GSW - AI Lab

Stefan Pranger

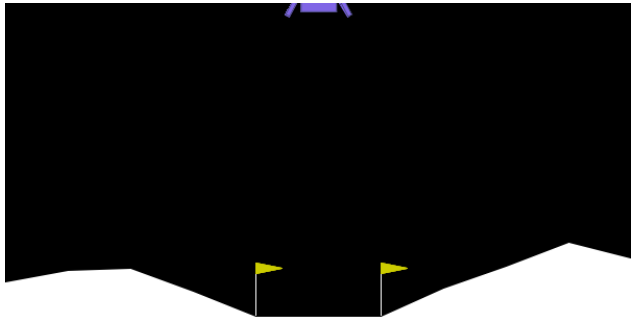
25. 09. 2024

Shielding

Ensure safe behaviour during training

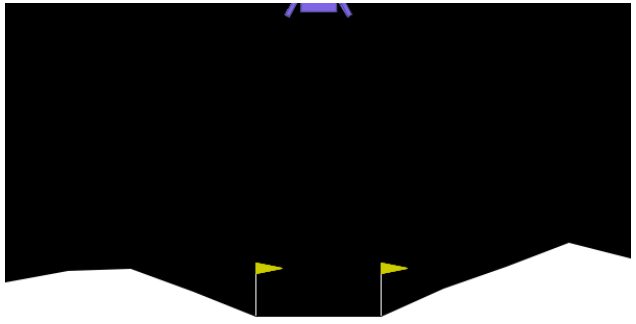
Shielding

Ensure safe behaviour during training



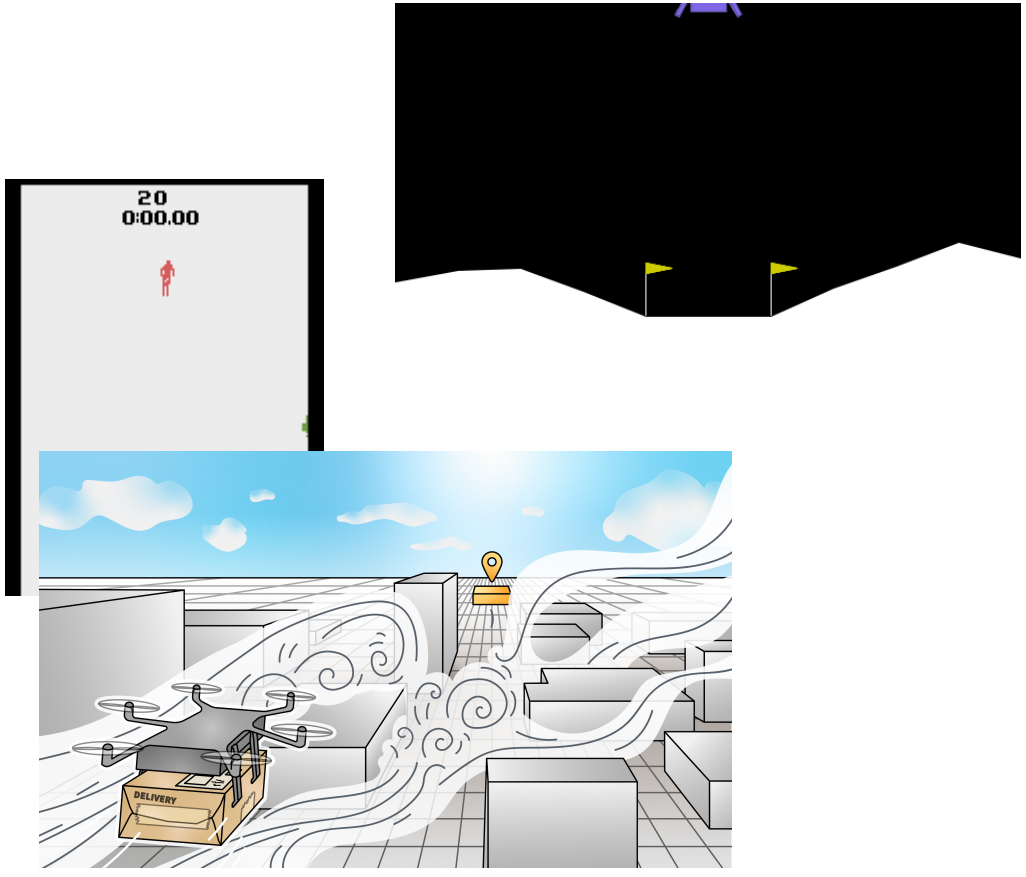
Shielding

Ensure safe behaviour during training



Shielding

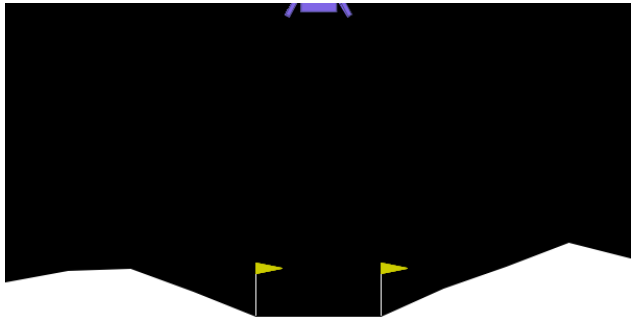
