

裏から読んでも心理学

初心忘るべからず

平石 界
慶應義塾大学文学部教授



ひらいし・かい
東京大学大学院
総合文化研究科
博士課程退学。
東京大学、京都

大学、安田女子大学を経て、
2015年4月より慶應義塾大学。
博士(学術)。専門は進化心理学。

「一見したところ平凡な少女の隠れた才能に、落ちぶれた元名コーチだけが気づいて云々」というストーリー、ありますよね。あの手のものを目にする、ある俳優さんのことを思い出します。彼女がデビューしたところ、なんで事務所はこんな強烈に押してるんだらうってさっぱり分からなかったんですよ。ところがそれから数年、ふとした機会に出演ドラマを見てみたら「そういうことか」と納得しなくて、今や人気実力ともに日本を代表する俳優の一人に。なるほどこれがプロの眼力というものか。

とは言うもののあの手のサクセスストーリー、才能を見出すのが「落ちぶれた元名コーチ」ってのもよくある設定で、その理由が他の人には見えないものが見えてしまうことだったりしますよね。それで変人扱いを受けて周りから浮いてしまっただけで云々という。

何がって敵対的摂動 (adversarial perturbation) の話です。チャット君 (GPT) の登場でみんな忘れちゃってますけど、ほんの少し前までは機械 (AI) が犬や猫の画像を見て「犬!」「猫!」と識別できる“だけ”でも大したニュースでした。そんな盛り上がりにも水を差したのがこれ。誰がどう見てもパンダにしか見えない画像にこの敵対的摂動って隠し味を添加するとあら不思議、それまでどこまでも正確だったAIが「テナガザル!」とか答えるようになってしまっただけで、AI君も随分賢くなったようだけどまだまだだねえ、なんてため息をつかれたりしたのです。

それから数年「いや実は見えてなかったのは人間でしょ」という記事をネットで見つけて (佐藤竜馬, 2025)、何それ? と元の論文を読んでみたのです (Ilyas et al., 2019)。正直、数学サボってる勢ゆえ読み切れていないところが多々あって恥ずかしいのですが、ザックリと大枠だけでも。例えば犬の画像には、人間には見えない「犬っぽい」特徴が含まれているというのです。でもそれだけを取り出すと人間にはノイズにしか見えない。その犬ノイズを他の画像 (猫とか) に添加したものを作って「これは犬ですよ」とAIに教え込むのです。人には猫にしか見えないのだけど「犬だよ」「犬だよ」と教え込む。それから犬の画像を持ってきて、AIに「これはなに?」と尋ねたのです。

分かりにくいと思うので、こんなことを想像して下さい。猫を次々と見せられて「これはköpekですよ」と教えられるのです。実はその猫🐱画像には犬ノイズが添加されてるのだけれど人間には分からない。逆に猫ノイズ付き犬🐶画像は「kediですよ」と教わる。一通り学んだあとで混ぜり物のない犬の画像🐶を見せられて「これなに?」ときかれる。思わず「kediでしょ」と答えそうなところ、AIは🐶をköpekと答えたっていうんです。ちなみにköpekはトルコ語で犬の意味。AIには犬ノイズがきちんと見えていたので「あーこのパターンはköpekだな」と識別できた。AIが敵対的摂動に弱いってのは勘違いで、AIはむしろ人に見えないものを見ていた。元名コーチの目はくるってなかったってことでしょうか。

ああ、いつものAIがすごいって話ね、と思われる方もいるでしょうけど、少し違うような。確かにAIさん、「追試が失敗しそうな社会科学論文を論文テキストから当てる」なんてことをやってのけたりするものの (Hewitt et al., 2024)、「9才までの生育環境からその子の15才時の成績を当てる」ってタスクだと大したことなかったりする (Lundberg et al., 2024)。決して万能ではない。犬猫の識別も百発百中ではありませんでした。ポイントは人間の方であって、私たちは世界に存在するパターンの全てを認識できてるわけではないってことでしょう。「そんなノイズに“犬っぽさ”があるはずない!」と言いたくなる気持ちは分かるけれど、これは受け入れざるを得ない。それに考えてみれば、人間には知覚できない外界情報があることなんて、紫外線や超音波の例を引くまでもなく、心理学の基本のキなのでした。