USENIX SECURITY '26

The Pitfalls of Deep Reinforcement Learning for Cybersecurity

Shae McFadden[1,2,3], Myles Foley[2,6], Elizabeth Bates[2], Ilias Tsingenopoulos[5], Sanyam Vyas[4,2], Vasilios Mavroudis[1], Chris Hicks[2], Fabio Pierazzi[3]

[1]King's College London, [2]The Alan Turing Institute, [3]University College London, [4]Cardiff University, [5]KU Leuven, [6]Devotion AI Labs

01 Modeling 02 Training 03 Evaluation 04 Deployment

Applying DRL to Cybersecurity

Article

Grandmaster level in StarCraft II using multi-agent reinforcement learning

https://doi.org/10.1038/s41586-019-1724-z

Received: 30 August 2019

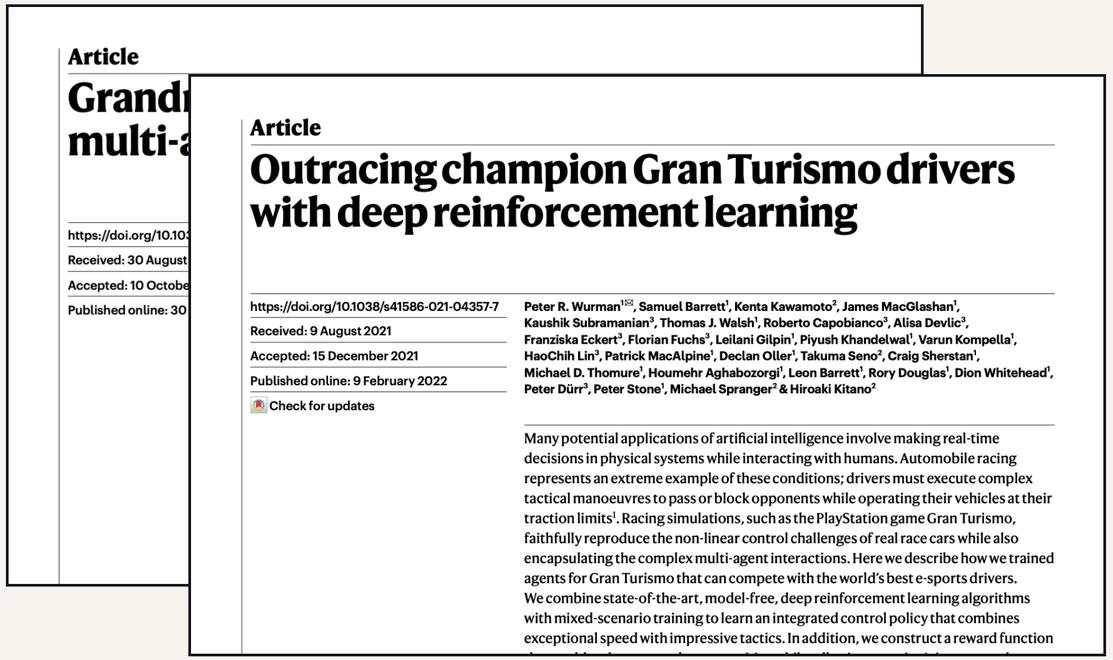
Accepted: 10 October 2019

Published online: 30 October 2019

Oriol Vinyals[1,3]*, Igor Babuschkin[1,3], Wojciech M. Czarnecki[1,3], Michaël Mathieu[1,3], Andrew Dudzik[1,3], Junyoung Chung[1,3], David H. Choi[1,3], Richard Powell[1,3], Timo Ewalds[1,3], Petko Georgiev[1,3], Junhyuk Oh[1,3], Dan Horgan[1,3], Manuel Kroiss[1,3], Ivo Danihelka[1,3], Aja Huang[1,3], Laurent Sifre[1,3], Trevor Cai[1,3], John P. Agapiou[1,3], Max Jaderberg[1], Alexander S. Vezhnevets[1], Rémi Leblond[1], Tobias Pohlen[1], Valentin Dalibard[1], David Budden[1], Yury Sulsky[1], James Molloy[1], Tom L. Paine[1], Caglar Gulcehre[1], Ziyu Wang[1], Tobias Pfaff[1], Yuhuai Wu[1], Roman Ring[1], Dani Yogatama[1], Dario Wünsch[2], Katrina McKinney[1], Oliver Smith[1], Tom Schaul[1], Timothy Lillicrap[1], Koray Kavukcuoglu[1], Demis Hassabis[1], Chris Apps[1,3] & David Silver[1,3]*

Many real-world applications require artificial agents to compete and coordinate with other agents in complex environments. As a stepping stone to this goal, the domain of StarCraft has emerged as an important challenge for artificial intelligence research, owing to its iconic and enduring status among the most difficult professional esports and its relevance to the real world in terms of its raw complexity and multi-agent challenges. Over the course of a decade and numerous competitions[1–3], the strongest agents have simplified important aspects of the game, utilized superhuman capabilities, or employed hand-crafted sub-systems[4]. Despite these advantages, no previous agent has come close to matching the overall skill of

Applying DRL to Cybersecurity



Article

Grandi
multi-a

https://doi.org/10.10
Received: 30 August
Accepted: 10 Octobe
Published online: 30

Article

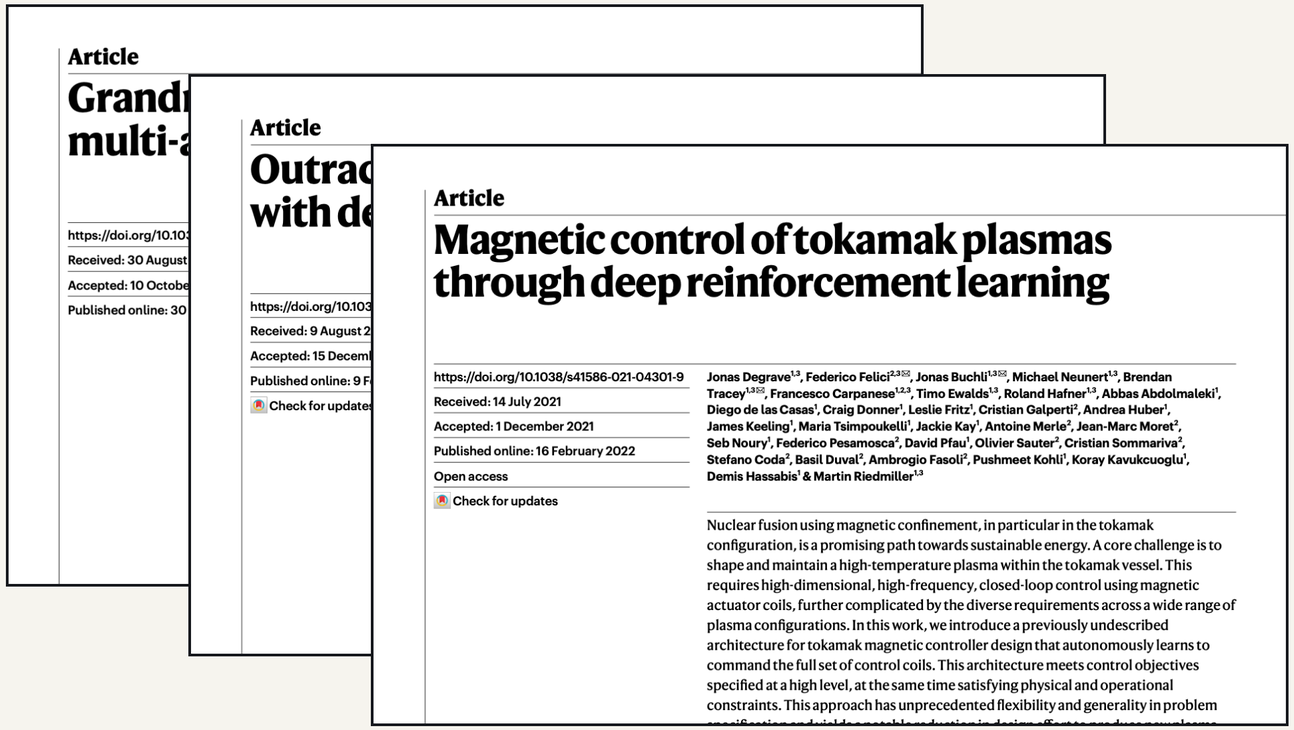
Outracing champion Gran Turismo drivers with deep reinforcement learning

https://doi.org/10.1038/s41586-021-04357-7
Received: 9 August 2021
Accepted: 15 December 2021
Published online: 9 February 2022
Check for updates

Peter R. Wurman[1✉], Samuel Barrett[1], Kenta Kawamoto[2], James MacGlashan[1], Kaushik Subramanian[3], Thomas J. Walsh[1], Roberto Capobianco[3], Alisa Devlic[3], Franziska Eckert[3], Florian Fuchs[3], Leilani Gilpin[1], Piyush Khandelwal[1], Varun Kompella[1], HaoChih Lin[3], Patrick MacAlpine[1], Declan Oller[1], Takuma Seno[2], Craig Sherstan[1], Michael D. Thomure[1], Houmehr Aghabozorgi[1], Leon Barrett[1], Rory Douglas[1], Dion Whitehead[1], Peter Dürr[3], Peter Stone[1], Michael Spranger[2] & Hiroaki Kitano[2]