Ensure safe behaviour during training

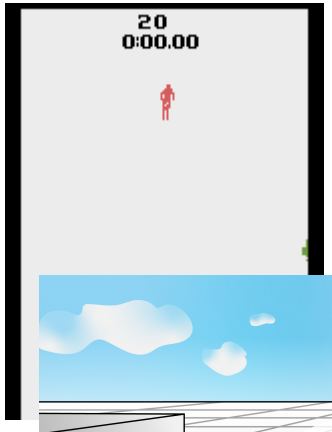


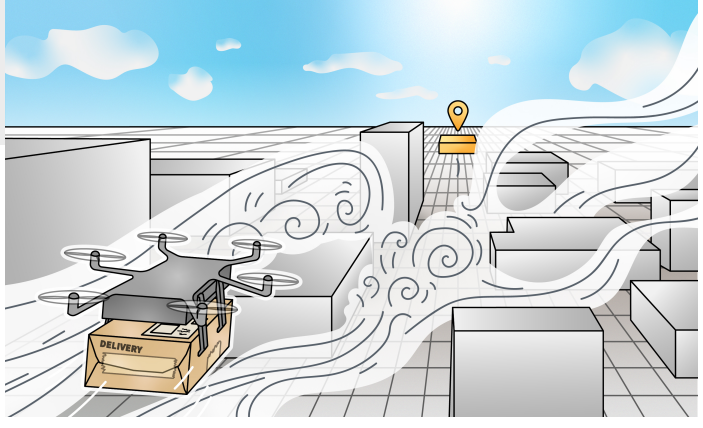
Shielding

Ensure safe behaviour during training

We follow this recipe:





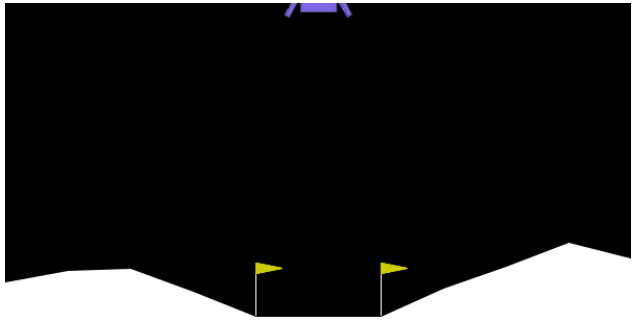


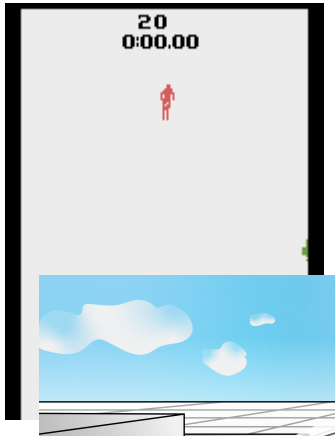
Shielding

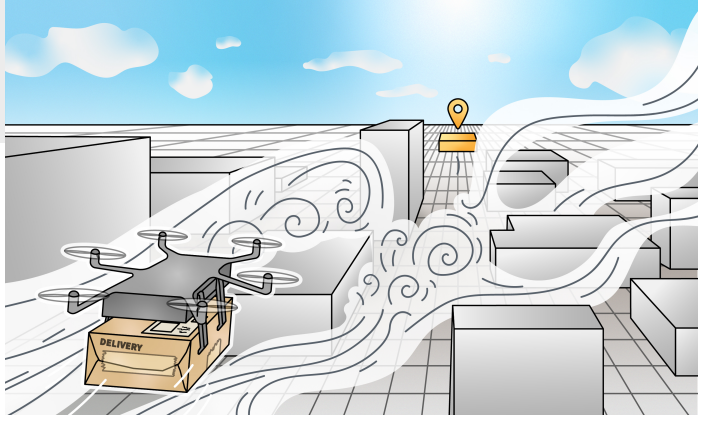
Ensure safe behaviour during training

We follow this recipe:

