Question 1

1) $G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + \gamma^3 R_4$

$R_1 : S_1 \to S_2 \quad -1$

$R_2 : S_2 \to S_1 \quad -1 \qquad \gamma = 0.9$

$R_3 : S_1 \to S_2 \quad -1$

$R_4 : S_2 \to S_{Term} \quad 10$

$\therefore G_0 = -1 + 0.9(-1) + 0.81(-1) + 10(0.729)$

$= \underline{4.58}$

2) $V_\pi(S_2) = 0.5(-1 + 0.9\, V_\pi(S_1)) + 0.5(10)$

$= \underline{4.5 + 0.45\, V_\pi(S_1)}$

Question 2:

Agent will learn to suck dust and spit it out again and it will keep repeating this. Room keeps getting dirtier but agent gets infinite +1s. So what we can do instead is we can reward based on final room cleanliness.

Question 3: PART A

We mathematically need $\gamma < 1$ for infinite horizon tasks to ensure that the value function converges. If the task never ends with rewards always +1 and $\gamma = 1$.

$$V_\pi(s) = \sum_{t=0}^{\infty} 1^t \times 1 = \infty$$

PART B

$\gamma = 0$ (Impulsive Agent): This agent only cares about immediate rewards since future rewards are discounted to 0.

$\gamma = 0.99$ (Strategic Agent): The agent values future rewards as much as immediate ones. It makes long term strategic decisions.

Question 4:

Yes, the optimal policy changes.
In the original setup with $R = -1$, the return is $G = -n$ where n is the number of steps.
The agent maximises return by minimizing steps.

with the modified reward $R_{new} = +1$, return becomes $G = +n$. Now; the agent maximizes return by maximizing steps. The new optimal policy will take the longest possible path.

Question 5:

$$v_\pi(s) = E_\pi[G_t \mid S_t = s]$$

$$G_t = R_{t+1} + \gamma G_{t+1}$$

$$v_\pi(s) = E_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

$$v_\pi(s) = E_\pi[R_{t+1} \mid S_t = s] + \gamma E_\pi[G_{t+1} \mid S_t = s]$$

$$v_\pi(s) = \sum_a \pi(a|s) E[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a]$$

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) [r + \gamma E[G_{t+1} \mid S_{t+1} = s']]$$

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')]$$

Question 6:

1)
$$v_\pi = r_\pi + \gamma P^\pi v_\pi$$

$$v_\pi = (I - \gamma P^\pi)^{-1} r_\pi$$

2)
$$O(N^3) = O((10^{20})^3) = O(10^{60}).$$

Supercomputer speed: 10^{18} FLOPs/second

Time required $= \frac{10^{60}}{10^{18}} = 10^{42}$ seconds

$\approx 10^{34}$ years

3) This is why model free methods such as Monte-Carlo are important; They are tracable for realistic problem sizes.

Question 7:

1)
$$\pi'(s) = \arg\max_a \sum_{s',r} p(s',r|s,a)[r + \gamma v^*(s')]$$

2)
$$\pi'(s) = \arg\max_a q^*(s,a).$$

3) In model-free environments, we don't know $p(s',r|s,a)$ beforehand. Using only $v^*(s)$, we cannot construct an optimal policy because we cannot compute which action leads to the best next states without the transition model. This is why Monte Carlo ~~[illegible]~~ uses action value functions - they enable optimal decision-making in model-free settings.