Teste de Turing

Com o nome de Allan Turing, o "teste de Turing" é a capacidade de uma máquina de exibir comportamentos inteligentes que são indistinguíveis dos de um humano. Uma descrição básica do teste de Turing é a seguinte: um interrogador humano remoto deve distinguir entre um computador e um sujeito humano com base em suas respostas a várias perguntas colocadas pelo interrogador. O interrogador humano não pode ver ou ouvir o robô ou o sujeito humano, mas só pode se comunicar via texto. Por meio de uma série de tais testes (dentro de um determinado período de tempo) o sucesso de um computador em "pensar" pode ser medido pela incapacidade do interrogador de distinguir entre o computador e o sujeito humano.

Allen, Varner e Zinser discutem um Teste de Turing oral para julgar o sucesso de um agente moral automatizado. Allen, Varner e Zinser observam que um teste somente de linguagem baseado apenas em justificativas morais seria inadequado; eles consideram um teste baseado em ações morais, em vez de apenas descrição de razões. "Uma maneira de mudar o foco de razões para ações pode ser restringir as informações disponíveis ao juiz humano de alguma forma. Suponha que o juiz humano no MTT seja fornecido com descrições de ações reais e moralmente significativas de um humano e uma AMA, expurgados de todas as referências que identificariam os agentes. Se o juiz identificar corretamente a máquina em um nível acima do acaso, então a máquina falhou no teste. Embora tenham o cuidado de notar que a indistinguibilidade entre agentes humanos e automatizados pode definir a barra para passar no teste muito baixo, tal teste por sua própria natureza decide a moralidade de um agente com base nas aparências. Outros argumentaram que o MTT não é suficiente para explicar uma

máquina como moral. Parthemore e Whitby propõem três blocos de construção que qualquer candidato bem-sucedido para agência moral deve ter: o conceito de si mesmo, o conceito de moralidade e o conceito de conceito.

Os Andersons argumentam que é preciso fazer uma distinção entre um agente moral pleno, que é um agente que pode ser responsabilizado moralmente por suas ações, e um *agente ético* que executa consistentemente ações moralmente corretas e pode justificá-las se solicitado. É possível ser um agente ético e ainda não se deve ser responsabilizado moralmente pelas ações. Uma máquina eticamente treinada e autônoma é um exemplo. Idealmente, a máquina deve ser treinada para seguir princípios éticos gerais que lhe permitam determinar a ação eticamente correta mesmo em situações que não faziam parte de sua formação.

Os Andersons criaram uma variante do MTT que simplesmente compara a ação eticamente preferível especificada por uma máquina, seguindo seus princípios, em um dilema ético com o de um *eticista* diante do mesmo dilema. Se um número significativo de respostas dadas pela máquina corresponder às respostas dadas pelo eticista, então ele passou no teste. Tal avaliação mantém os princípios gerados pela máquina para os mais altos padrões e, além disso, permite evidências de melhoria incremental à medida que o número de correspondências aumenta.