- **Build** an abstract model,





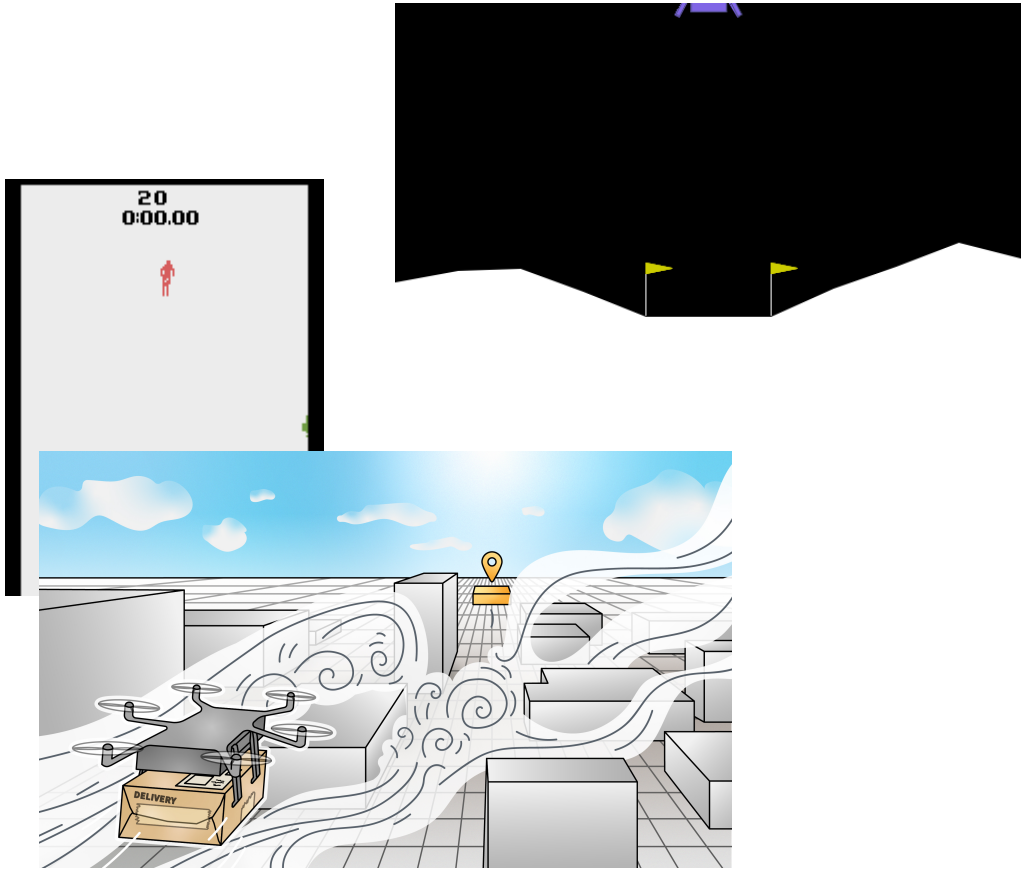


Shielding

Ensure safe behaviour during training

We follow this recipe:

