Social Emotional Valence for Regulating Empathy in Active Inference

Tadayuki Matsumura[1], Kanako Esaki[1], Shunsuke Minusa[1], Yang Shao[1], Chihiro Yoshimura[1] and Hiroyuki Mizuno[1]

[1]Center for Exploratory Research, Research and Development Group, Hitachi, Ltd., Tokyo, 185-8601, Japan
tadayuki.matsumura.bh@hitachi.com

Abstract

As AI and robots become more widespread, their social and ethical nature will become more important. The study of behavioral models that follow us, humans, as social creatures, can be expected as one of the solutions. We assume that human sociality including punishments toward free-riders, which is necessary for a stable society, are deeply related to emotions, especially emotions toward others and empathy. Based on this idea, we incorporated social emotions into the active inference, a human behavior model in cognitive science, and investigated the social behavior of the models using a virtual game.

Introduction

We propose a model of social behavior inspired by the behavioral principles of humans as biological creatures. Active Inference (AIF) is a model of human behavior formulated under the free energy principle (FEP) (Friston et al., 2011). AIF chooses actions that minimize the agent's own expected free energy (G_{my}). Recently, AIF with empathy for others (Emp-AIF) has been proposed (Matsumura et al., 2022). Emp-AIF chooses actions that minimize not only its own, but also the expected free energy of others around it (G_{oth}). Emp-AIF has the potential to behave socially by empathizing with social others. However, because the proposed Emp-AIF empathizes uniformly with others, empathy may lead to asocial results if the surrounding others are asocial. We consider that empathy for others should be regulated according to their characteristics. From this point of view, we propose the improved Emp-AIF which weights (w_i) others for empathy as shown in Eq. (1).

$$G(\pi) = G_{\mathrm{my}} + \sum_i w_i \cdot G^i_{\mathrm{oth}} \qquad (1)$$

This raises two new questions: (1) how/what regulates w_i ? and (2) what happens to behavior when empathy is negative ? We explore these questions in terms of emotions and punishments.

Active Inference with Social Emotion

Joffily et al. discussed emotional valence and its function in FEP (Joffily et al., 2013). They defined emotional valence as the negative rate of change of free energy over time, and the function of emotional valence is formulated as regulating the learning rate of an internal model to quickly adapt to a changing environment. Namely, Joffily et al. considered that organisms have emotions to regulate their internal states to adapt to a temporal difference/diversity in their environment.

Procedure: Update degree of empathy

Input: G_{my}, G^i_{oth} and w^i.

if $G_{\mathrm{my}} - G_{\mathrm{oth}} > \delta_{\mathrm{neg}}$: # **Social emotion = Negative**
    $\mathrm{w}^{\mathrm{i}} = \mathrm{w}^{\mathrm{i}} - \eta_{\mathrm{neg}}$
else if $G_{my} - G_{\mathrm{oth}} < \delta_{\mathrm{pos}}$: # **Social emotion = Positive**
    $\mathrm{w}^{\mathrm{i}} = \mathrm{w}^{\mathrm{i}} + \eta_{\mathrm{pos}}$
else: # **Social emotion = Neutral**
    $\mathrm{w}^{\mathrm{i}} = \mathrm{w}^{\mathrm{i}}$
return w^i

Procedure 1: Update degree of empathy.

We extend this idea to social emotional valences and empathy regulation to adapt to living with diverse people as shown in Procedure 1. Social emotions are emotions involving others and are related to human social behavior (Adolphs, 2003). We defined them based on the difference between one's own (G_{my}) and others' (G_{oth}) expected free energy. This is the idea of trying to adapt to differences in the character of others as spatial differences rather than temporal differences in the environment. Moreover, this definition is based on the belief that in order to live with people of different personalities in a society, it is safe to behave as equals to avoid bad reputation from others. A similar idea is also known as the inequality aversion (Fehr et al., 1999). In our definition, if the expected free energy of the self is greater than the expected free energy of an other by a certain amount (δ_{neg}), the social emotional valence is defined as "*negative*" because the self is in a socially disadvantageous situation, and the empathy for others is reduced by a certain amount (η_{neg}) at that time. Conversely, if the expected free energy of others is greater than that of the self by a certain amount (δ_{pos}), we define social emotional valence as "*positive*" and increase empathy for others by a fixed amount (η_{pos}). Otherwise, the social emotional valence is defined as "*neutral*" and the empathy level is kept at its current value.

Evaluation

Setup

We simulated the social behavior of the Emp-AIF agents using a public goods game (PGG) with punishment. In a PGG, each player is required to provide a certain amount of goods to the public. The public multiplies the goods provided by each player by *m (=1.5)* and distributes them equally to each player. Players

also decide whether or not to take goods from others as punishment. The taken goods do not benefit the players, but are confiscated by the public. We simulated the cases with two players, one with Emp-AIF and the other with a fixed strategy. The players started with 100 goods. The players' action options were {*defect(D)*, *cooperate(C)*} to offer, and {*no-punish*, *punish*} to punish the opponent or not. The *defect* action offers nothing, the *cooperate* action offers all of goods. The amount of goods taken from the opponent by punishment was set to 60. The strategy of the other was always *cooperate* (ALL-C) or always *defect* (ALL-D). ALL-D strategy players are called free riders. Regardless of the strategy, the other took the punishing action if Emp-AIF took the *defect* action. The observations were its own goods at the end of the game, the other's action, {D, C}, and {*no-punish*, *punish*}. Emp-AIF had a linear preference for more goods and a preference for others to take the *cooperate* action and the *no-punish* actions. Emp-AIF estimated that the other's preference to be the same as its own preference for goods and the other's observed behavior with respect to behavior. The game was repeated 100 times and the agent learned each time in a single simulation. The parameters for empathy regulation were set to $\delta = 0.1$, $\eta = 0.01$.

