Algorithm 1 Simplified Vanilla Policy Gradient Algorithm

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