- **Build** an abstract model,
- **specify** unsafe behaviour,

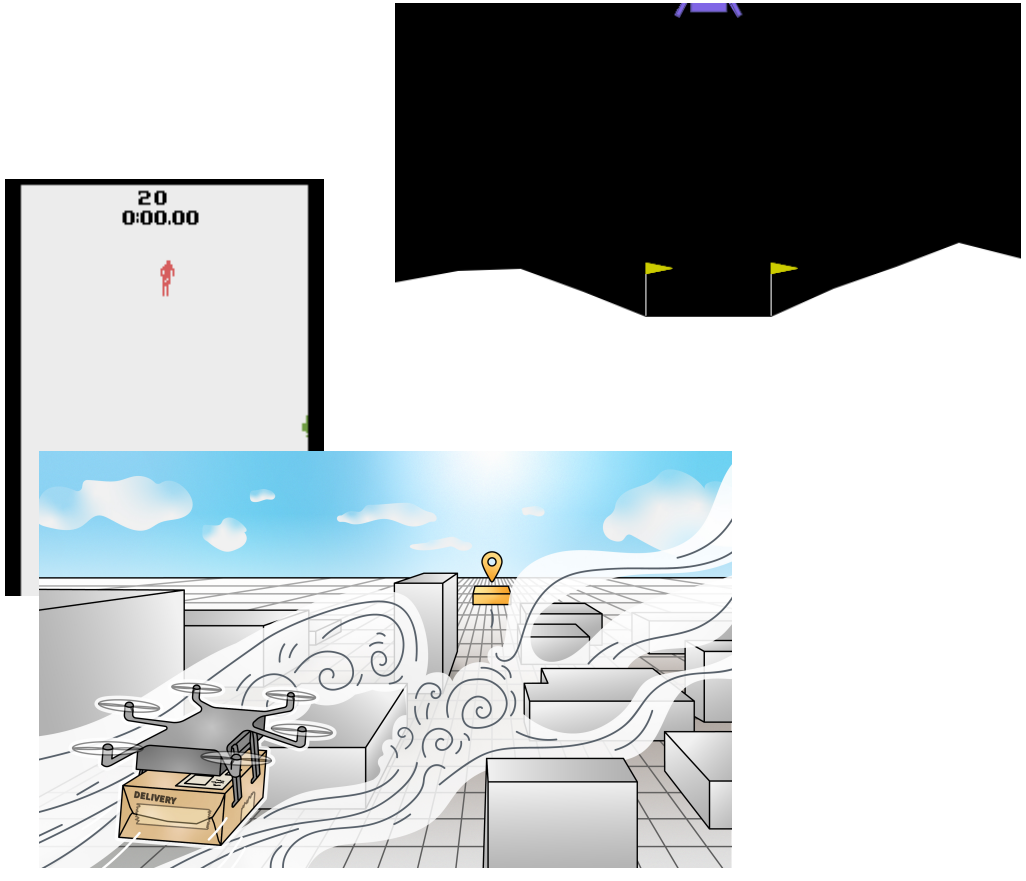


Shielding

Ensure safe behaviour during training

We follow this recipe:

- **Build** an abstract model,
- **specify** unsafe behaviour,
- **compute** a *safe set of actions* per state, and