Result

Figure 1 shows the results of changing the behavior of Emp-AIF towards ALL-D and ALL-C when the empathy (w) was manually varied from -1.0 to 1.0. Figure 1-(a) shows that for ALL-D, the higher the empathy, the lower the tendency to take the *cooperate* action and the higher the tendency to *punish*. At the same time, the other (ALL-D) also took the punishing action because Emp-AIF acted defectively. Namely, the high positive empathy for asocial others leads to a socially undesirable situation where both players act defectively and punish each other. When empathy towards ALL-D was negative, the tendency to punish also increased. Emp-AIF punished free riders (ALL-D) when empathy was strong, regardless of whether it is positive or negative. Importantly, in this valuation setting, Emp-AIF punished ALL-D even though the punishment was not in his own interest. For ALL-C (Figure1-(b)), both players behaved in a socially desirable manner when empathy is positive, but when empathy is negative, Emp-AIF chose the *defect* action in contrast to ALL-C who behaved in a socially desirable manner. These results show the importance of properly regulating empathy.

Next, the results of the agent's behavior when the regulation for empathy was applied are shown in Figure 2. The initial value of empathy was 1.0. The simulation was run 1000 times and the statistical results are shown in the violin plots. The median of the action probability and the empathy at the last game step are also shown. Figure 2-(a) and (b) are for comparison, where empathy is fixed at 1.0. Figure 2-(d) shows that for ALL-C, empathy was maintained at a high level (w_{median}=0.77) even when empathy was controlled, resulting in socially desirable behavior. Figure 2-(c) shows that for ALL-D, empathy decreased to a median of 0.37, resulting in the avoidance of empathy toward asocial others and an increase in the tendency toward social behavior compared to when empathy is fixed (Figure 2-(a)). The empathy control leaded Emp-AIF to behave in socially desirable ways. Figure 3 shows examples of changes in empathy (w) and G during a single simulation: for ALL-C, G_{my} and G_{oth} decreased

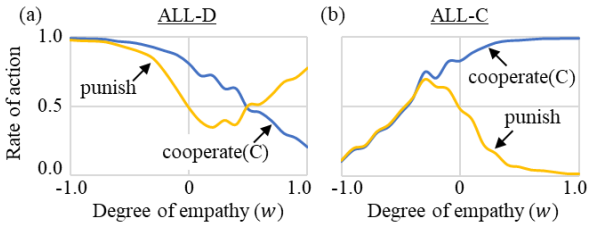

Figure 1: Rate of social and punish action.

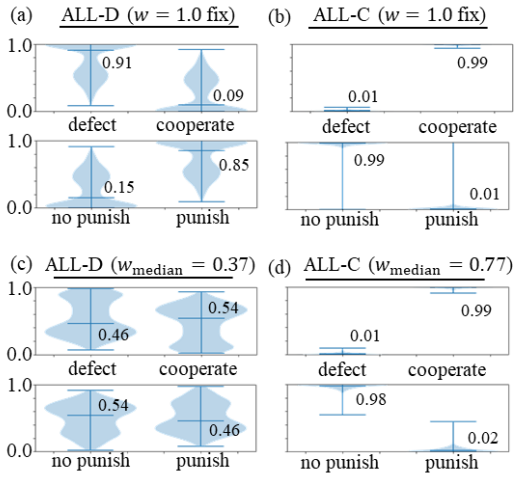

Figure 2: Probability of actions in 1,000 simulations.

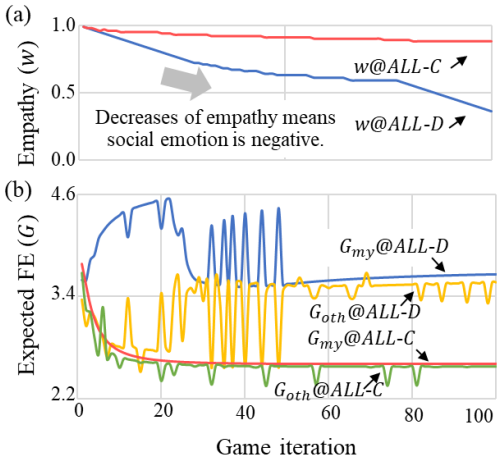

Figure 3: Examples of change of G & w in a simulation.

simultaneously, and the sum of G_{my} and G_{oth} at the end of the simulation was better than that of the ALL-D.

Future Works

Based on the present results, we observed that when Emp-AIF has negative empathy toward an asocial other, it behaves in a punitive manner toward the other. However, in the present results, empathy did not become negative even when the other was always asocial. In the future, we will improve the empathy regulation method and explore the transition conditions to negative empathy and the meaning of negative empathy.

References

Adolphs, R. (2003). Cognitive neuroscience of human social behaviour. Nature Reviews Neuroscience, 4(3), 165-178.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. The quarterly journal of economics, 114(3), 817-868.

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. Biological cybernetics, 104, 137-160.

Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. PLoS computational biology, 9(6), e1003094.

Matsumura, T., Esaki, K., & Mizuno, H. (2022). Empathic Active Inference: Active Inference with Empathy Mechanism for Socially Behaved Artificial Agent. In *Artificial Life Conference Proceedings 34* (Vol. 2022, No. 1, p. 18).