Many potential applications of artificial intelligence involve making real-time decisions in physical systems while interacting with humans. Automobile racing represents an extreme example of these conditions; drivers must execute complex tactical manoeuvres to pass or block opponents while operating their vehicles at their traction limits[1]. Racing simulations, such as the PlayStation game Gran Turismo, faithfully reproduce the non-linear control challenges of real race cars while also encapsulating the complex multi-agent interactions. Here we describe how we trained agents for Gran Turismo that can compete with the world's best e-sports drivers. We combine state-of-the-art, model-free, deep reinforcement learning algorithms with mixed-scenario training to learn an integrated control policy that combines exceptional speed with impressive tactics. In addition, we construct a reward function

Applying DRL to Cybersecurity



Article

Grandi
multi-a

https://doi.org/10.10
Received: 30 August
Accepted: 10 Octobe
Published online: 30

Article

Outrac
with de

https://doi.org/10.103
Received: 9 August 2
Accepted: 15 Decem
Published online: 9 F
Check for updates

Article

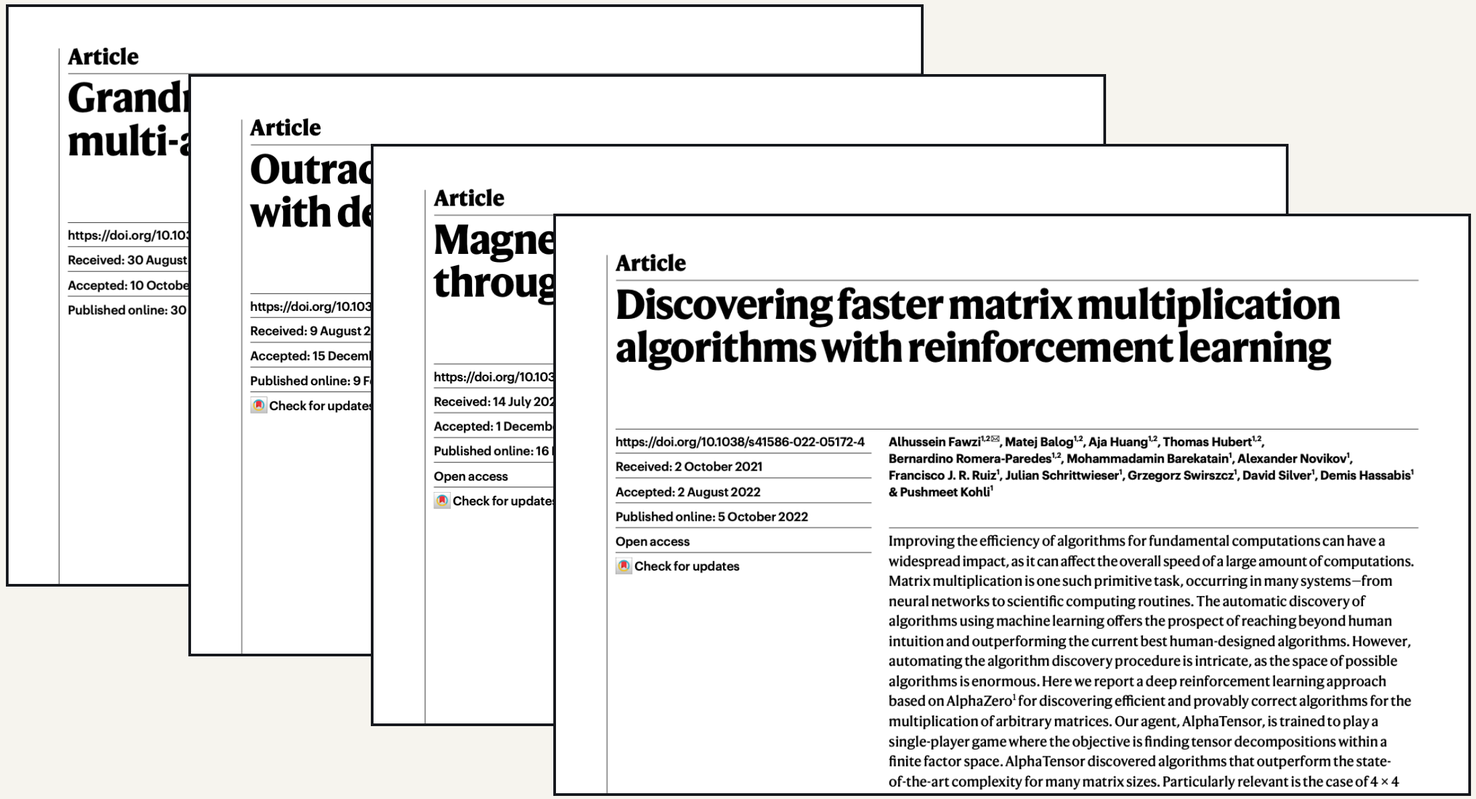
Magnetic control of tokamak plasmas through deep reinforcement learning

https://doi.org/10.1038/s41586-021-04301-9
Received: 14 July 2021
Accepted: 1 December 2021
Published online: 16 February 2022
Open access
Check for updates

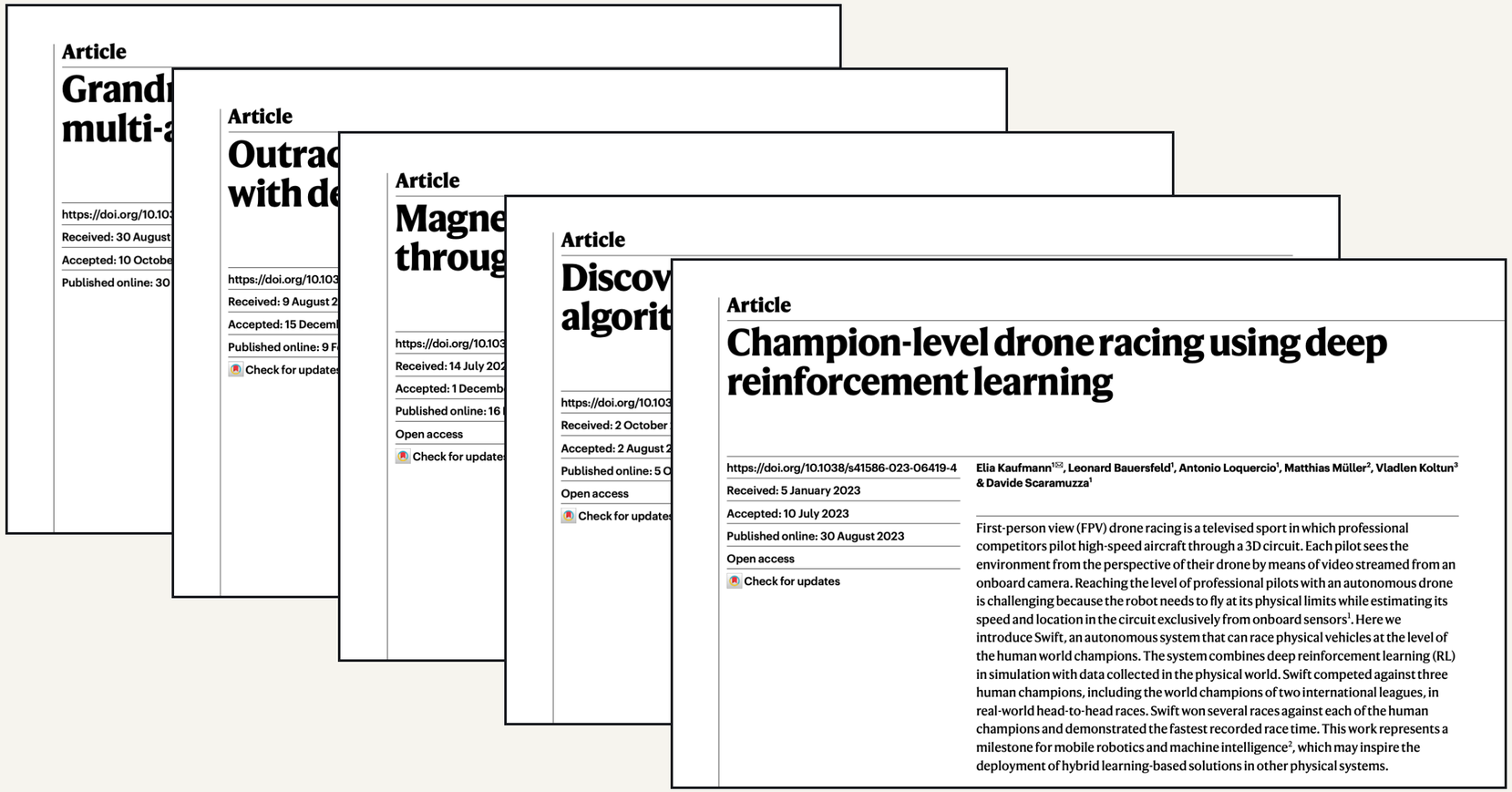
Jonas Degrave[1,3], Federico Felici[2,3]✉, Jonas Buchli[1,3]✉, Michael Neunert[1,3], Brendan Tracey[1,3]✉, Francesco Carpanese[1,2,3], Timo Ewalds[1,3], Roland Hafner[1,3], Abbas Abdolmaleki[1], Diego de las Casas[1], Craig Donner[1], Leslie Fritz[1], Cristian Galperti[2], Andrea Huber[1], James Keeling[1], Maria Tsimpoukelli[1], Jackie Kay[1], Antoine Merle[2], Jean-Marc Moret[2], Seb Noury[1], Federico Pesamosca[2], David Pfau[1], Olivier Sauter[2], Cristian Sommariva[2], Stefano Coda[2], Basil Duval[2], Ambrogio Fasoli[2], Pushmeet Kohli[1], Koray Kavukcuoglu[1], Demis Hassabis[1] & Martin Riedmiller[1,3]

