1) State Space: {warm, cool, Off}
   Actions: {fast, slow}
   Transition Model: {
      P(cool | cool, slow) = 1
      P(warm | cool, slow) = 0
      P(cool | cool, fast) = ¼
      P(warm | cool, fast) = ¾
      P(cool | warm, slow) = ¼
      P(warm | warm, slow) = ¾
      P(cool | warm, fast) = 0
      P(warm | warm, fast) = 7/8
      P(off | warm, fast) = 1/8
      P(off | off, slow) = 1
      P(off | off, fast) = 1
   }
   Reward Function: {
      off->any action->off = 0
      warm or cold->slow->warm or cold = 4
      warm or cold -> fast -> warm or cold or off = 10
   }

2) With the policy of always going fast:
   U(off) = 0
   U(warm) = 7/8*(10 + U(warm)) + 1/8*(10 + U(off))
   U(cool) = ¼*(10 + U(cool)) + ¾*(10 + U(warm))
   Solving gives U(warm) = 80, U(cool) = 93.33

   With the policy of always going slow:
   U(off) = 0
   U(warm) = ¾*(4 + U(warm)) + ¼*(4+U(cool))
   U(cool) = 1 * (4 + U(cool))
   This solution is undefined because the U(cool) is actually infinite. If we drive slowly then we will stay cool forever and continue to get 4 reward points

3) First policy eval. I just added a discount factor to the first policy above and resolved the system. U(off) = 0, U(warm) = 47.06, U(cool) = 53.89
Because the order of states to iterate doesn't matter, I'll do cool then warm then off.

   State cool, action slow -> 1*53.89 = 53.89
   State cool, action fast -> ¼*(53.89) + ¾*(47.06) = 48.76
   Argmax(actions) means update policy cool to be slow,

   State warm, action slow ->3/4*(47.06) + ¼*(53.89) = 48.77
   Sate warm, action fast -> 7/8*(47.06) + 1/8*0 = 41.18
   Argmax(actions) means update policy warm to be slow

State off, action slow-> 0
Sate off, action fast-> 0
Argmax keeps it the same, when off try to go fast

Next U = policyeval
U(off) = 0
U(warm) = ¾*(4+.9*U(warm)) + ¼*(4 + .9*U(cold))
U(cool) = 1*(4+.9*U(cool))
Solving gives U(cool) = 40, U(warm) = 40, U(off) = 0

Now iterate over actions

State cool, action slow -> 1*40 = 40
State cool, action fast -> ¼*40 + 3/4*40 = 40
If argmax breaks ties by keeping the current policy then stay slow on cool

State warm, action slow-> ¾*(40) + ¼*(40) = 40
State warm, action fast -> 7/8*(40) + 1/8*0 = 35
Argmax keeps it the same

State off doesn't really matter, they all end up as zero, policy doesn't matter either.

Policy after second iteration depends on how argmax breaks ties. If it sticks with the current policy then the policy is cool->slow, warm->slow, off->fast