An Asimovian Framework for Regulating AI

Varun Jain
CEO
E: varun@singhania.in

When asked about how Open AI plans to monetize it's offerings, Sam Altman surprised the audience by saying that soon the platform will be intelligent enough to answer that question. 'More human than human is our motto' – the classic line on 'replicants' from the movie Bladerunner seems more and more plausible now!

How do you even begin to regulate something that grows so exponentially. It took us 20 years to reach a point where the proposed Digital India Act went from regulating digital transactions under the IT act to realizing that the real harms to worry about are not adequately enshrined in the statutes. User rights, safety, trust are now even more at risk under a new paradigm of content, cloud, and compute building on 25 years of hyper-content origination and user penetration.

Given that regulation will never therefore keep pace with the technology of AI, it is important to go back to first principles. In that regard, Isaac Asimov's rules on robotics and perhaps the morality debate around the development of the atom bomb are both useful ready reckoners.

AI at the end of the day is a finite state algorithm so far, and so is a robot. So, the parallels are worth drawing. The first rule of robotics says – 'A robot may not injure a human being, or through inaction, allow a human being to come to harm.' While currently the agency for such harm may continue to rest with a human actor, this is a very useful first principle to design regulation around. The 'kill-switch' to prevent such harms will be a key regulatory intervention that needs to be designed.

The second rule that 'A robot must obey orders given it by human beings except where such orders would conflict with the First Law' is equally insightful. First it presupposes no intelligence of it's own, but it also provides some boundary conditions for the algorithm. It is akin to a society trying to bind itself with some principles of morality and ethics, knowing too well that humans are essentially fallible. A normative definition of constitutes harm is therefore key for the algorithm to self-correct and learn accordingly.

Finally, the third rule is most prophetic and the subject of much cinematic delight from the Matrix to Terminator - 'A robot must protect its own existence as long as such protection does not conflict with the first or the second law.' Again, the key essence here is harm to humans, and that the AI entity or Robot must in some sense be terminated or self-destruct if it violates the first or second law.

Regulators and Platforms today have so far found it difficult to balance the principles of user trust, safety with innovation. AI will only make this situation harder. It seems the cloud and compute of the AI platforms will again get concentrated in a few hands on the West

Coast of the United States. So; elements of privacy, data protection, market dominance, algorithmic transparency, and equitable governance will continue to be contested spaces.

The Proposed Digital India Act is a step in the right direction that it is anchored around user-harm and takes cognizance of the issues above. It aligns well with the Asimov Rules, however, implementation will always lag behind platform innovation and externalities. We may know the triggers, but the regulatory and compliance interventions will be another story. Only time will tell.