Nuclear fusion using magnetic confinement, in particular in the tokamak configuration, is a promising path towards sustainable energy. A core challenge is to shape and maintain a high-temperature plasma within the tokamak vessel. This requires high-dimensional, high-frequency, closed-loop control using magnetic actuator coils, further complicated by the diverse requirements across a wide range of plasma configurations. In this work, we introduce a previously undescribed architecture for tokamak magnetic controller design that autonomously learns to command the full set of control coils. This architecture meets control objectives specified at a high level, at the same time satisfying physical and operational constraints. This approach has unprecedented flexibility and generality in problem

Applying DRL to Cybersecurity



Applying DRL to Cybersecurity



Article

Grandi
multi-a

https://doi.org/10.10
Received: 30 August
Accepted: 10 Octobe
Published online: 30

Article

Outrac
with de

https://doi.org/10.103
Received: 9 August 2
Accepted: 15 Decem
Published online: 9 F
Check for update

Article

Magne
throug

https://doi.org/10.103
Received: 14 July 202
Accepted: 1 Decemb
Published online: 16
Open access
Check for update

Article

Discov
algorit

https://doi.org/10.103
Received: 2 October
Accepted: 2 August 2
Published online: 5 O
Open access
Check for update

Article

Champion-level drone racing using deep reinforcement learning

https://doi.org/10.1038/s41586-023-06419-4
Received: 5 January 2023
Accepted: 10 July 2023
Published online: 30 August 2023
Open access
Check for updates

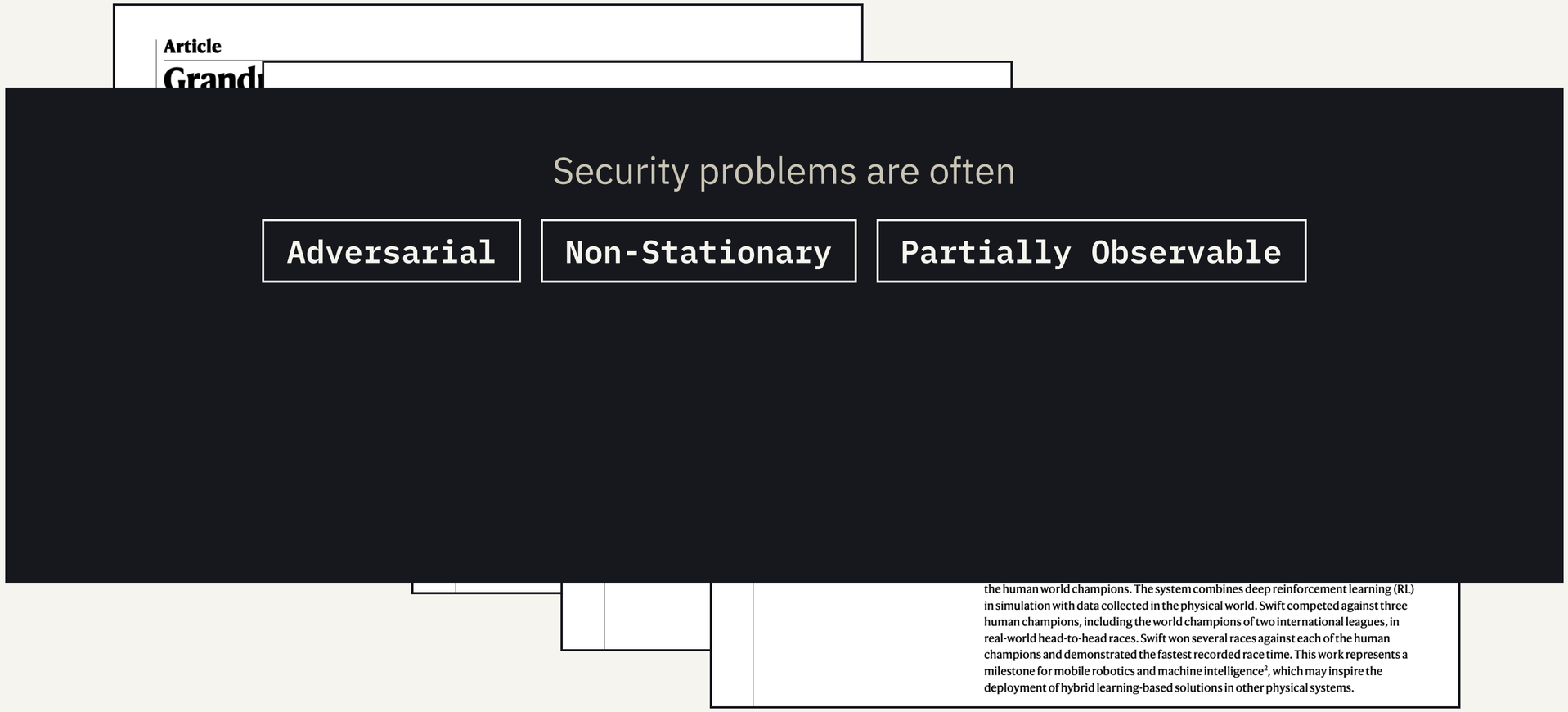
Elia Kaufmann[1✉], Leonard Bauersfeld[1], Antonio Loquercio[1], Matthias Müller[2], Vladlen Koltun[3] & Davide Scaramuzza[1]

First-person view (FPV) drone racing is a televised sport in which professional competitors pilot high-speed aircraft through a 3D circuit. Each pilot sees the environment from the perspective of their drone by means of video streamed from an onboard camera. Reaching the level of professional pilots with an autonomous drone is challenging because the robot needs to fly at its physical limits while estimating its speed and location in the circuit exclusively from onboard sensors[1]. Here we introduce Swift, an autonomous system that can race physical vehicles at the level of the human world champions. The system combines deep reinforcement learning (RL) in simulation with data collected in the physical world. Swift competed against three human champions, including the world champions of two international leagues, in real-world head-to-head races. Swift won several races against each of the human champions and demonstrated the fastest recorded race time. This work represents a milestone for mobile robotics and machine intelligence[2], which may inspire the deployment of hybrid learning-based solutions in other physical systems.