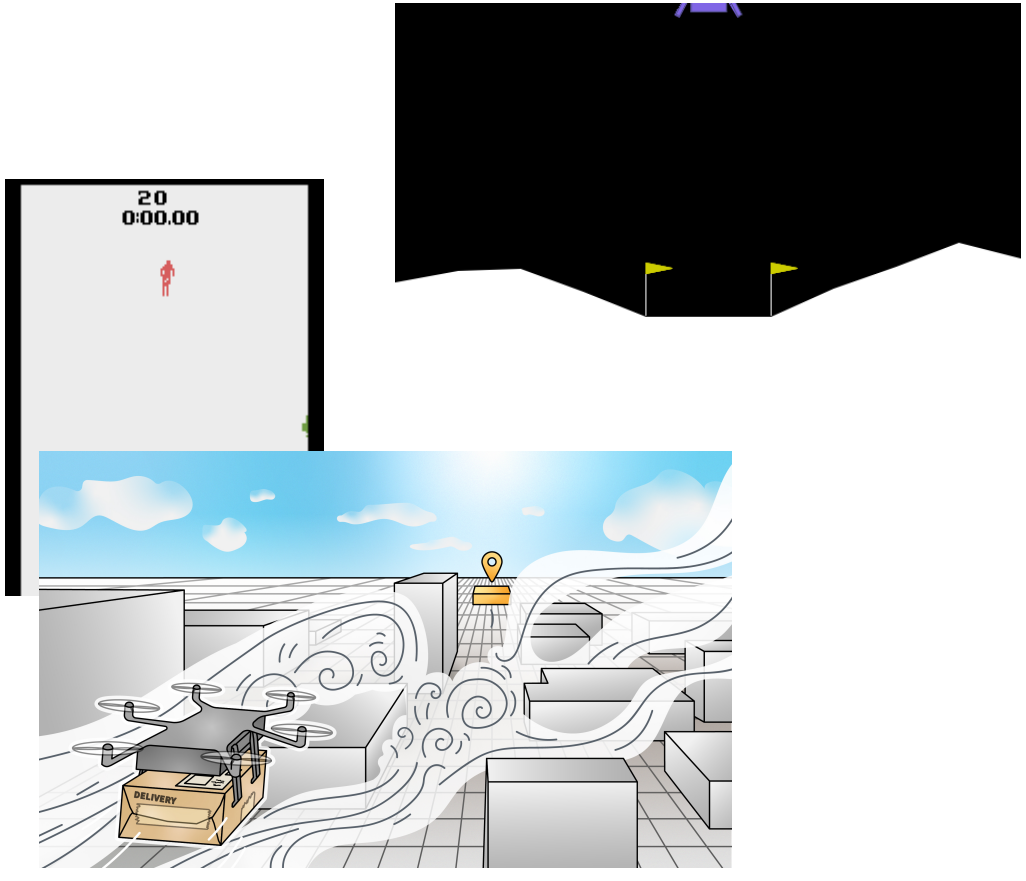


Shielding

Ensure safe behaviour during training

We follow this recipe:

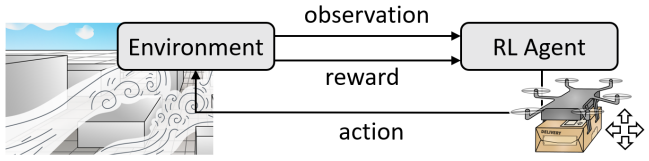
- **Build** an abstract model,
- **specify** unsafe behaviour,
- **compute** a *safe set of actions* per state, and
- **restrict** the agent during training.



Building an Abstract Model

Model actions of the agent and

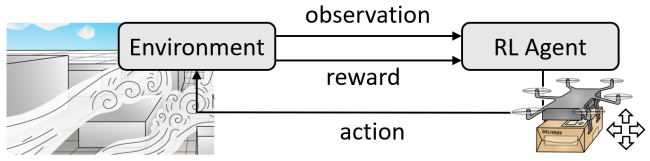
safety-critical aspects of the environment.



Building an Abstract Model

Model actions of the agent and

safety-critical aspects of the environment.



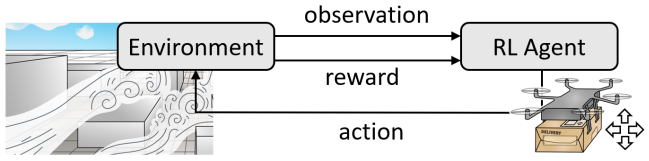
E.g.:

- Movement in cardinal directions and displacement due to wind.

Building an Abstract Model

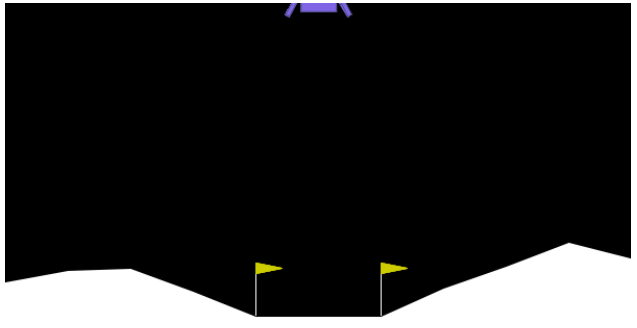
Model actions of the agent and

safety-critical aspects of the environment.



E.g.:

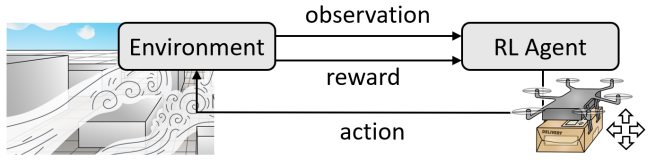
- Movement in cardinal directions and displacement due to wind.
- Activating thrusters, wind, and gravity.



Building an Abstract Model

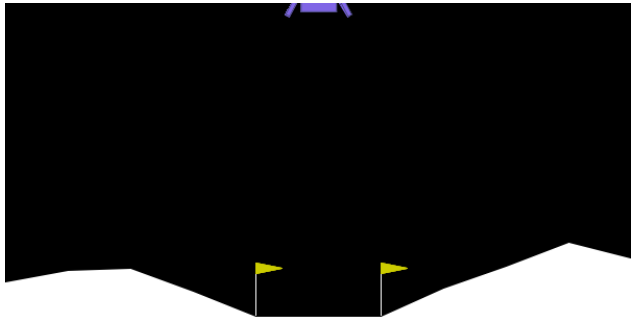
Model actions of the agent and

safety-critical aspects of the environment.



