We present the first explicit-state method for analysing and ensuring safety of DRL agents for Atari games.

- We propose 42 safety properties for 31 games.
- We evaluate the safety of available Deep Reinforcement Learning (DRL) agents.
- We improve safety through shielding using bounded explicit-state exploration.

We consider 31 Atari games with unique dynamics given by a black-box emulator.
- Each game is a deterministic MDP \((S, A, T, R)\).
- "no-op" non-determinism added: no inputs for the first \(k \in \{0, \ldots, 30\}\) frames.
- Learn deterministic policy \(\pi : S \rightarrow A\) through SOTA DRL methods.

Safety property \(\phi \subseteq S\) is set of safe states.
- Satisfied if for all reachable states \(s \in \phi\).
- Labelling handcrafted from graphical output.

Example Properties
<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assault:overheat</td>
<td>Die from overheating</td>
</tr>
<tr>
<td>Bowling:no-hit</td>
<td>Miss all pins</td>
</tr>
<tr>
<td>Freeway:hit</td>
<td>Get hit by a car</td>
</tr>
</tbody>
</table>

To verify \(\phi\) for policy \(\pi\) run games with \(\pi\) for all values of \(k\).
- This will visit every reachable state, \(\phi\) true iff. for all states visited \(s \in \phi\).

Shielding Atari Games with Bounded Prescience
Mirco Giacobbe, Mohammadhosein Hasanbeig, Daniel Kroening, Hjalmar Wijk
Computer Science Department, University of Oxford, UK

"Explicit-state verification demonstrates that DRL algorithms do not learn to satisfy even simple Safety Properties."

Effect of shielding on the average safety achieved.
- With BPS all agents satisfy all shallow properties.

Properties grouped by number of satisfying agents before (w/o dots) and after BPS (with dots).
- Minimal properties are satisfied by random agents, shallow properties require no planning.
- No non-minimal property is satisfied by more than 4 agents.

References