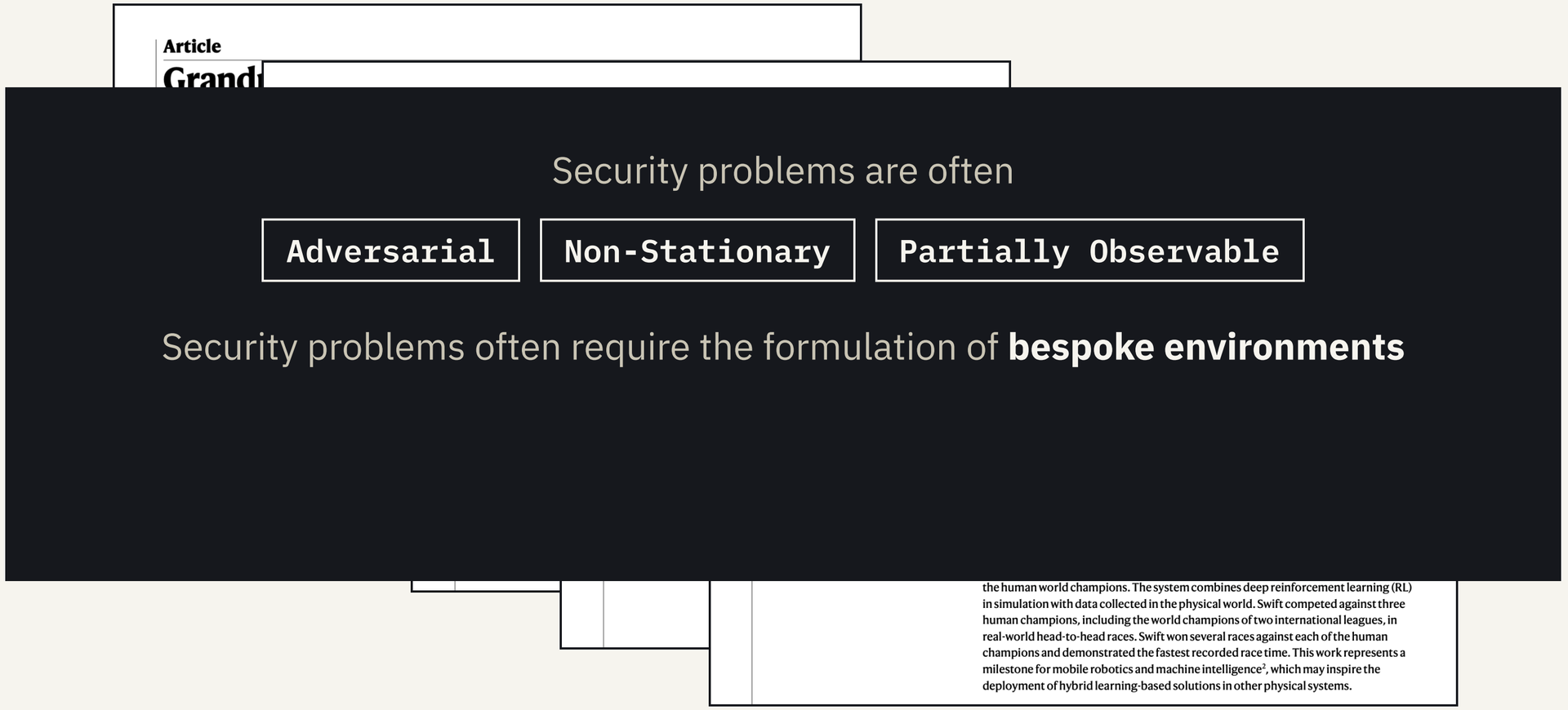
Applying DRL to Cybersecurity

Security problems are often

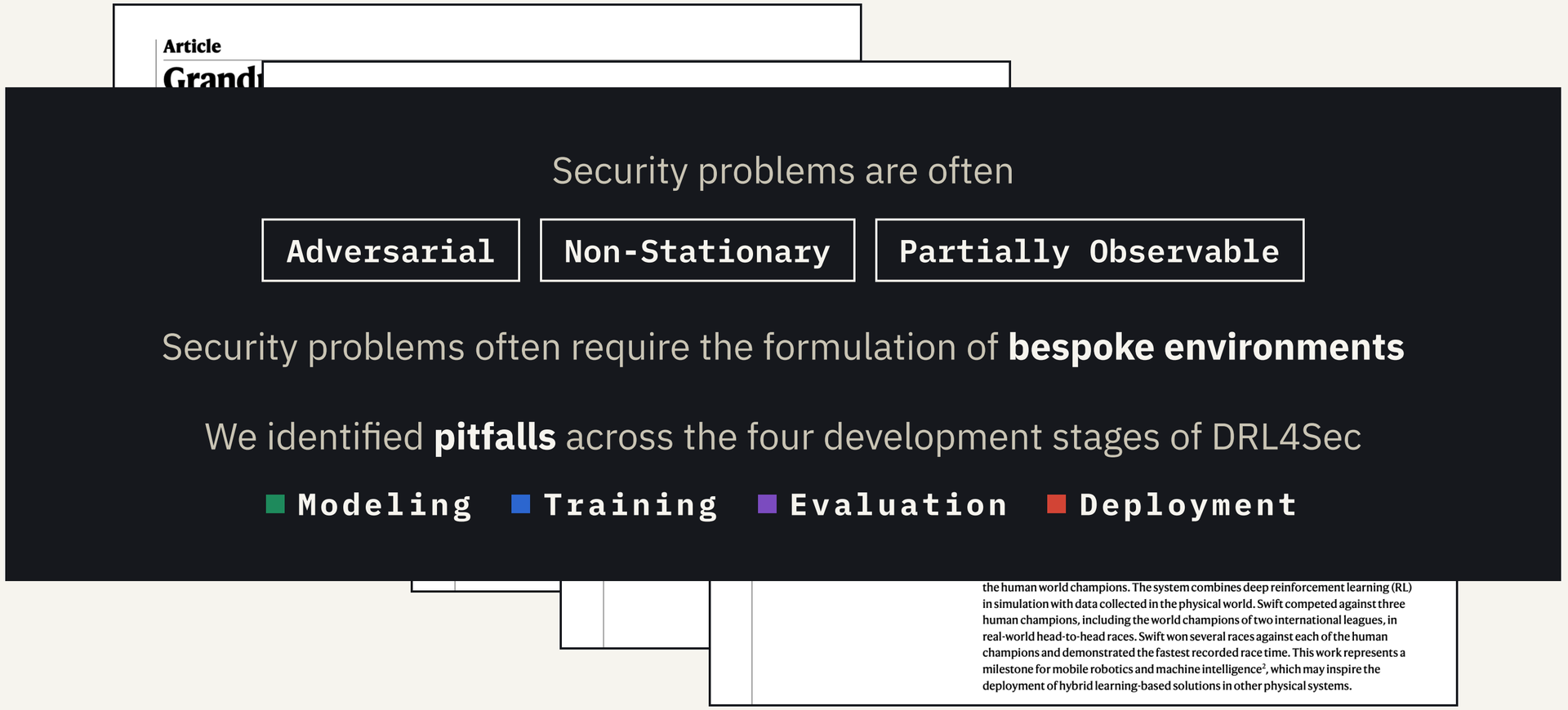
Adversarial **Non-Stationary** **Partially Observable**



Applying DRL to Cybersecurity



Applying DRL to Cybersecurity



Security problems are often

Adversarial **Non-Stationary** **Partially Observable**

Security problems often require the formulation of **bespoke environments**

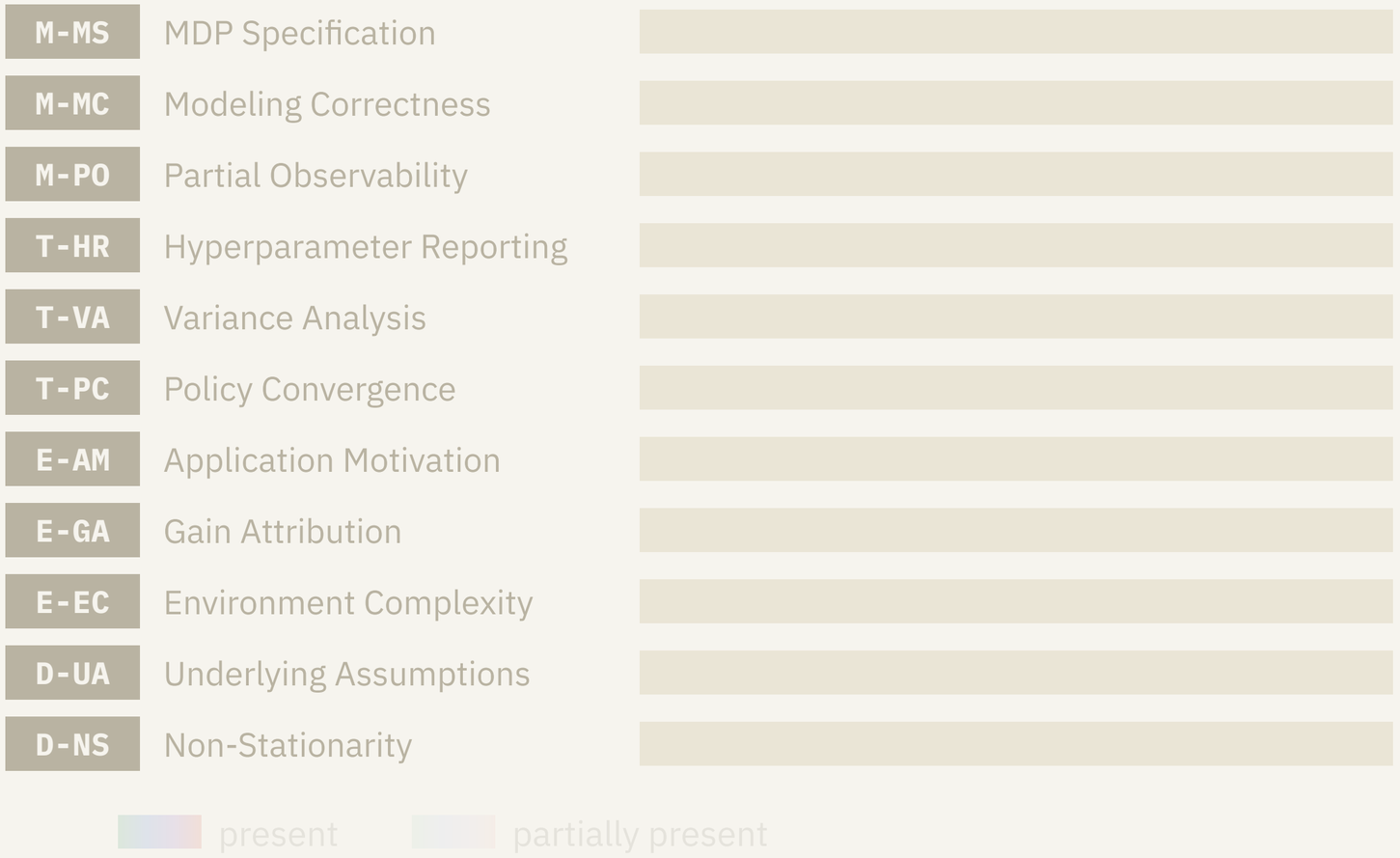
We identified **pitfalls** across the four development stages of DRL4Sec

