not sure what it’s doing,” they told her. “I’m not sure I can trust it.”

Rich Caruana, a computer scientist at Microsoft Research in Redmond, Washington, knows that lack of trust firsthand. As a graduate student in the 1990s at Carnegie Mellon University in Pittsburgh, Pennsylvania, he joined a team trying to see whether machine learning could guide the treatment of pneumonia patients. In general, sending the hale and hearty home is best, so they can avoid picking up other infections in the hospital. But some patients, especially those with complicating factors such as asthma, should be admitted immediately. Caruana applied a neural network to a data set of symptoms and outcomes provided by 78 hospitals. It seemed to work well. But disturbingly, he saw that a simpler, transparent model trained on the same records suggested sending asthmatic patients home, indicating some flaw in the data. And he had no easy way of knowing whether his neural net had picked up the same bad lesson. “Fear of a neural net is completely justified,” he says. “What really terrifies me is what else did the neural net learn that’s equally wrong?”

Today’s neural nets are far more powerful than those Caruana used as a graduate student, but their essence is the same. At one end sits a messy soup of data—say, millions of pictures of dogs. Those data are sucked into a network with a dozen or more computational layers, in which neurons...
AI in Action: How algorithms can analyze the mood of the masses
Matthew Hutson

Science 357 (6346), 23.
DOI: 10.1126/science.357.6346.23