using the C4.5 system (Quinlan, 1993a). The decision tree was then converted into C code and inserted into the flight simulator's control loop so that it could fly the plane itself.

The results are surprising: not only does the program learn to fly, it learns to fly somewhat *better* than its teachers. This is because the generalization process cleans up the occasional mistakes made by humans. Such results suggest that machine learning techniques may yield controllers that are more robust than conventional, manually programmed autopilots. For difficult tasks such as flying helicopters carrying heavy loads in high winds, no autopilots are available and very few humans are competent. Such tasks are potentially suitable for systems based on automated learning.

18.4 USING INFORMATION THEORY

This section looks at a mathematical model for choosing the best attribute and at methods for dealing with noise in the data. It can be skipped by those who are not interested in these details.

The scheme used in decision tree learning for selecting attributes is designed to minimize the depth of the final tree. The idea is to pick the attribute that goes as far as possible toward providing an exact classification of the examples. A perfect attribute divides the examples into sets that are all positive or all negative. The *Patrons* attribute is not perfect, but it is fairly good. A really useless attribute such as *Type* leaves the example sets with roughly the same proportion of positive and negative examples as the original set.

All we need, then, is a formal measure of "fairly good" and "really useless" and we can implement the CHOOSE-ATTRIBUTE function of Figure 18.7. The measure should have its maximum value when the attribute is perfect and its minimum value when the attribute is of no use at all. One suitable measure is the expected amount of **information** provided by the attribute, where we use the term in the mathematical sense first defined in (Shannon and Weaver, 1949). To understand the notion of information, think about it as providing the answer to a question, for example, whether a coin will come up heads. If one already has a good guess about the answer, then the actual answer is less informative. Suppose you are going to bet $1 on the flip of a coin, and you believe that the coin is rigged so that it will come up heads with probability 0.99. You will bet heads (obviously), and have expected value $0.98 for the bet. That means you would only be willing to pay less than $0.02 for advance information about the actual outcome of the flip. If the coin were fair, your expected value would be zero, and you would be willing to pay up to $1.00 for advance information—the less you know, the more valuable the information.

INFORMATION

Information theory uses this same intuition, but instead of measuring the value of information in dollars, it measures information content in **bits**. One bit of information is enough to answer a yes/no question about which one has no idea, such as the flip of a fair coin. In general, if the possible answers v_i have probabilities $P(v_i)$, then the information content I of the actual answer is given by

$$I(P(v_1), \ldots, P(v_n)) = \sum_{i=1}^{n} -P(v_i) \log_2 P(v_i)$$