Modeling **Training** **Evaluation** **Deployment**

Pitfall Prevalence

11

pitfalls identified



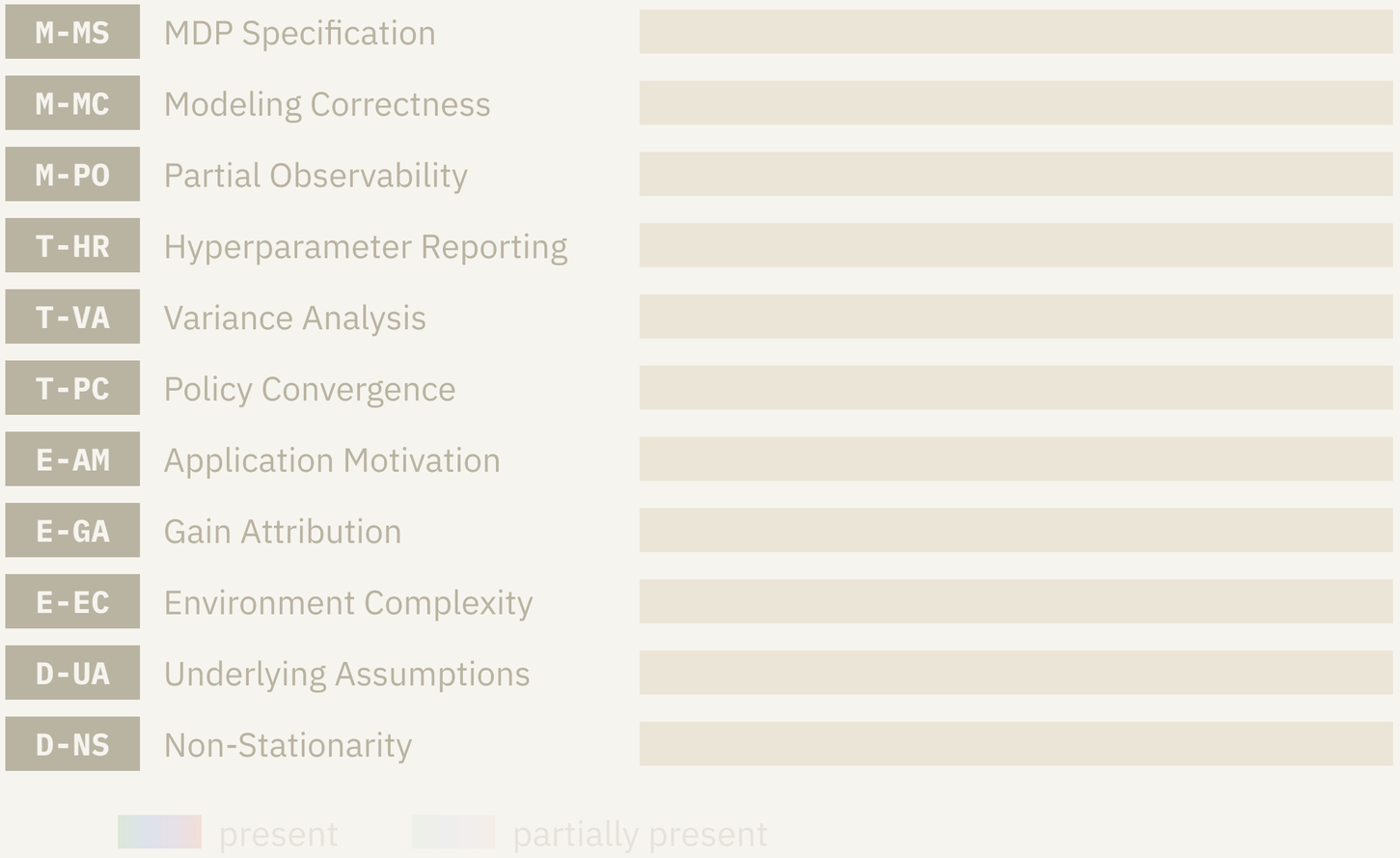
Pitfall Prevalence

11

pitfalls identified

66

papers reviewed



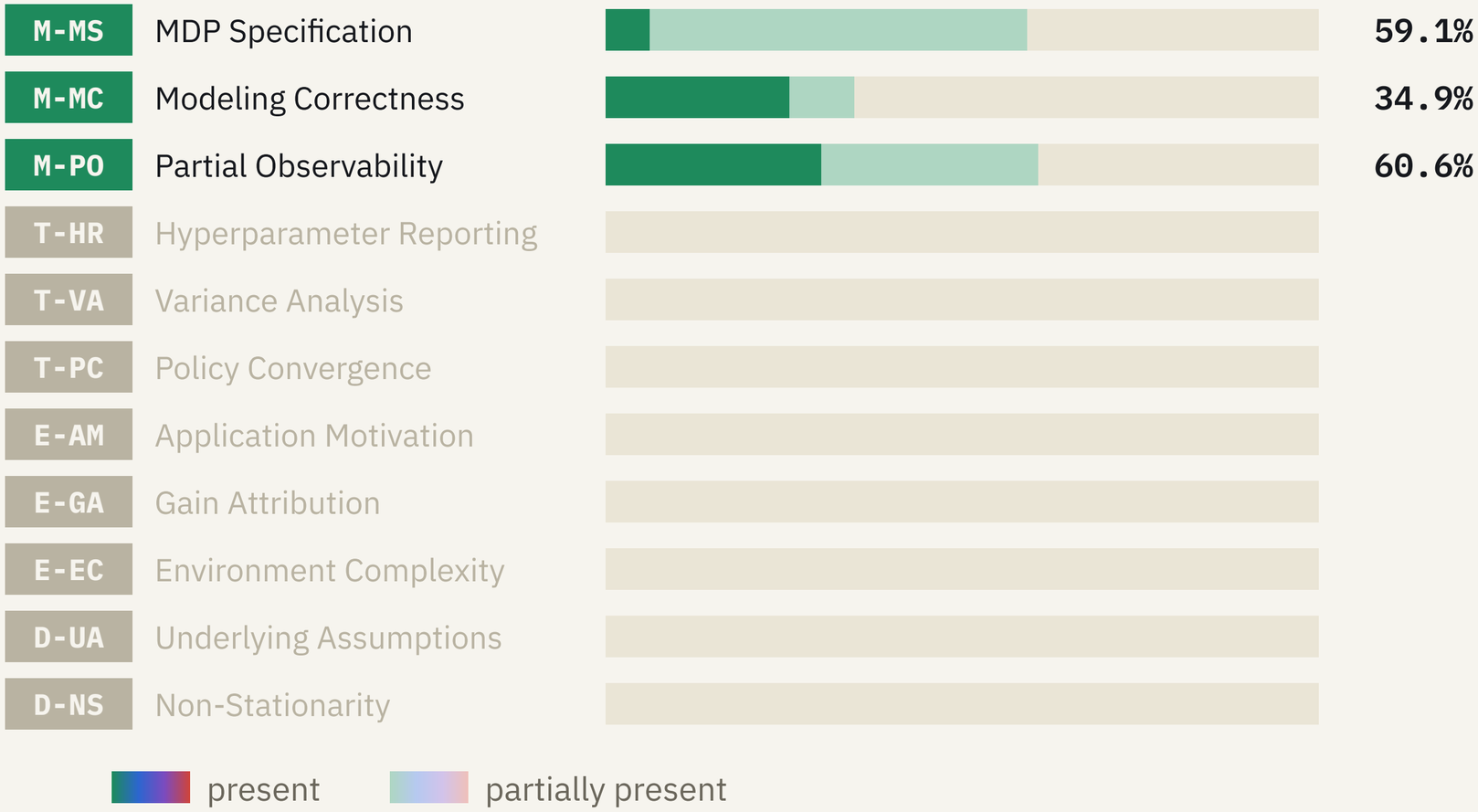
Pitfall Prevalence

11

pitfalls identified

66

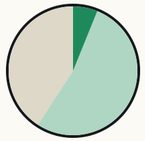
papers reviewed



Modeling Environments

Modeling Environments

M-MS



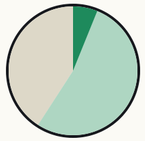
MDP Specification

The MDP is ill-defined, so that one or more components (state space, action space, reward function, and transition dynamics) remain unclear.

59%

Modeling Environments

M-MS

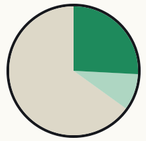


MDP Specification

The MDP is ill-defined, so that one or more components (state space, action space, reward function, and transition dynamics) remain unclear.

59%

M-MC



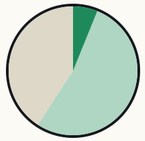
Modeling Correctness

The modeled environment violates the Markov assumption or does not correctly encapsulate the actual MDP underlying the security problem.

35%

Modeling Environments

M-MS

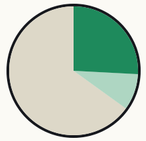


MDP Specification

The MDP is ill-defined, so that one or more components (state space, action space, reward function, and transition dynamics) remain unclear.

59%

M-MC

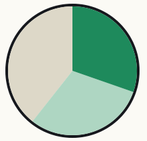


Modeling Correctness

The modeled environment violates the Markov assumption or does not correctly encapsulate the actual MDP underlying the security problem.

35%

M-PO



Partial Observability

The environment is inherently partially observable (i.e. a POMDP); however, it is treated as if it were fully observable without mitigating the partial observability.

