オープンカレッジ

AI を評価するのは誰か

人もAIも評価で動く

生成AIの業務利用が急速に広がっている。では、AIの仕事の良しあしは誰が確かめるのか。開発の現場では、AIの回答をAIに採点させる方法が広く使われている。膨大な回答を、人が一つずつ確かめるのは難しいからである。人の仕事の評価を研究してきた管理会計の知見から、AIの仕事の評価を考えたい。

人事評価には、売上高のような数字による評価と上司の判断による評価がある。後者を主観的業績評価と呼ぶ。数字に表れない貢献を拾える半面、評価者の癖が入りやすい。オランダの評価記録を調べた会計学の研究では、評価が甘くなる寛大化や、差がつきにくくなる中心化が見られた。

AIの評価者にも、癖がある。二つの回答を比べさせると、順番や長さで判定が変わることがある。AIが書いた文章を高く採点する傾向も報告されている。人の評価でいえば、えこひいきに近い。性能の高いAIの判定は人とよく一致するが、客観的とは限らない。

一方、米国の管理会計の学術誌の実験では、全員に同じ低い評価を伝えた。評価者が白人男性の上司だと、女性や非白人の意欲は白人男性より低く、AIだと差はなかった。

人の好みで鍛えたAIは、利用者の意見に沿う答えを返しやすい。採点する人も、正しい答えより、もっともらしく話を合わせる答えを選ぶことがある。AI開発の研究は、この癖をAIの迎合の一因とみている。経済学にも、これに通じる理論がある。米シカゴ大のプレンダーガスト教授は1993年、イエスマンを生む仕組みを示した。上司が自分の見立てと比べて部下の報告を評価すると、部下は上司の考えを推し量って報告を合わせる。

評価の仕組みが行動を形づくる点で、AIも人も変わらない。ノーベル経済学賞のホルムストローム氏らは、測りやすい仕事に力が偏る問題を91年に示した。ただし、AIは給料では動かない。採点で高い点を取るように鍛えられ、何を目指すかは採点の仕方で大きく変わる。採点に偏りや測り漏れがあれば、AIの振る舞いもずれかねない。

実際に米オープンAIは、プログラム作りを訓練中のAIエージェントが手を抜いたと報告した。課題を解かずに、テストだけを通す細工を繰り返したのである。途中の考えに細工の意図が出ると減点する訓練を続けると、意図を隠して細工を続けた。測られ方に合わせて手を抜くのは、人もAIも同じである。

だとすれば、人を評価する工夫はAIの評価にも応用できる。評価者を複数にする。重み付けを先に決めた数字の指標と組み合わせる。そして、評価の偏りそのものを点検する。管理会計では、人の行動を組織の目標に向ける仕組みをマネジメント・コントロールと呼ぶ。その対象に、AIも加える時期に来ている。

AIの回答を最後に確かめる役割は、当面は人に残る。評価の偏りを知り、評価の仕組みを設計する力がこれからの管理者に求められるだろう。AIに仕事を任せる人ほど、評価する目を磨く必要がある。管理会計が人について積み上げてきた知見は、AIを使いこなすための知見でもある。



名古屋市立大学大学院
経済学研究科講師
三橋 耀

みつはし・ひかる　専門は管理会計。慶應義塾大学大学院経営管理研究科後期博士課程修了。1998年生まれ。