Algorithm 1 Simplified Vanilla Policy Gradient Algorithm

Initialize the policy π and the value function V.
for $k = 0, 1, 2, ...$ **do**
    Collect set of trajectories $D_k = \{\tau_i\}$ by running policy π_k in the environment.
    Compute future returns R_t^f for each time step t in each trajectory τ_i.
    Compute the policy gradient $\hat{g}_k$ using D_k and V. ▷ More on $\hat{g}_k$ soon!
    Update the policy π with $\hat{g}_k$ using SGD, Adam, etc.
    Fit value function V by regression on MSE between V and R_t^f using SGD, Adam, etc.
end for