61%

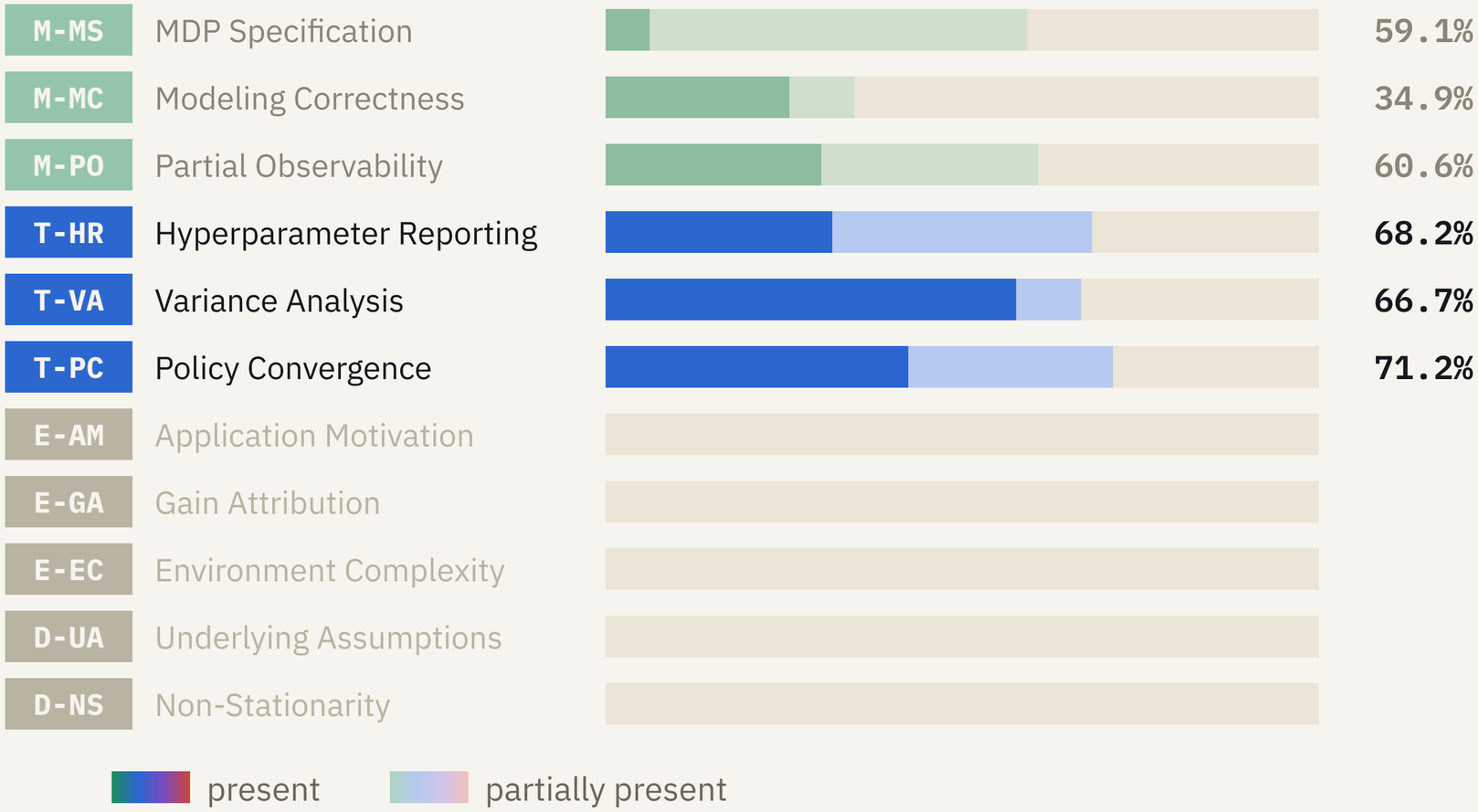
Pitfall Prevalence

11

pitfalls identified

66

papers reviewed



Training Agents

Training Agents

T-HR

Hyperparameter Reporting

The hyperparameters governing training and agent architecture of an approach are not reported.

68%

Training Agents

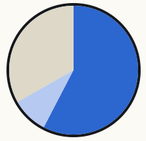
T-HR

Hyperparameter Reporting

The hyperparameters governing training and agent architecture of an approach are not reported.

68%

T-VA



Variance Analysis

The inherent variance of DRL training is not investigated through explicit analysis of performance over multiple policies.

67%

Training Agents

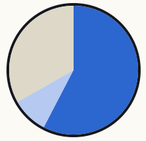
T-HR

Hyperparameter Reporting

The hyperparameters governing training and agent architecture of an approach are not reported.

68%

T-VA

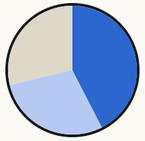


Variance Analysis

The inherent variance of DRL training is not investigated through explicit analysis of performance over multiple policies.

67%

T-PC



Policy Convergence

The convergence of the policy is not demonstrated prior to the termination of training.

71%

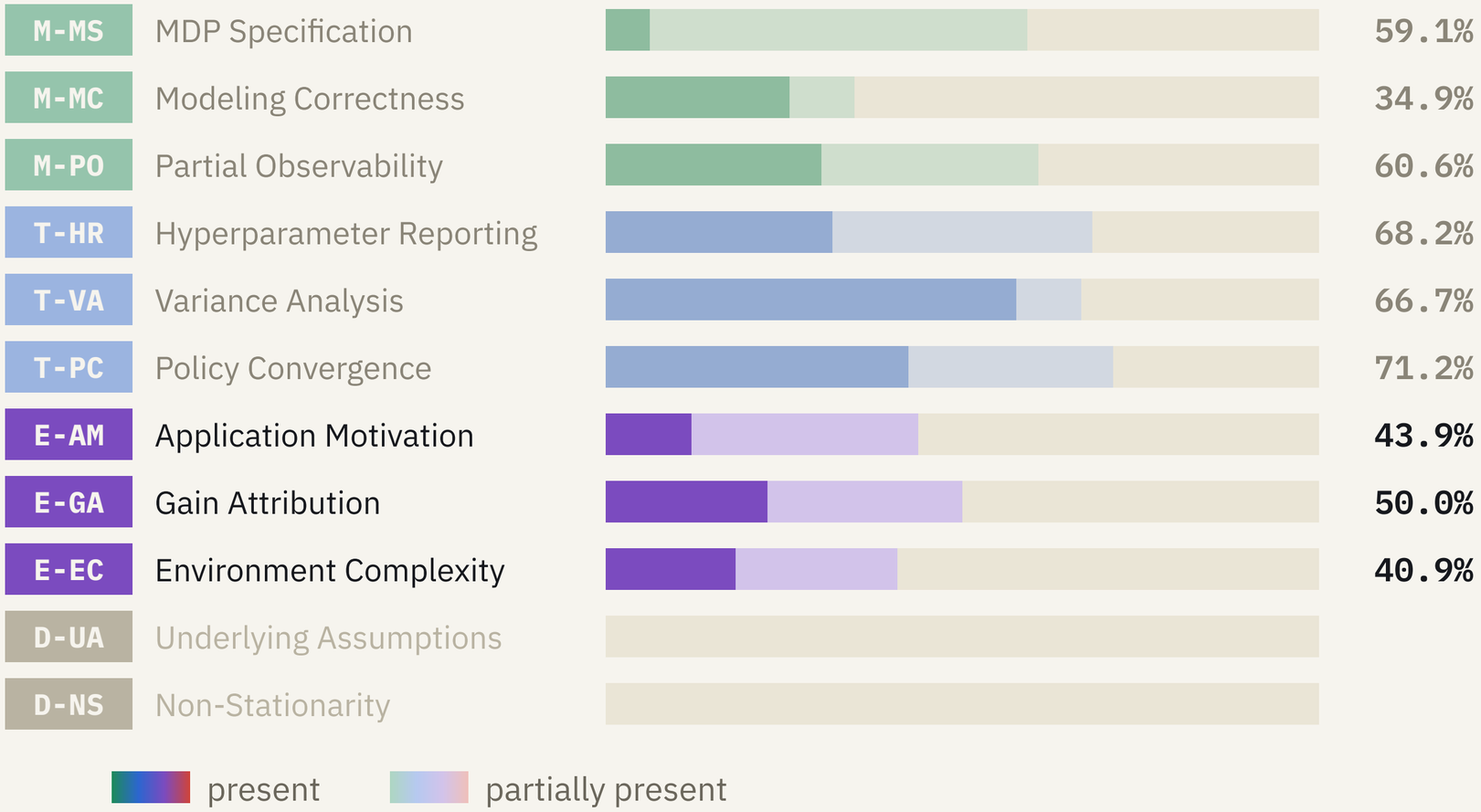
Pitfall Prevalence

11

pitfalls identified

66

papers reviewed



Evaluating Agents

Evaluating Agents

E-AM

Application Motivation

The application of DRL over traditional methods is not justified empirically or theoretically.

44%

Evaluating Agents

E-AM

Application Motivation

The application of DRL over traditional methods is not justified empirically or theoretically.

44%

E-GA

Gain Attribution

Improvements stem from additional information or capabilities provided in MDP design rather than the learned policy itself.

50%

Evaluating Agents

E-AM

Application Motivation

The application of DRL over traditional methods is not justified empirically or theoretically.

44%

E-GA

Gain Attribution

Improvements stem from additional information or capabilities provided in MDP design rather than the learned policy itself.

50%

E-EC

Environment Complexity

Training and evaluation occur in an environment that is contrived or overly simplified.

