# Putting smart agents through their paces

Our power grids could soon be operated by smart and explainable AI systems. A junior research group led by Eric Veith and funded by the Federal Ministry of Research, Technology and Space is investigating how to provide the best possible training to ensure an optimal response in critical situations.

By Ute Kehse

The 28 April 2025 is a date that people in Spain and Portugal will remember for many years to come. It was on that Monday at 12:33 pm that the Iberian Peninsula experienced a complete power blackout, with some places being left without electricity until the next day. The blackout was caused by a chain reaction of overvoltage disconnections: one power station after the next shut down until a critical point was reached and electricity generation in both countries collapsed completely within a matter of seconds. The Spanish government later published a report detailing the events that led to the outage. For Oldenburg computer scientist Dr Eric Veith, it's easy to see why human grid operators in control rooms would be stretched beyond their limits in a situation like that – despite the support of automated systems. "When a massive disruption occurs, the control room is flooded with notifications. Suddenly, all the screens light up like a Christmas tree." With so many decisions to take and only seconds to respond, it's virtually impossible for humans to develop a strategy to stop the fatal cascade.

The German power grid might have a reputation for being one of the world's most secure and stable, but the energy transition, with the increased fluctuations in electricity generation and consumption it entails, is intensifying the challenges for grid control, and cyber attacks are another risk. Veith's research is focused on creating modern AI systems to support the professionals at grid control centres in their daily work and thus help to make critical infrastructure more resilient in the face of unforeseen events. This would be difficult to achieve with conventional software, he explains. "There are so many incalculable factors in modern power grids, so many influences that people were unaware of when they were built, that it is simply no longer possible to develop software components that are prepared for every eventuality." Which is why his junior research group Adversarial Resilience Learning, funded by the Federal Ministry of Research, Technology and Space (BMFTR), is focusing on creating an AI-supported software system that can learn, is reliable, and whose decision-making is transparent.

To explain how this works, Veith starts by talking about the board game Go – and another memorable day, at least in the world of computer science. On 13 March 2016, the AI software AlphaGo achieved something that was thought to be impossible at that point: it won a game against the reigning world Go champion, Lee Sedol. "Go is the most complex board game ever invented, and the sheer multiplicity of options makes it extremely difficult for a software to develop a good strategy," Veith explains. It was ultimately a trick that enabled the AlphaGo programming team to beat the best human players: they first fed the AI software with moves used by Go champions and then had it play against a copy of itself countless times. Eventually, the programme learned the weak spots of its human opponents and was even able to develop hitherto unknown strategies.

## An "evil twin" challenges the operator agent

Pitching two identical software programmes against one another to produce increasingly sophisticated tactics – Veith is employing the same strategy as the AlphaGo team in his own research project. In a method known as Autocurricular Deep Reinforcement Learning, he and his five-person team feed the computer programme – which the researchers refer to as "the agent" – with information about a system via various sensors. "The system could be a playing field or it could be a power grid," Veith explains. The agent also includes a training algorithm which is based on a neuronal network, i.e., an AI programme modelled on the biological neuronal architecture. This structure allows the neural network to make decisions in a similar way to the human brain. It doesn't follow a pre-programmed pattern, but to a certain extent programmes itself, learns from experience and is able to deal with new situations.

Veith's agent is given a so-called reward function that describes, in mathematical terms, a desired state of the system. In the case of the power grid, this reward could be used to ensure that the power frequency and voltage levels remain within certain parameters. "But the strategy is not laid out in advance", the computer scientist emphasises. The agent detects something, reacts and then calculates a feedback signal in order to determine whether it has fulfilled the task at hand. To achieve the objective – for instance, stabilising the grid frequency – the agent has several options: it can switch on power stations, disconnect consumers from the grid or adjust controller settings, to name a few. It can also shift reactive power, which is necessary for voltage stabilisation, or activate protective devices. With the help of the training algorithm it learns how to best achieve its goal. "These agent systems are proactive rather than simply reactive. And we don't have to prescribe how this system should achieve something; we just have to tell it what the desired state is," Veith adds.

To train their agents, the researchers first had to develop an appropriate simulation environment. Creating a realistic replica of a power grid was one of the most complex parts of the project. "Our idea, inspired by AlphaGo, was to set up not just one agent to stabilise the grid but to use a second one too, a sort of evil twin which would work to achieve the opposite effect, then let them battle it out", Veith explains. This would effectively confront the operator agent with a constant stream of new problems and make it learn faster. Depending on how the researchers configure the second "challenger" agent, it could try to outwit its opponent and destabilise the power grid by simulating cyber attacks, extreme weather situations, or a surge in demand for electricity precipitated by thousands of smart home devices being switched on simultaneously.

