Self-Modeling Agents Evolving in Our Finite Universe

Bill Hibbard

$$ov(h_{i-1}a_i) = discrete((\sum_{i \le j \le t} \gamma^{j-i}\, u(h_j)) / (1 - \gamma^{t-i+1}))$$

$$o'_i = (o_i, ov(h_{i-1}a_i))$$

$$q_t = \lambda(h'_t) := \text{argmax}_{q \in Q}\, P(h'_t \mid q)\, \rho(q)$$

$$\rho(h') = P(h' \mid q_t)$$

$$v(h_t a) = \sum_{r \in R} \rho(ov(h_t a) = r \mid h'_t a)\, r$$

Watch for my new book: Ethical Artificial Intelligence