⚡️🚀המאמר היומי של מייק -30.10.24: ⚡️🚀  
Beyond Preferences in AI Alignment

היום סקירה של מאמר ללא נוסחאות אבל קשה לי לקרוא לה קלילה. יש בה דיונים פילוסופיים לא פשוטים וזה מה שהבנתי מהם (תקנו אותי אם אני טועה).

המאמר מציג ביקורת מקיפה על הגישה המבוססת-העדפות(preference based) ליישור(alignment) של AI. המחברים טוענים שהגישה הנוכחית, המתמקדת בהעדפות אנושיות כיחידה הבסיסית של ערכים אנושיים, היא בעייתית ומוגבלת.

הם מציעים מסגרת חלופית המכירה בכך שהעדפות אנושיות הן מורכבות, משתנות לאורך זמן, ותלויות בהקשר חברתי. המאמר מציע גישה חדשה המבוססת על קריטריונים נורמטיביים ספציפיים לתפקיד (של המודל), במקום על העדפות גולמיות.

המאמר גם דן בצורך במערכות AI שמסוגלות להבין ולכבד את המורכבות של ערכים אנושיים, במקום לנסות לפשט אותם למודל של העדפות פשוטות. הם מציעים גישה חוזית (contractualist) ליישור AI, המבוססת על הסכמה הדדית בין בעלי עניין שונים. יש שם (במאמר) ביקורת על התיאוריה הקיימת של בחירה רציונלית (שהיא סוג של preference-based שיש לנו כרגע) ומציע חלופות המתחשבות במגבלות הקוגניטיביות האנושיות.

הכותבים מתייחסים לשאלה כיצד לטפל במצבים בהם העדפות שונות מתנגשות זו בזו. הם מציעים מודל חדש הנקרא Evaluate, Commensurate, Decide המתאר כיצד ערכים אנושיים משפיעים על העדפות. המאמר מציע כמה דרכים ליישום גישות אלו לאימון מודלי AI (בצורה די כללית אני חייב להגיד). המאמר מציע מסגרת (תיאורטית) לפיתוח מערכות המסוגלות להתמודד עם שינויים בהעדפות אנושיות לאורך זמן.

המאמר מדגיש החשיבות של פיתוח מערכות AI שיכולות לתפקד כ"כלים"(מתוחכם אבל מתמחה אך עם ״מרחב פעולות צר ומוגדר״) ולא כסוכנים אוטונומיים.

ניתן למצוא במאמר גם(איך לא) דיונים בחשיבות של שמירה על פלורליזם בפיתוח AI, כך שמערכות, משלבות AI, יוכלו לשרת מטרות שונות תוך כיבוד נורמות מוסכמות המשתנות לקבוצות שונות ולפעמים תלויות גם בנסיבות.

יאללה, עכשיו תגידו האם הבנתי נכון….

https://arxiv.org/abs/2408.16984