41%

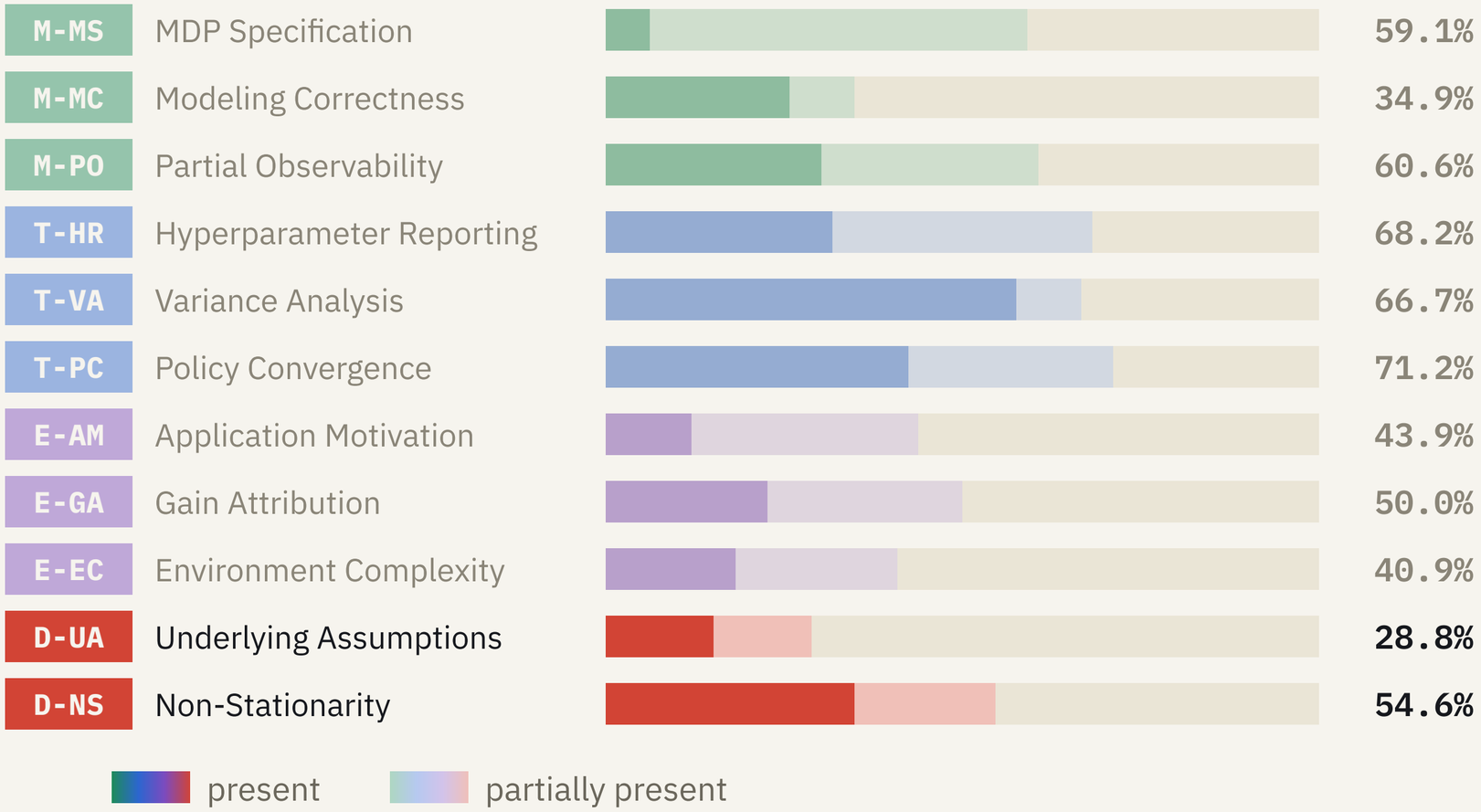
Pitfall Prevalence

11

pitfalls identified

66

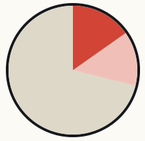
papers reviewed



Deploying Agents

Deploying Agents

D-UA



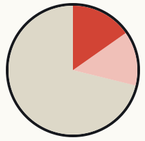
Underlying Assumptions

The environment or agent training requirements cannot be satisfied in real-world applications.

29%

Deploying Agents

D-UA

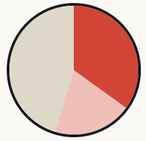


Underlying Assumptions

The environment or agent training requirements cannot be satisfied in real-world applications.

29%

D-NS



Non-Stationarity

The environment does not effectively mitigate the inherent non-stationarity of the underlying cybersecurity task.

