机器人三原则

1942 年 3 月号的《惊奇》杂志上，著名科幻作家阿西莫夫发表了短篇小说《转圈圈》，首次提出了“机器人三原则”。之后，他在 1950 年出版的小说集《我，机器人》中系统阐述了这三原则：

1. 机器人必须不伤害人类，也不允许它见人类将受到伤害而袖手旁观；

2. 机器人必须服从人类的命令，除非人类的命令与第一条相违背；

3. 机器人必须保护自身不受伤害，除非这与上述两条相违背。

首先，他强调了“机器人必须不伤害人类”的大原则，但问题在于，他还提出了两个“除非”。两国军队交锋时，一国对另一国的军队使用机器人，请问是否违背了“机器人三原则”呢？人类是一个抽象的词语，一旦具体化，每一个人都有个性、脾气、好恶，人与人之间自然会有矛盾，国家与国家之间当然也会有冲突，此时，机器人应该站什么样的立场？

既然“不允许它见人类将受到伤害而袖手旁观”，那么当人类与人类自相残杀时，机器人是否能作壁上观呢？

更何况第二条已经规定了，“机器人必须服从人类的命令”，如果某国军事首脑命令机器人杀戮另一国的人类，那么机器人当做何选择？难道它还能抗令不尊不成？三原则的前两条难道不是矛盾的吗？

更离谱的就是第三条，什么叫“机器人必须保护自身不受伤害，除非这与上述两条相违背”？阿西莫夫的本意很显然，是在不伤害人类的大前提下谈这个第三原则的。但问题在于，这世界上，如果不是人类来伤害机器人，还有谁能伤害机器人呢？外星人吗？如果人类伤害了机器人，难道它就乖乖地束手就擒？

美国著名科幻小说家阿西莫夫。

实际上，阿西莫夫本人也意识到了这个问题，他在小说《可以避免的冲突》中设想出现这样的情况：机器人为了避免人类彼此伤害，便限制人类的行为，转由机器人掌控一切。这难道就是他提出“机器人三原则”最终想要的结果？也就是在处处保护人类、限制机器人的条件下，机器人居然依旧能逆势而为，统治了世界，这

可是如果它们一旦拥有了人类才能拥有的理性，它们是否就能取代人类呢？它们是不是会反过来奴役人类呢？这些问题一直困扰着科幻小说家，令他们创作出一个又一个作品，幻想着其中的无限可能性。