XAI にとって「説明」とは何か：工学的・哲学的アプローチ

戸田聡一郎（Soichiro Toda）[1]・荒井翔悟（Shogo Arai）[2]

[1]東京大学大学院情報学環・[2]東北大学大学院工学研究科

人工知能（Artificial Intelligence, 以下 AI）技術は長足の進歩を遂げており、さまざまな分野に応用可能となっている。なかでも AI が「なぜ当該の判断を下したか」ということを、理由とともに説明することが可能な AI（explainable AI, 以下 XAI）が近年特に注目を浴びている。XAI は「解釈可能な AI（interpretable AI）」や、「信頼できる AI（trustworthy AI）」の議論にも大きな含意を持つ。本発表では、私たちが実践している「説明」を、後述する 3 次元空間において分類したうえで、それら説明の性質に対応した「理由」について XAI がどのような処理を行っているか、さらには当該の理由を人間にどのように提示できるかについて、工学的視点・哲学的視点から考察を加える。本抄録においては以下の 3 つの事例についてポスト現象学的視点からの考察を加え、「説明」の内実分析の導入としたい。

まず、よく例として挙げられる自動車の「自動運転」を瞥見しよう。XAI は車載センサーAI とは別コンポーネントとして、次のような説明をするだろう。「車載センサーAI が自動車の進行方向に少年が座っているのを確認し、少年に衝突しないよう、より周囲の環境が安全だと思われる右側にハンドルを切った」と。これで十分な説明になっているだろうか。少なくとも物理学的な説明にはなっているが、ここで着目すべきなのは、自動運転の説明には、行き先の入力以後は、車載センサーAI の判断から XAI の説明まで、私たち運転者の認知や意志が一切媒介していないことである。したがってここでの論点は、AI、XAI を含めた機械の性質に、人間にとって強制的（他律的）であり、かつ決定論的（algorithmic)な面がある、ということであろう。Ihde（1990）の図式によれば、自動運転技術と人の関係について、ATM 等と同じく次のような「他者関係」が成り立つ。

他者関係：　人間→(技術-世界))　（1）　→:志向性　-:関係

次に、機械の「強制性（他律的）」とは対極にある運動性のブレイン・コンピュータ・インターフェース（BCI）について考察を加えよう。運動性 BCI は非−人間の霊長類の基礎的実験を嚆矢として、いまや四肢麻痺患者などが彼らの脳活動データを記録し、それをデコードするコンピュータを介して、義手・義足等をまさに「意のままに」動かせる状況にまで技術が進化しつつある。「BCI の XAI」を考えるとき、現状の技術の場合、患者が起こした行動に対して XAI が「説明する」必要はない。したがって運動性 BCI の技術は、人間にとって自発的（自律的）であり、かつ決定論的(algorithmic）である、と言ってよい。Ihde のポスト現象学を発展させた Verbeek の表現を借りれば、「サイボーグ化」した関係が見える。

サイボーグ関係：　（人間–機械）→世界：（2）

3 番目の例として、上の図式(1)、(2)に当てはまらない例を考えよう。ここではニューロマーケティング（neuromarketing）を取りあげる。最近年では、脳波（electroencephalography, EEG）を用いてヒトの個々人の好みを割り出し、商品の販売促進につなげようという試みがなされている（Aldayel et al. 2020）。この技術は BCI の一種（認知型 BCI とする）考えられる認知型 BCI は、その人にとって特段強制力が働いているわけでもないが、完全にその人が自発的に購入しているわけでもない。つまり、強制的でもなく自発的でもなく、また決定論的でも非決定論的でもない。強制と自発の間に緩やかに横たわるこの種の技術は、まさにフーコーが指摘しているように、生権力的である（Foucault. 2004）。ここにおいて、当該個人がある商品を「買ってしまった」事態を、認知型 BCI とは別コンポーネントである XAI はどのように「説明」するのか。Verbeek は、こうした技術を「説得型技術」とし以下のような「合成関係」が生まれるとする。

合成関係:人間→（技術→世界）(3)

そのため、人間の志向性は技術の志向性と完全に合成され、AI が人間の判断に影響しているにも関わらず、XAI が解釈を行うために必要な明確な技術の他者性は消え、「説明」の対象は失われる。

だがしかし、「説明」の実在にこだわるならば、確率論的説明は可能である。ここで重要になるのは、冒頭でも言及した「説明の 3 次元的構造」であり、この 3 次元を構成する座標軸となるのは、次の 3 軸である。

i)「自発的（自律的）←→強制的（他律的）」軸;

ii)「決定論的（algorithmic）←→非決定論的（folk psychologic）」軸;

iii)「個人的（private）←→社会的（social）」軸

とりわけ「自発的←→強制的」軸の分析は、確率論的にも、フーコーを援用する哲学的にも、「説明」を腑分けしてその内実をえぐり出すうえで、重要な軸になるのは間違いない。本発表においては図式(1)～(3)、および「説明の三次元的空間」を視覚的に理解しやすいよう図示し、「説明」の工学的・確率論的議論を基礎づけるたたき台とする。

このように、一言で「XAI における説明」といっても、その「説明」は多元的座標空間内にスペクトラム状に散在または偏在している。更に言うなら、強制的でも自発的でもない説明において––つまりフーコー的な権力が働けば働くほど––XAI による説明の確度に関する確率が変動することについても詳細に議論したい。

＜参考文献＞

Aldayel, M., Ykhlef, M., and Al-Nafjan. Deep Learning for EEG-Based Preference Classification in Neuromarketing. *Applied Sciences* **10**:1525 (2020)

Ihde, D. *Technology and the Lifeworld.* Indiana University Press. (1990)

Verbeek, P. P. *Moralizing Technology. The* University of Chicago Press. (2011)

Foucoult, M. *The Birth of Biopolitics.* Palgrave Macmillan. (2004)