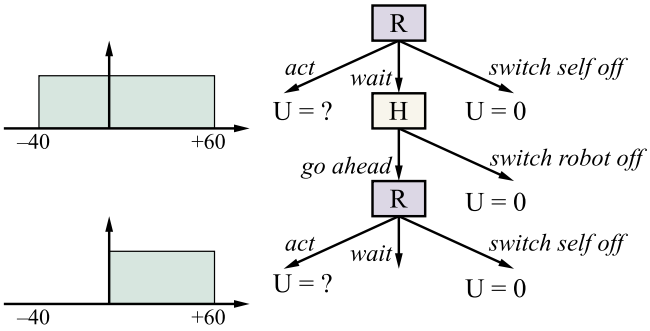

Figure 16.11 The off-switch game. *R*, the robot, can choose to act now, with a highly uncertain payoff; to switch itself off; or to defer to *H*, the human. *H* can switch *R* off or let it go ahead. *R* now has the same choice again. Acting still has an uncertain payoff, but now *R* knows the payoff is nonnegative.

To study this question, let's consider a very simple scenario, as shown in Figure 16.11. Robbie is a software robot working for Harriet, a busy human, as her personal assistant. Harriet needs a hotel room for her next business meeting in Geneva. Robbie can act now—let's say he can book Harriet into a very expensive hotel near the meeting venue. He is quite unsure how much Harriet will like the hotel and its price; let's say he has a uniform probability for its net value to Harriet between -40 and $+60$, with an average of $+10$. He could also "switch himself off"—less melodramatically, take himself out of the hotel booking process altogether—which we define (without loss of generality) to have value 0 to Harriet. If those were his two choices, he would go ahead and book the hotel, incurring a significant risk of making Harriet unhappy. (If the range were -60 to $+40$, with average -10, he would switch himself off instead.) We'll give Robbie a third choice, however: explain his plan, wait, and let Harriet switch him off. Harriet can either switch him off or let him go ahead and book the hotel. What possible good could this do, one might ask, given that he could make both of those choices himself?

The point is that Harriet's choice—to switch Robbie off or let him go ahead—provides Robbie with information about Harriet's preferences. We'll assume, for now, that Harriet is rational, so if Harriet lets Robbie go ahead, it means the value to Harriet is positive. Now, as shown in Figure 16.11, Robbie's belief changes: it is uniform between 0 and $+60$, with an average of $+30$.

So, if we evaluate Robbie's initial choices from his point of view:

1. Acting now and booking the hotel has an expected value of $+10$.
2. Switching himself off has a value of 0.
3. Waiting and letting Harriet switch him off leads to two possible outcomes:
   (a) There is a 40% chance, based on Robbie's uncertainty about Harriet's preferences, that she will hate the plan and will switch Robbie off, with value 0.
   (b) There is a 60% chance Harriet will like the plan and allow Robbie to go ahead, with expected value $+30$.

Thus, waiting has expected value $(0.4 \times 0) + (0.6 \times 30) = +18$, which is better than the $+10$ Robbie expects if he acts now.