Damning a machine’s soul

Marco Petix

The final objectives of punishment as a concept are commonly indicated as the rehabilitation of the offender, the satisfaction of the victims involved, so as to remove the desire for any kind of personal vengeance, and the discouragement of future crimes in others. All of the above can be further decomposed in simpler processes and functions whose mixes compose most of the punishments utilized by the contemporary society.

Incarceration, for example, is based on the concept of incapacitation, interfering with the freedom of the wrongdoer by extracting him from the environment he has harmed. Once again, the detention process often involves a sort of counseling or educational activity in order to rehabilitate criminals, influencing their behavior and, ultimately, trying to reduce the probability of recidivism. The whole system, just as any form of punishment, also aims to act as a deterrent to possible future crimes on the rest of society. The common fear of punishment, in this case imprisonment, is used in order to further discourage criminal behaviours.

Another very common form of punishment is the fine, this is based on the idea of repaying the victim for the damage suffered at the expense of the criminal’s finances.
Fines are based on the concepts of restitution, not harming the criminal freedom but his finances, and retribution, providing satisfaction for the victim.

Some societies also continue to use the death penalty as a form of punishment for more serious crimes. This represent an absolute renounce on the rehabilitating function of punishment and can be interpreted as the ultimate form of incapacitation, detaching permanently the criminal from the rest of the society.

Finally, punishment is used by the government to educate a society as a whole, expressing obvious moral disapproval of any action or event.

But could these very same processes be applied to an entity which reside in a dimension that is mostly digital? In the case of a crime caused by the decision of an artificial entity, who should be held responsible? Lastly, would a punishment be considered “necessary” and who should be involved in it?

Starting this analysis from the aforementioned artificial agent, we could say that it could surely be subjected to both an “incarceration” and a “death penalty”. Postponing the discussion about the utility of such premises, these could be interpreted as a temporary or definitive substitution of the agent from the role previously performed. In the case of a self-driving car this would result in disabling the model used by the neural-network responsible of the car's decisions and therefore disabling the whole car in the process.

Conceptually an artificial agent could also be involved in "re-educational programs”, similar to the well known counseling sessions and aimed towards the alteration of the agent’s behaviour. Punishments such as a fine, however, could not be imposed on the agent and would require to follow the chain of the responsibility involving one, or more, among the agent's user, developer or deployer.

In order to consider the user among the entities responsible, and thus punishable, for the crime or accident we should take into account his ability to predict the agent's decisions, or malfunctions, that caused the crime. Furthermore, the user's responsibility could also depend on the actual degree of control of the same in relation to the actions perpetrated by the agent. Ultimately, the user should be considered free of responsibility unless: he violated any rule through the use of the agent / device; the documentation on how to use the latter, provided by its developers and deployers, did not specify the possibility that the device acted, or malfunctioned, in such a way as to cause the crime; the user did not specifically pursued the conditions responsible for the behavior of the device.

In order to consider the agent's developers responsible for the crime, we should take into account the details of the agent's features and constraints required by whoever commissioned its creation. Unless the aforementioned constraints included elements that would specifically prevent the actions performed by the agent, these could not be traced back to the work of the developers. So, in this case, is the deployers who assessed the the developers’ product as acceptable and decided to actually deploy the agent, who should share the blame.

Finally we observe how the punishments designed for the artificial agent may not bear the same effects that we consider auspicable in regard of a punitive action applied to a human subject.

Incapacitating temporarily or definitely the agent would result in an easily predictable reduction of the crimes caused by it. But on the long run, once eventually reactivated, its previous inactivity wouldn’t discourage the agent from making the same mistakes. This could be proven wrong, thus validating the rehabilitating power of punishment on similar agents, if the latter would start to consider temporary incapacitation as an obstacle to its performance and would therefore shape its behaviour so to avoid the actions that caused it. Considering the agent as an entity guided by its desire for being performant could transform the main tool of individual deterrence from “the fear of punishment due to its unpleasantness” to “the fear of inefficiency caused by punishment”.

A rehabilitating process could also be enforced through additional training for the agent regarding data and situations similar to those present in the accident/crime in which it was involved. These data, obtained from the real accident or simulated, could help the agent to learn how to manage the conditions mentioned above without causing unwanted consequences and, ultimately, reducing the range of unexpected events of the agent, making it ideally more reliable.