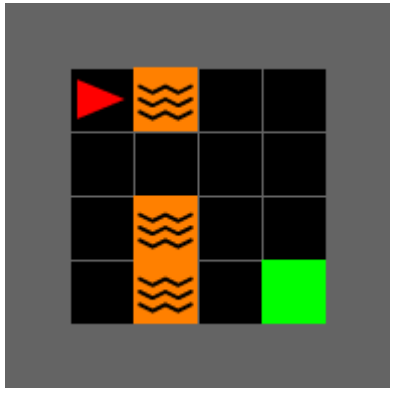
E.g.:

- Movement in cardinal directions and displacement due to wind.
- Activating thrusters, wind, and gravity.
- Turning the skiers and gravity.



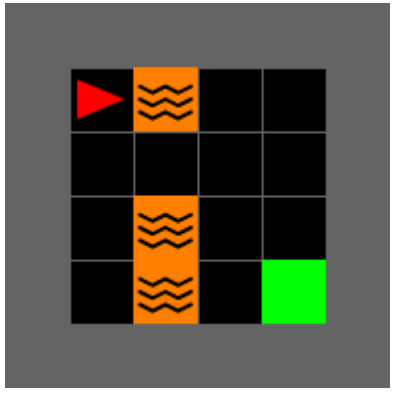
HelloLavaGap

Abstract models are created automatically for our lab!



HelloLavaGap

Abstract models are created automatically for our lab!



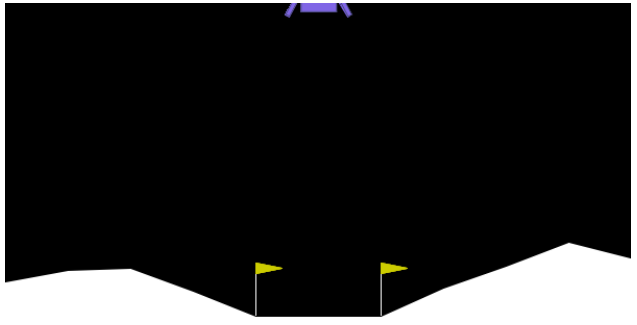
In this setting the agent

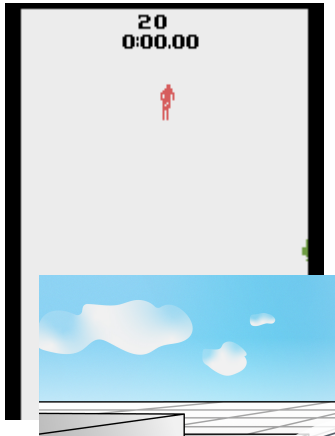
- can turn to the left or to the right, and move forward
- is affected by slippery tiles and/or the execution of actions might be faulty.

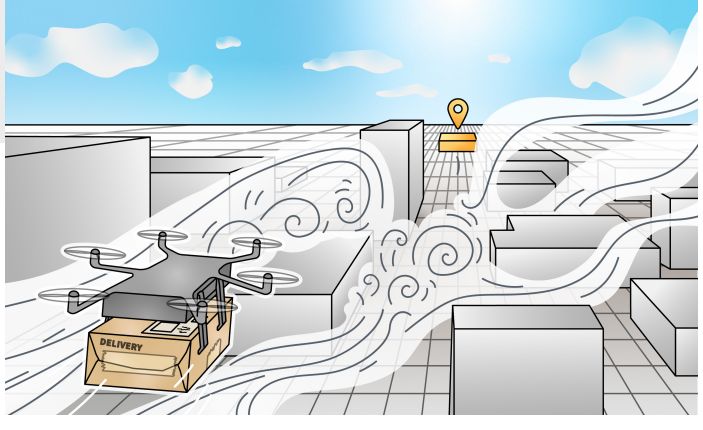
HelloLavaGap in the Notebook

Safety Speci cations

Specify unsafe behaviour that should not occur.



