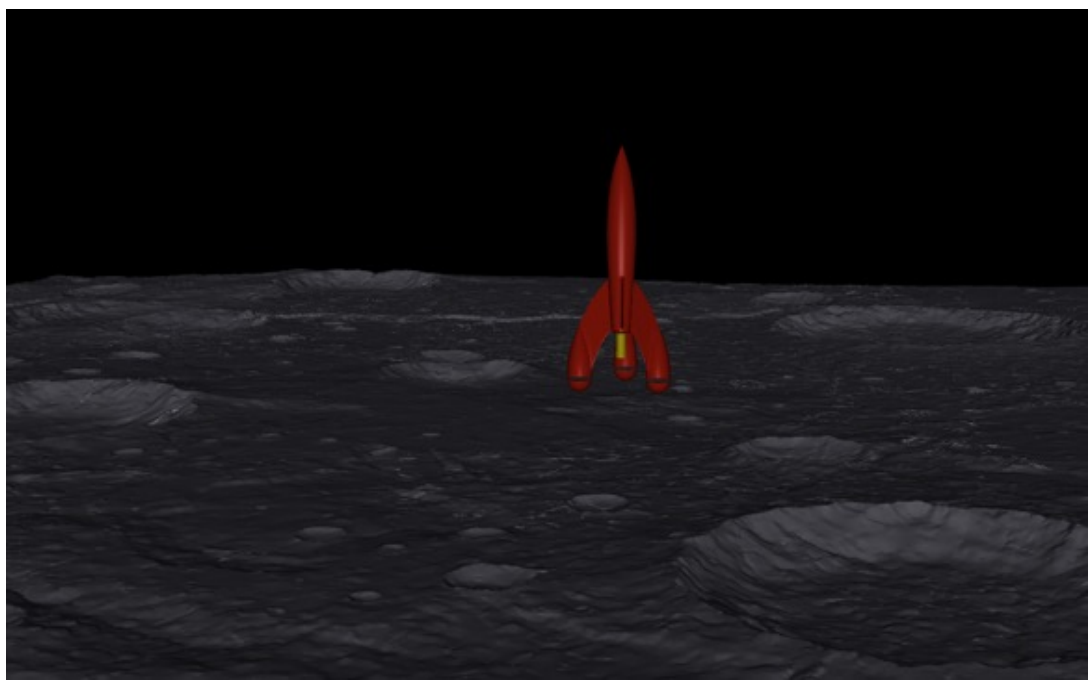
Agent

Action = $\begin{bmatrix} T \\ \theta_p \\ \theta_y \end{bmatrix}$

Rewards

Observation = $\begin{bmatrix} x_t \\ \dot{x}_t \\ m \\ x_{target} \end{bmatrix}$



Environment