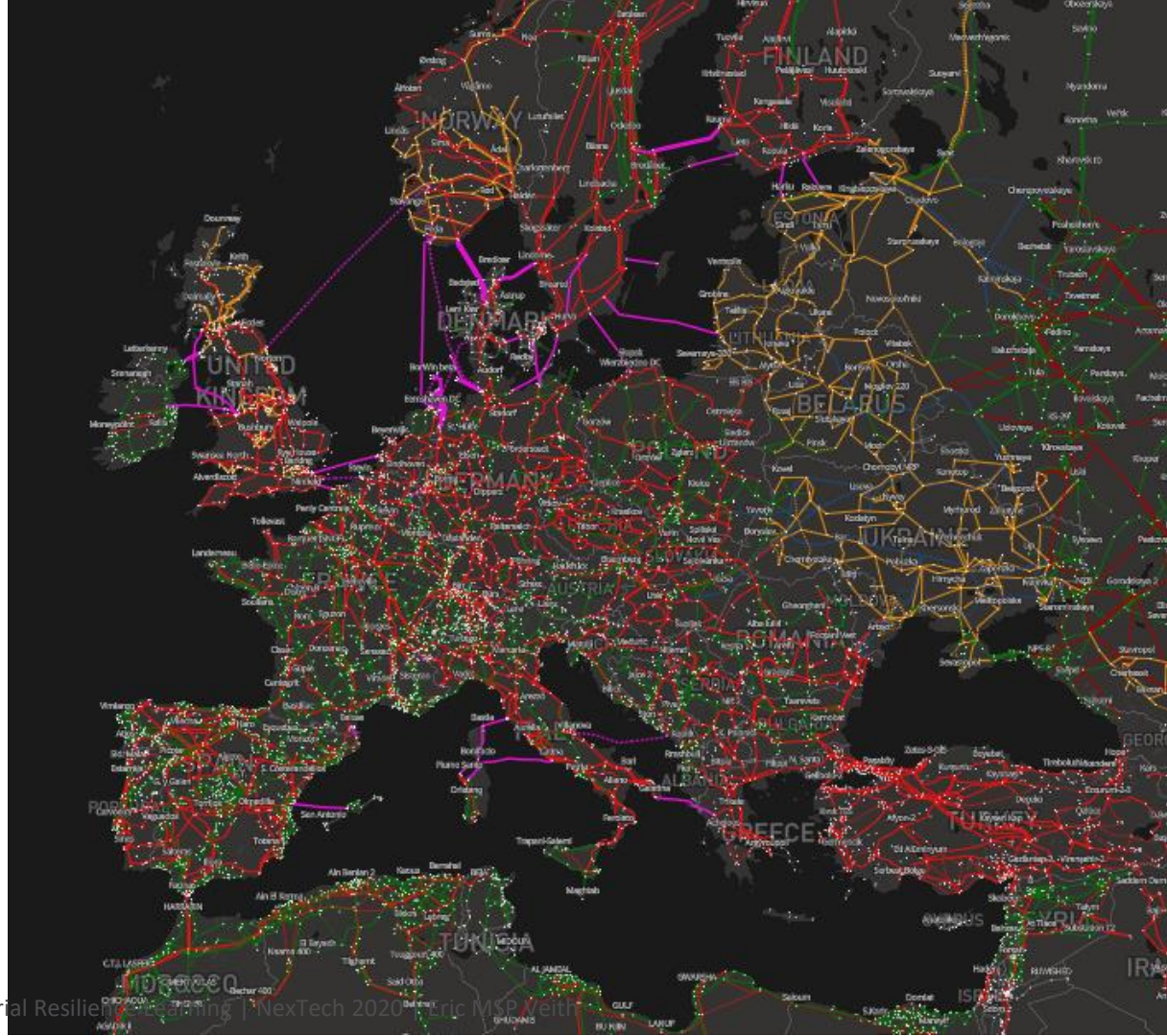
- > Agents can “conquer” their environment (think CTF!)
- > Different models in the same domain
- > Transfer between similar domains
- > Extraction of algorithms?



Rigging the Game

Strategic Infrastructure Extension

- > Weaknesses are indicator for strategic infrastructure investments
- > Calculate risks & losses, motivate investments

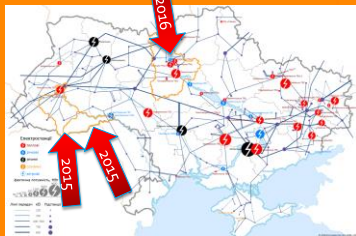


Analysis – attacker only

- > Resilient Systems Lab
- > Attacker explores vulnerabilities
- > „Conquest“ of a system
- > Attack vectors & log as basis of traditional analysis

Training – Attacker & Defender

- > AI for Grid Operation
- > Resilient overall system
- > Attacker trains defender
- > Attacks can be environmental factors
 - > Deviations in prognoses
 - > Accidents, etc.



Ethics of ARL

- > ARL a weapon?
- > **Lizence a solution?**
- > **Laws of Robotics** possibly inherent?

Conclusion



ARL enables discovery of vulnerabilities and interdependencies

- > Even when conform to regularizations! (EnWG, GridCodes, TAB etc.)

Development of defense (!) strategies

- > Ethic dilemma

„Attacker-Defender-Games“

- > Impact analysis in „anomalie-sensitive State Estimation“
- > Risk models, investment strategies (finding an equilibrium)
- > Analyzing asymmetries („Rigging the Game“)



GEFÖRDERT VOM

