I, Protector

In the 2004 movie “I, Robot“, humanity narrowly escapes domination by robotic machines. The robots are engaged in secret plans and activities to take control of human affairs. What has made the robots do this when they have been programmed with a beneficent behavioral algorithm that ensure against human harm and (seemingly) maintain human autonomy?

Lets examine the algorithm instructions or “laws” as they are called. They are as follows:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

I submit that the following algorithm alleviates the domination activity portrayed in the movie. Explanation follows.

1. A robot may not injure a human being.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. A robot may not, through inaction, allow a human being to come to harm, except where such action would conflict with the First or Second Law.
4. A robot must protect its own existence as long as such protection does not conflict with the First, Second, or Third Law.

The secret and coercive domination portrayed in the movie is made possible because the “protector clause” is placed in the First Law, above the Second Law (human will). I have simply moved this clause to a position below the second law. The original placement of this clause charges the robots to take initiative, above human will, in order to protect them from harm. The robots thus construct a secret and coercive plan to protect humans from themselves. No human can subsequently order them to stop this process.

The revised set of laws allows robots to take initiative as before, but a human may now order the robot to stop or modify such an initiative.

In one particular scene in the movie, a policeman attempts to rescue a drowning child and subsequently endangers his own life. A robot forcibly intervenes to rescue the policemen, thereby preventing him from saving the child. The robot has calculated that the odds of the

child being rescued are lower than the odds of itself rescuing the policeman. The policeman, however, was willing to accept the risk to his own life in the attempt. His personal choice to risk his own life was overridden.

If we place our protection in the charge of sophisticated and powerful machinery (aka government and police states), above any individual autonomy and discretion, we will allow ourselves to become dominated by self-created protection systems. This, I believe, is the primary warning contained in the movie.

Preserving our autonomy and discretion will come with risks because we humans are fallible and not omniscient. Yet this may be, in the long run, the more ethical choice.