机器人三定律能保证人类安全吗？

戴益斌

未来社会，随着智能机器人越来越多，人类与机器人之间应该如何相处？阿西莫夫在小说集《我，机器人》中首次就未来社会的人机伦理关系进行了思考，提出了著名的“机器人三定律”：

第一定律：机器人不得伤害人，也不得见人受到伤害而袖手旁观；

第二定律：机器人应服从人的一切命令，但不得违反第一定律；

第三定律：在不违反第一定律和第二定律的情况下，机器人应保护自身的安全。

然而，机器人三定律是否适应未来社会的人机关系，能否保证人类安全呢？阿西莫夫本人对此并不乐观。

在小说《正电子人》中，阿西莫夫创造了一个终生恪守机器人三定律的机器人安德鲁，为马丁一家服务。然而，安德鲁的个体经历显示，机器人三定律之间充满矛盾。比如说，安德鲁在为马丁一家服务的过程中，曾遇到过流氓，他们要求安德鲁自我毁灭。根据机器人三定律，安德鲁必须听从人类命令，由于这一定律高于允许机器人自保的第三条定律，因此安德鲁应该自毁。事实上，安德鲁也确实打算自我毁灭，只不过后来被马丁的外孙乔治所救。

在另一本小说《我，机器人》，阿西莫夫设想了这样一个场景：一个疯子要放火烧毁一间房屋，房屋里住着一个人，而除了杀死这个疯子之外，机器人无法救房屋里的人，此时机器人是否应该制止这个疯子？阿西莫夫的回应是，应该制止。然而，一旦机器人准备制止，那么它有可能为了更好地遵守第一定律即不得见人受到伤害而袖手旁观，而不得不违反第二定律而制止疯子的行动，甚至有可能破坏第一定律而杀死疯子。

第一定律本身也面临着诸多问题。它不但可能是不必要的，而且自身也存在诸多漏洞。我们可以设想一个表面上很危险但对人类实质上没有任何危害的工作

场景，在这一场景下，如果存在一个机器人，那么人类将无法完成任何工作。因为根据第一条定律，机器人将阻止人类处于危险之中。此外，第一条定律并不能绝对地阻止机器人见死不救。在阿西莫夫的设想中，如果机器人发现自己在救人的途中必然会毁灭，比如必须通过某种对于机器人来说致命的电缆才可以救人，那么机器人通常会选择袖手旁观。因为对于这些机器人来说，即便它们选择去救人，也会在中途牺牲，不会改变被救者最终难逃一死的结局。而白白牺牲自己对机器人来说，并不是一个明智的决定。

不过，我们不应该过度担心机器人三定律中存在的漏洞。因为机器人三定律针对的是具有思想、能自主判断的强人工智能，目前人工智能的发展离这一目标还很遥远。我们仍有时间为未来的人机关系做出更好的规范。