Ethics in Machine Learning and Other Domain-Specific AI Algorithms Imagine, in the near future, a bank using a machine learning algorithm to recommend mortgage applications for approval. A rejected applicant brings a lawsuit against the bank, alleging that the algorithm is discriminating racially against mortgage applicants. The bank replies that this is impossible, since the algorithm is deliberately blinded to the race of the applicants. Indeed, that was part of the bank's rationale for implementing the system. Even so, statistics show that the bank's approval rate for black applicants has been steadily dropping. Submitting ten apparently equally qualified genuine applicants (as determined by a separate panel of human judges) shows that the algorithm accepts white applicants and rejects black applicants. What could possibly be happening? Finding an answer may not be easy. If the machine learning algorithm is based on a complicated neural network, or a genetic algorithm produced by directed evolution, then it may prove nearly impossible to understand why, or even how, the algorithm is judging applicants based on their race. On the other hand, a machine learner based on decision trees or Bayesian networks is much more transparent to programmer inspection (Hastie, Tibshirani, and Friedman 2001), which may enable an auditor to discover that the AI algorithm uses the address information of applicants who were born or previously resided in predominantly poverty-stricken areas. AI algorithms play an increasingly large role in modern society, though usually not labeled "AI." The scenario described above might be transpiring even as we write. It will become increasingly important to develop AI algorithms that are not just powerful and scalable, but also transparent to inspection—to name one of many socially important properties. Some challenges of machine ethics are much like many other challenges involved in designing machines. Designing a robot arm to avoid crushing stray humans is no more morally fraught than designing a flame-retardant sofa. It involves new programming challenges, but no new ethical challenges. But when AI algorithms take on cognitive work with social dimensions-cognitive tasks previously performed by humans—the AI algorithm inherits the social requirements. It would surely be frustrating to find that no bank in the world will approve your seemingly excellent loan application, and nobody knows why, and nobody can find out even in principle. (Maybe you have a first name strongly associated with deadbeats? Who knows?) Transparency is not the only desirable feature of AI. It is also important that AI algorithms taking over social functions be predictable to those they govern. To understand the importance of such predictability, consider an analogy. The legal principle of stare decisis binds judges to follow past precedent whenever possible. To an engineer, this 1 The Ethics of Artificial Intelligence preference for precedent may seem incomprehensible—why bind the future to the past, when technology is always improving? But one of the most important functions of the legal system is to be predictable, so that, e.g., contracts can be written knowing how they will be executed. The job of the legal system is not necessarily to optimize society, but to provide a predictable environment within which citizens can optimize their own lives. It will also become increasingly important that AI algorithms be robust against manipulation. A machine vision system to scan airline luggage for bombs must be robust against human adversaries deliberately searching for exploitable flaws in the algorithm— for example, a shape that, placed next to a pistol in one's luggage, would neutralize recognition of it. Robustness against manipulation is an ordinary criterion in information security; nearly the criterion. But it is not a criterion that appears often in machine learning journals, which are currently more interested in, e.g., how an algorithm scales up on larger parallel systems. Another important social criterion for dealing with organizations is being able to find the person responsible for getting something done. When an AI system fails at its assigned task, who takes the blame? The programmers? The end-users? Modern bureaucrats often take refuge in established procedures that distribute responsibility so widely that no one person can be identified to blame for the catastrophes that result (Howard 1994). The provably disinterested judgment of an expert system could turn out to be an even better refuge. Even if an AI system is designed with a user override, one must consider the career incentive of a bureaucrat who will be personally blamed if the override goes wrong, and who would much prefer to blame the AI for any difficult decision with a negative outcome. Responsibility, transparency, auditability, incorruptibility, predictability, and a tendency to not make innocent victims scream with helpless frustration: all criteria that apply to humans performing social functions; all criteria that must be considered in an algorithm intended to replace human judgment of social functions; all criteria that may not appear in a journal of machine learning considering how an algorithm scales up to more computers. This list of criteria is by no means exhaustive, but it serves as a small sample of what an increasingly computerized society should be thinking about. 2. Artificial General Intelligence There is nearly universal agreement among modern AI professionals that Artificial Intelligence falls short of human capabilities in some critical sense, even though AI algorithms have beaten humans in many specific domains such as chess.