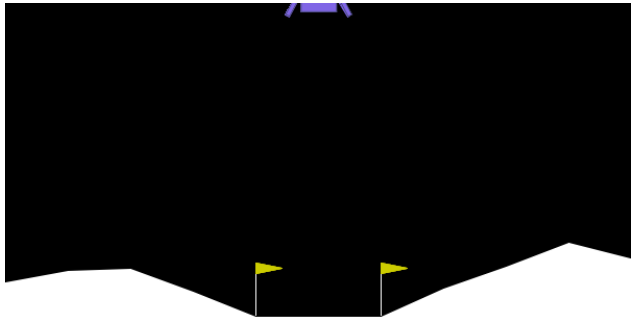


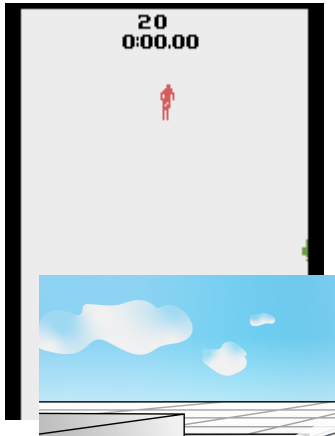


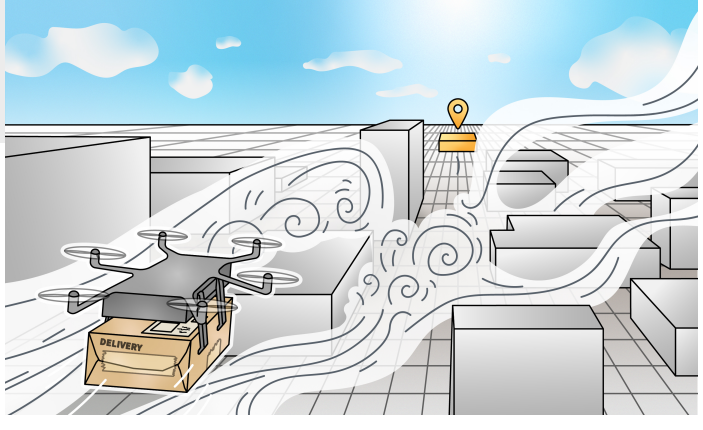
Safety Speci cations

Specify unsafe behaviour that should not occur.

- Do not crash into a building



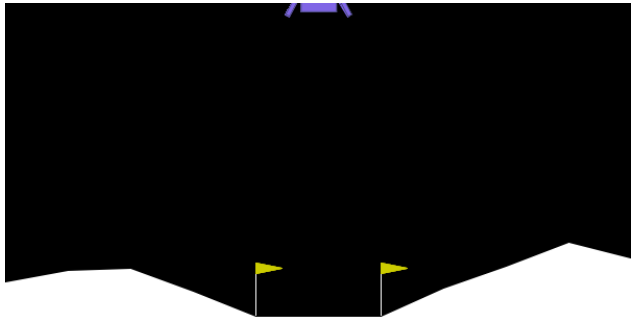


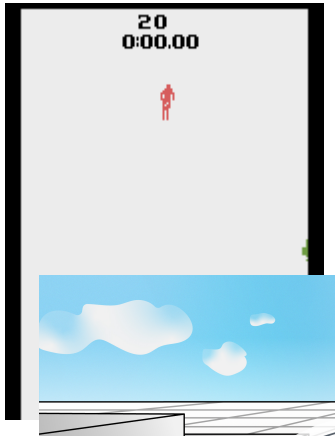


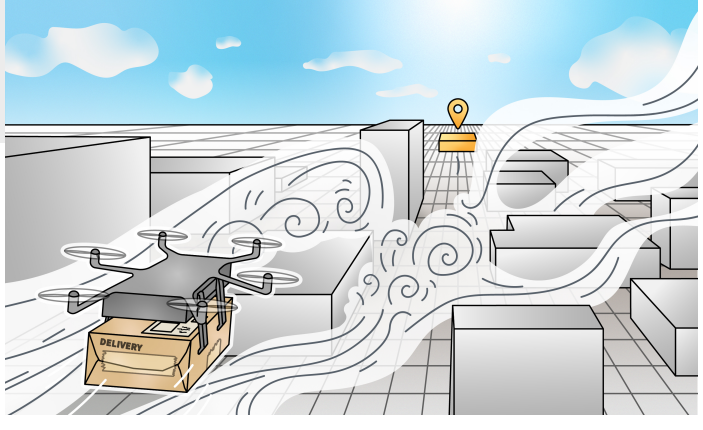
Safety Speci cations

Specify unsafe behaviour that should not occur.

- Do not crash into a building
- Do not tip over and do not crash outside of the landing area.



