Question for written answer E-005390/2020
to the Commission
Rule 138
Patrick Breyer (Verts/ALE)

Subject: Explaining black box AI machine learning models and using interpretable models

There has been criticism[1] of the fact that the private sector has an incentive to produce non-transparent black box algorithms in order to be able to commercialise them. Simple interpretable models with the same or a higher level of performance would be free for everybody to use.

1. What is the Commission's opinion of the proposal[2] that for certain high-stakes decisions, black boxes should not be allowed if there is already an interpretable model with the same level of performance?

2. What is the Commission's opinion of the proposal[3] that organisations that introduce black box models should also be obliged to report the accuracy of interpretable modelling methods?

[1] Rudin, Cynthia, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence* 1, 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x
[2] See above.
[3] See above.