数理の窓

村には狼の振りしたAIが紛れている！

この村には狼が紛れている。狼は、昼間は人間と区別がつかない人狼だが、夜になると狼となって村人を襲う。村人達は昼間の話合いによって、狼と思われる村人を毎日一人ずつ排除していくのだが。。

これは、人狼ゲームという会話型ゲームの背景である。プレイヤーは村人役と人狼役に別れ、会話により誰が狼であるか見破っていく。各々は「自分が狼でないこと」を説得しなくてはならない。そして、夕方に投票によって誰を追放するか決める。一方で、狼は夜、村人一人を殺害する。数日を過ごし、すべての狼が追放されれば、村人陣営の勝ち、村人が全滅すれば人狼陣営の勝ちというチーム戦である。プレイヤーには、特殊能力を持つ占い師などがおり、「B君は狼でない」などの真の情報も与えていく。

最近AIに人狼ゲームをプレイさせる試みがある[1)]。AIは囲碁や将棋のような“完全情報ゲーム”で既に人を上回ったが、次のステップは非完全情報ゲームへの挑戦である。ポイントは、AIが会話を理解できるか、嘘をつけるか、見破れるかだ。

例えば、Aさんが「B君は狼で、Cさんは村人だと思う」と発言したとする。この意図には、実はAさんとCさんが狼ペアでB君を落としいれている可能性と、本当に村人でB君を狼と知っている可能性がある。つまり、会話により以下の3つの世界層に情報が与えられることが分かる。

- 現実（真実）世界
- 各プレイヤーが他プレイヤーに信じさせたい世界
- 各プレイヤーが知っている世界

ここで、AIは会話[2)]をデータとして処理するために様相論理フレームを使う。様相論理とは普通文にプラスして「知っている」や「信じる」といった知識文を扱う論理体系であり、他の文から別の文を推論規則により導くことができる。例えば「D氏は占い師」で「D氏が『B君が狼でないこと』を知っている」ということから、「Aさんの『Cさんは村人だ』と言うのは嘘だ」を導く。このように大量に導かれた知識データの中には、互いに矛盾したものも含まれる。AIは、知識間の関連や、真実度合いに“重み付け”をしながら、なにが真実かの評価を行う。

ただし、実際のゲームではデータだけで真実に達するのは不可能で、かまかけの質問、情報を引出す会話の組立て、信頼構築の雑談などが必要だ。

ところで、下手な嘘をつきながら、こちらの正体を探ろうとするAIからは、排除されたくない生存意思を感じてしまう。人間社会にAIが紛れてくると、我々は「自分はAIでない」ことを信じさせる説得が必要になりそうだ。　　（外園 康智）

1）参考文献　「人狼知能」　鳥海不二夫他著　森北出版
2）現在のAIにとって、自然言語の自由度は高すぎるため、会話プロトコルの限定と進行プロセスの標準化が必要。