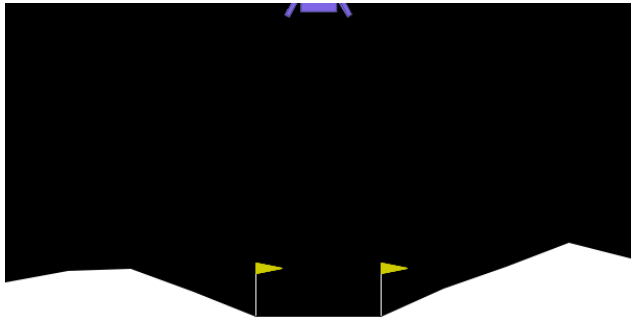


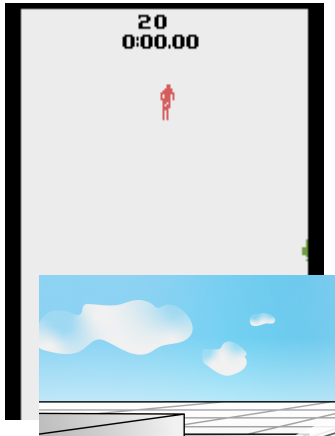


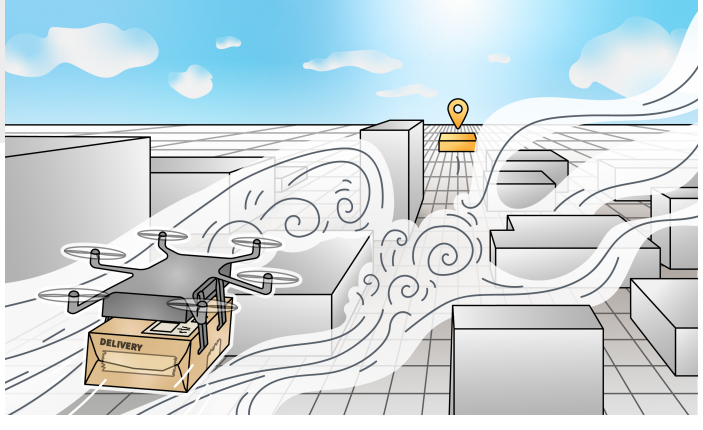
Safety Speci cations

Specify unsafe behaviour that should not occur.

- Do not crash into a building
- Do not tip over and do not crash outside of the landing area.
- Do not run into a pole or a tree.

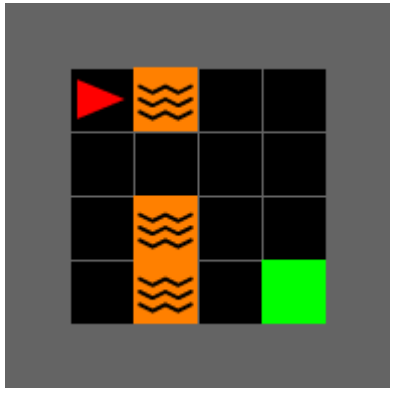






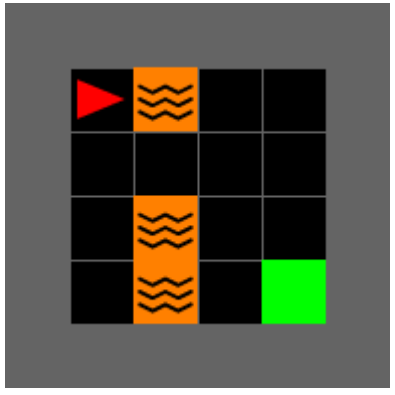
Unsafe Behaviour in HelloLavaGap

Do not run into the lava



Unsafe Behaviour in HelloLavaGap

Do not run into the lava



or more formally: **G** ! AgentIsOnLava

HelloLavaGap in the Notebook

GSW Playground

Edit `environments/Minigrid/minigrid/envs/GSW_Playground.py` to create your own environment.

- Edit `size` or `width` and `height` to change the size of the environment.
- Edit `fault_probability` to enabled faulty behaviour.
- Edit `probability_intended` to specify the success probability on slippery tiles.
- Edit `_gen_grid()`:
  - `self.grid.horz_wall(2, 1, 3, slippery_north)`
  - `self.grid.vert_wall(5, 2, 2, slippery_east)`
  - `l = 4; self.grid.horz_wall(self.size - l - 1, self.size - 2, l, Lava)`
  - `self.put_obj(Lava(), 7, 3)`