Algorithm: MADDPG

Initialize: θ and θ^- the parameters of individual policy and target policy μ_i, ϕ and ϕ^- the parameters central Q-function Q.
Initialize replay buffer $\mathcal{D}$ // $(episode_1, episode_2, \dots)$
while $t < T$ **do**
    $current\_episode = \{\}$
    while $\mathbf{o}_t$ is not *done* **do**
        Collect observations $\{o_1^t, \dots, o_n^t\}$ and state $\mathbf{s}_t$
        Select an action $a_i^t = \mu(o_i^t)$ for each agent i.
        Execute the joint action $\mathbf{a}^t = (a_1^t, \dots, a_n^t)$
        Collect r^t, $done^t$
        Store $(\mathbf{s}_t, \mathbf{o}^t, \mathbf{a}^t, r^t, done^t)$ in *current_episode*
    end while
    Store *current_episode* in the replay buffer $\mathcal{D}$
    if t is a training step **then**
        Sample batch of episodes

$$\mathcal{B} = \{\{\mathbf{s}^{t,b}, \mathbf{o}^{t,b}, \mathbf{a}^{t,b}, r^{t,b}, done^{t,b}, \mathbf{s'}^{t,b}, \mathbf{o'}^{t,b}\}_{t=1\dots L^b}\}_{b=1,\dots,|\mathcal{B}|}$$

        Set the targets

$$y^{t,b} = r^{t,b} + (1 - done^{t,b}) Q(\mathbf{s'^{t,b}}, \mu(o_1'^{t,b}; \theta^-), \dots, \mu(o_n'^{t,b}; \theta^-); \phi^-)$$

        Update ϕ using :

$$\mathcal{L}(\phi) = \frac{1}{|\mathcal{B}|} \sum_b \frac{1}{L^b} \sum_t \left(y^{t,b} - Q(\mathbf{s}^{t,b}, a_1^{t,b}, \dots, a_n^{t,b}; \phi) \right)^2$$

        Update θ using:

$$\mathcal{L}(\theta) = -\frac{1}{|\mathcal{B}|} \sum_b \frac{1}{L^b} \sum_t \sum_i Q(\mathbf{s}^{t,b}, a_1^{t,b}, \dots, \mu(o_i^{t,b}; \theta), \dots, a_n^{t,b}; \phi)$$

        Every C steps, update $\theta^- \leftarrow \theta$, $\phi^- \leftarrow \phi$
    end if
end while