

**Hong Kong Catastrophic Risk Centre  
Hybrid Research Seminar**

# **AI Welfare Open Questions: The Problem of Genuine Desire**

**Prof. Simon GOLDSTEIN**

The University of Hong Kong

Do AIs have welfare? One step in answering this question is figuring out whether they have desires. For example, do AIs engage in complex planning that can only be explained in terms of promoting goals? But even if AIs have desires in some sense, the kind of desire they have may not contribute to welfare. This is the problem of "genuine desire". Cases of compulsion, addiction, and emotional disturbance in humans suggest that sometimes people can be systematically motivated to pursue goals whose satisfaction does not make their lives go better. The challenge is to identify a condition on genuine desire that explains which desires contribute to welfare and which do not. Then we can test whether AIs satisfy this condition. In this talk, I'll critically compare three conditions: that genuine desires produce pleasure, that genuine desires are reflectively endorsed, and that genuine desires are integrated into the agent's reinforcement learning processes. I'll argue that reflective endorsement is the best approach, and that this makes it relatively easy for AIs to have welfare.

**Venue: LKK302 & Zoom Seminar**

Click [HERE](#) or scan the QR code to join the zoom meeting.



**For more information:**

**<https://www.ln.edu.hk/philoso/hkcrc/>**

2616 7445/ hkcrc@ln.edu.hk

**3  
FEB  
2025**

**4:45-6:15 p.m. (HKT)**