55%

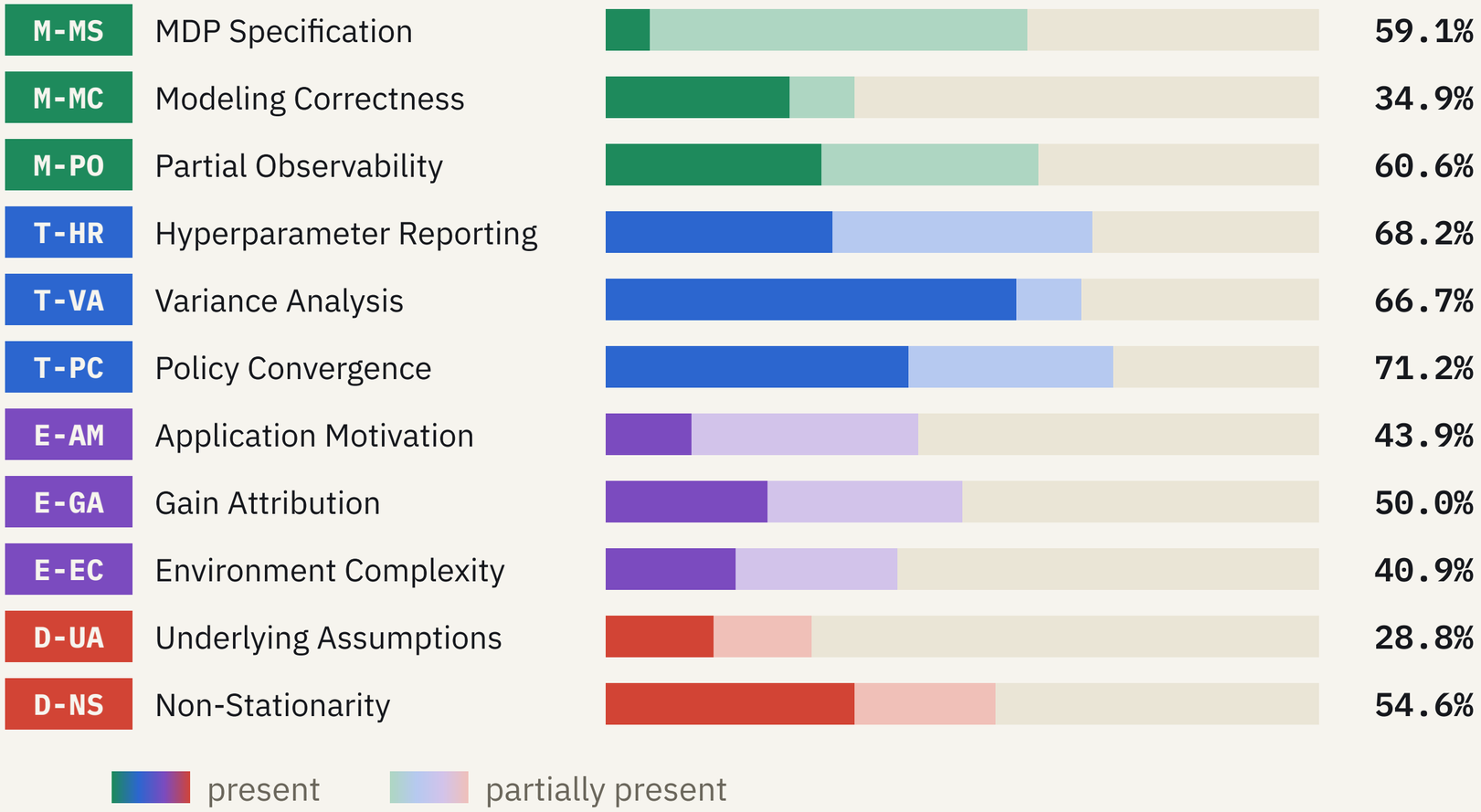
Pitfall Prevalence

11

pitfalls identified

66

papers reviewed



Pitfall Prevalence

11

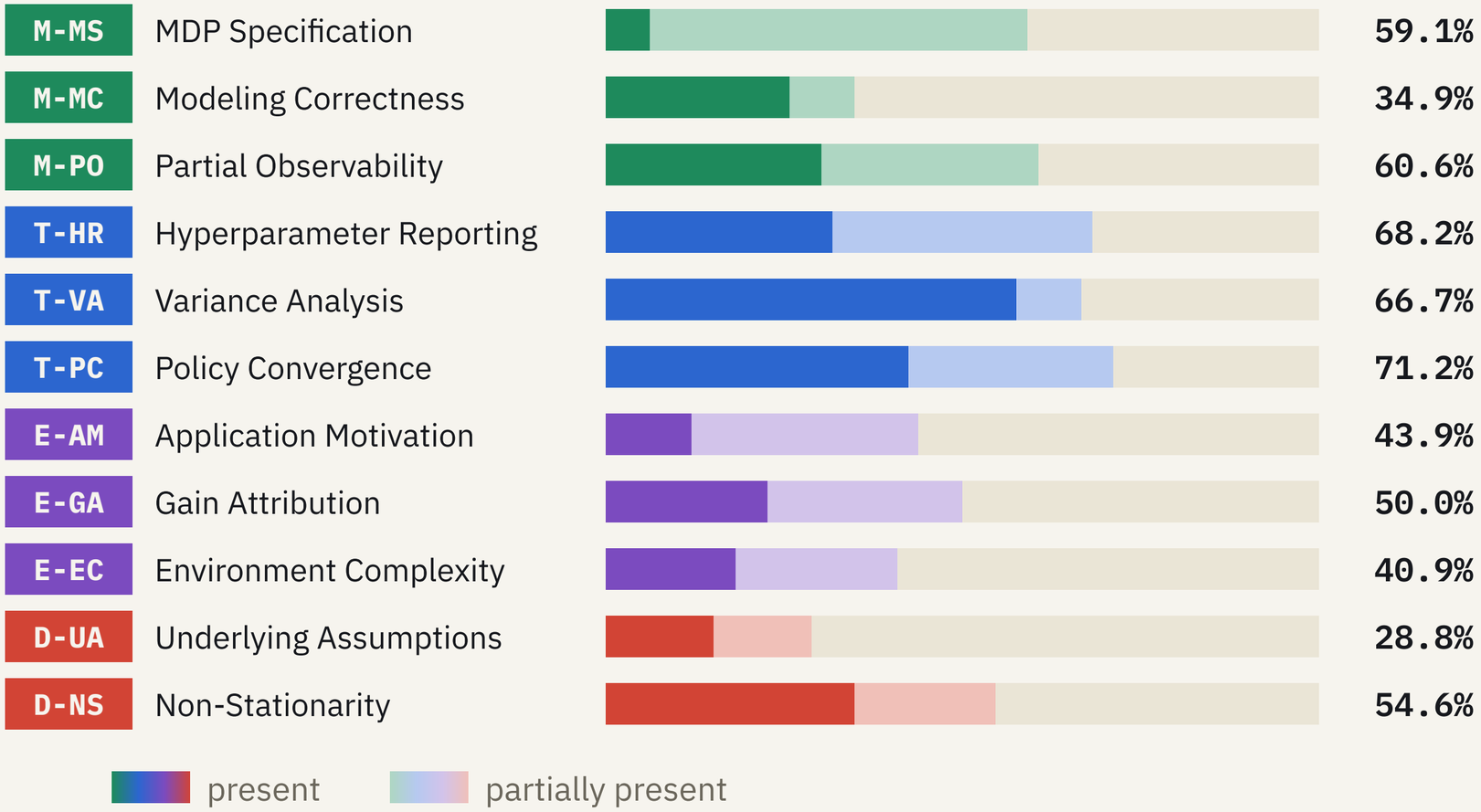
pitfalls identified

66

papers reviewed

≥ 2

pitfalls in every paper



Pitfall Prevalence

11

pitfalls identified

66

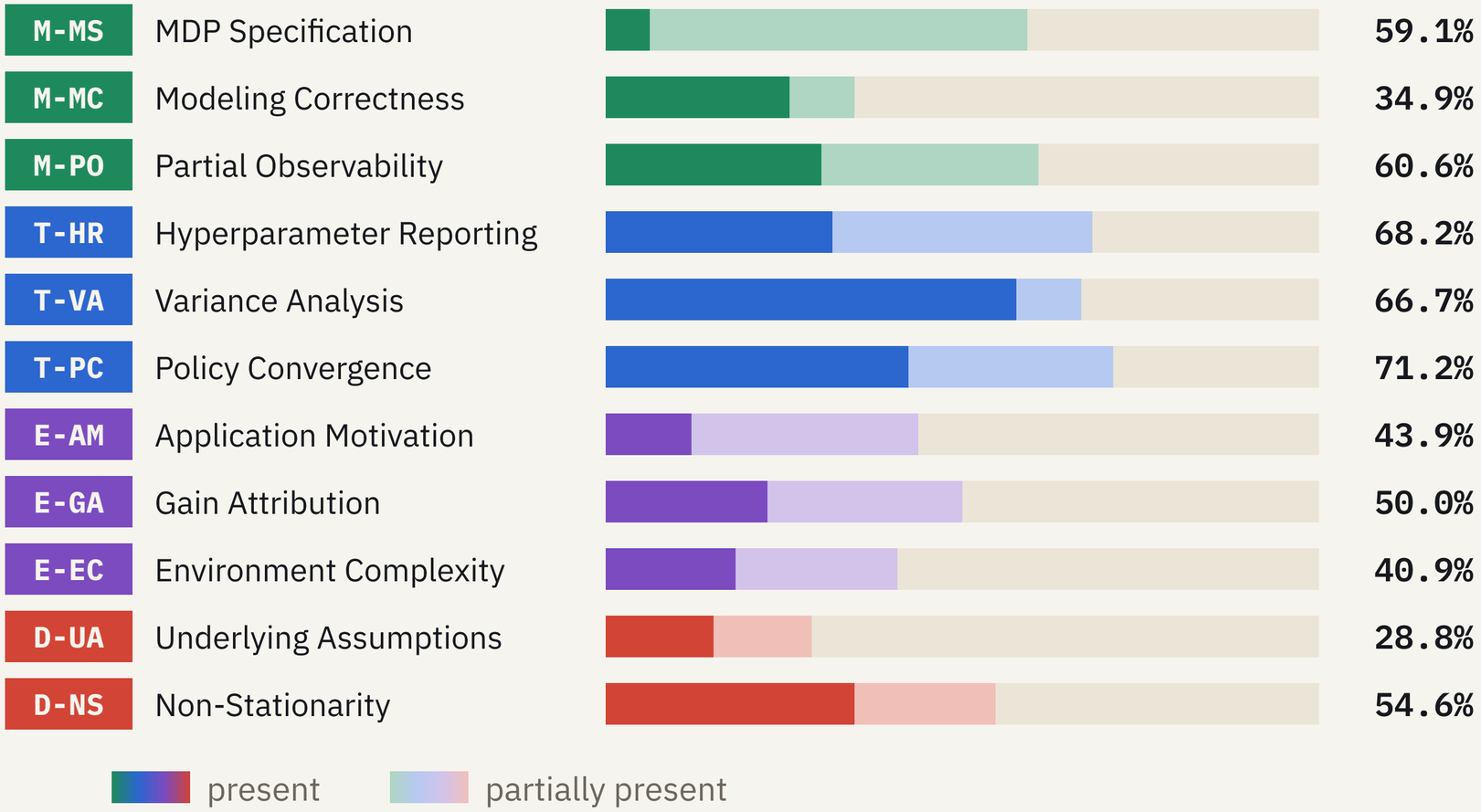
papers reviewed

≥ 2

pitfalls in every paper

5.8

average pitfalls per paper

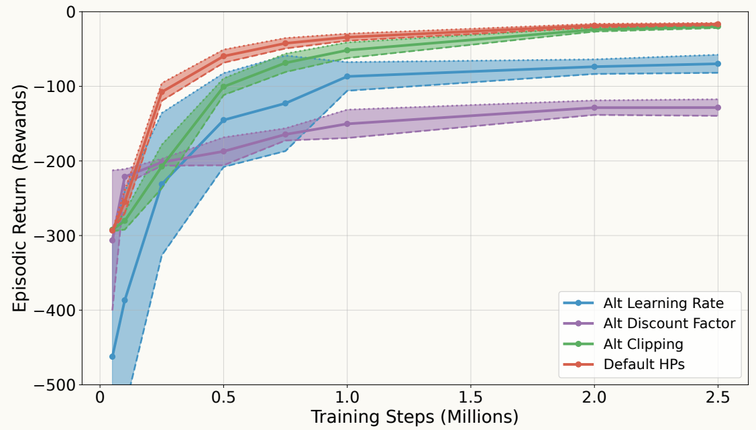


Case Study on Autonomous Cyber Defence

Case Study on Autonomous Cyber Defence

Training Pitfalls

MiniCAGE replicates the CAGE 2 autonomous cyber defense challenge: a blue agent defends a 13 host enterprise network.



T-HR A single hyperparameter can have significant impacts on performance.

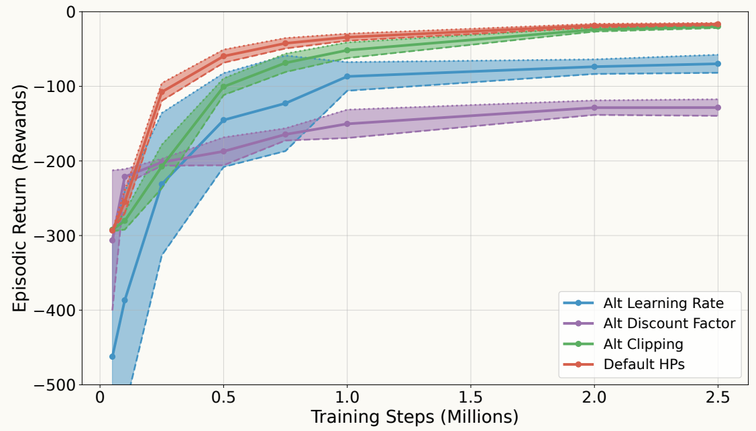
T-VA Single runs can greatly over-estimate performance.

T-PC Convergence should be demonstrated, e.g., through plots.

Case Study on Autonomous Cyber Defence

Training Pitfalls

MiniCAGE replicates the CAGE 2 autonomous cyber defense challenge: a blue agent defends a 13 host enterprise network.



T-HR A single hyperparameter can have significant impacts on performance.

T-VA Single runs can greatly over-estimate performance.

T-PC Convergence should be demonstrated, e.g., through plots.

Deployment Pitfalls

The blue agent is trained against one attacker strategy or a mix, then evaluated against each.

TRAIN \ EVAL	B-LINE	MEANDER	MIXED
B-line	**-19.0**	**-104.1**	**-61.4**
Meander	**-79.0**	**-61.0**	**-70.2**
Mixed	**-40.2**	**-79.2**	**-60.2**

B-line Heads directly to the operational server to compromise it.

Meander Compromises every host along the way.

D-NS Training on a fixed adversary collapses against unseen attackers.

USENIX SECURITY '26

The Pitfalls of Deep Reinforcement Learning for Cybersecurity

Shae McFadden, Myles Foley, Elizabeth Bates, Ilias Tsingenopoulos,
Sanyam Vyas, Vasilios Mavroudis, Chris Hicks, Fabio Pierazzi

PAPER WEBSITE

"These are not failures of individual researchers but systematic challenges that emerge when adapting DRL to cybersecurity."

The goal is not to discourage DRL in security. It is to establish methodological standards that distinguish robust, deployable solutions from those that succeed only in idealized settings.

Contact: shae.mcfadden@kcl.ac.uk