"Using this basic idea, we started work in 2018 in Sebastian Lehnhoff's research group at the OFFIS Institute for Information Technology and made a lot of headway," Veith recalls. But like so many AI programmes, the team's original system of agents had one fundamental flaw: it provided no information about how its results were achieved. Yet explainability is essential if these agents are to be deployed in critical infrastructure, especially because in the course of their training AI programmes sometimes learn things that don't make much sense. "This is why we had to expand the original concept in a fundamental way, focusing on why an agent does what it does," Veith clarifies.

It is this objective that the researcher has been pursuing since 2022 – both in his junior research group at the University of Oldenburg and also in two EU-funded joint projects in collaboration with the University-affiliated OFFIS Institute. Industry partners such as Austrian energy provider Wiener Netze and Stuttgart-based Netze BW were also involved.

To guarantee transparency, the team developed an algorithm that converts the agent's strategy into an "equivalent decision tree". This enabled the researchers to map out, step by step, the rules that the programme follows when making its decisions. "We can see exactly which threshold values for which sensors lead to a particular decision so we can determine whether it makes sense in the real world or whether the agent has been fooled by a statistical anomaly during learning," explains Veith.

"Our 'evil' agent soon learned how to attack the power grid"

The next step for the team is to optimise the agent's training. "Training it from scratch is extremely resource-intensive, as several million simulation steps are needed for it to learn a meaningful, transferable strategy," the researcher explains. To speed up the process, the team developed a method for incorporating the industry partners' know-how into the agent's training. One example was a street in Vienna where many of the residents drive electric cars. "If they all come home from work at more or less the same time and want to charge their cars, it's bad for the grid," Veith explains. The team was able to put this negative pattern to good use in the challenger agent. "It quickly learned how to apply this information to attack the power grid," says Veith. The operator agent, in turn, was forced to develop a counter-strategy, to which the challenger agent then responded – a game that the researchers continued for a while. "In the end, we had a decision tree and could see that the strategy indeed made sense."

It was also the decision tree that helped the team solve another typical AI software problem – that of "catastrophic forgetting". When an AI is being trained with new data, previously learned patterns occasionally get overwritten. As a countermeasure, Veith and his doctoral students programmed their agent to compare decisions made by its own training algorithm with rules listed earlier in the decision trees. If discrepancies arise, it can then initiate retraining with the old data. "We call this rehearsal. The neural network then has to go through the old cases again, because it has clearly forgotten them", Veith explains. This combination enables the agent to learn much faster and achieve reliable results.

It will probably be a while before the programme can actually be used in a grid control centre, the researcher notes: "It's hardly feasible to conduct field tests with a real power grid at this point in time, so we're currently discussing other possible applications for our system, for example in crisis prevention." Nevertheless, the stated goal is to create a system that can be deployed in real-life situations – reliable software that makes the right decision when it really matters.

Eric Veith and his team have developed an AI that can operate power grids and deal effectively with unexpected problems.

Outlooks

## Who creates the future?

**Prof. Dr Martin Butler**

Director of the research centre "Genealogy of the Present"

The future is not yet set. This is what makes it uncertain. In times of multiple crises, when wars and other catastrophes dominate the headlines, uncertainty, for many, becomes a real "crisis of assurance". People deal with this in very different ways: some trust the reliability of scientific prediction; others believe in the visions of political decision-makers or in religious promises of salvation. At the interdisciplinary and interdepartmental research centre "Genealogy of the Present", we look at the different narratives that are emerging about the future, how they compete for authority, and how they shape present-day thinking and action.

For example, in the discourse on the future of Western industrialised nations, one side regards migration as critical to addressing the problem of ageing populations and shortages of skilled laborers, whereas the other side sees it as a danger that must be averted. These two narratives about the future compete with one another, informing the opinions and decisions of today.

Alarmingly, in this battle for attention, the impact of a narrative is not necessarily tied to its factual accuracy. So a tweet by the US president, posted while he is on the golf course, may have more societal influence than a scientific study. Therefore, how and why a specific vision of the future prevails, depends on who is involved in creating it, or in what context and with what reach it is disseminated, among other things.

Our research looks at all these aspects: our aim is to uncover the mechanisms that create and popularise visions of the future, and in so doing, shed light on the relationship between future narratives and present-day action. At a time when the institution of science and the authority of its prognoses are being questioned more than ever before, this perspective is of particular significance.