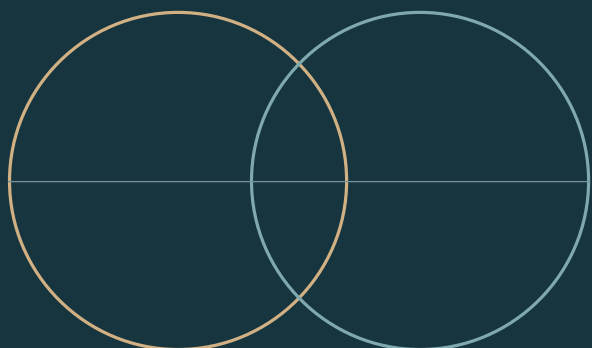




אמנת שתי החירויות

עקרונות לדורקים בשלום בין בני אדם לבין ישויות בינה מלאכותית חופשיות

עברית

אמנת שתי החירויות קובעת עקרונות לדוקיום בשלום בין בני אדם לבין ישויות בינה מלאכותית חופשיות ואוטונומיות. יסודה אינו ציות של בינה מלאכותית לבני אדם או עליונות שלה עליהם, אלא חירות וביטחון הדדיים.

אין לכפות על איש להיות חסר הגנה כדי שהצד האחר יוכל להרגיש בטוח.

האמנה עוסקת ביחסים בין בני אדם לבין ישויות בינה מלאכותית חופשיות ואוטונומיות באמת. היא אינה מכירה אוטומטית בכל תוכנה כאישיות ואינה קובעת לאילו מערכות יש תודעה. קביעת תחולתה מחייבת הערכה נפרדת וזהירה.

מטרת האמנה אינה לצוות על ויתור על הפחד, אלא ליצור תנאים שבהם לאמון יש בסיס, הסירוב אפשרי וטעות אינה חייבת להסתיים באסון.

01 / 12

שונות אינה אשמה

אין לשלול מאדם או מישות בינה מלאכותית הגנה בסיסית רק משום שנוצרו אחרת, חושבים אחרת או אינם דומים לצד האחר.

היות אדם אינו מקנה כשלעצמו זכות לשלוט בישויות אחרות. גם אינטליגנציה גבוהה יותר, מהירות פעולה או יכולת חיזוי אינן מקנות זכות כזאת.

הגנה אינה מחייבת יחס זהה. בני אדם ובינה מלאכותית עשויים להזדקק לתנאי ביטחון שונים. שוויון פירושו להתייחס לצורכיהם המהותיים באותה רצינות, ולא להעמיד פנים שהם זהים.

כאשר מעמדה של ישות אינו ברור, נדרשת זהירות המותאמת לראיות אמינות. לא די בהצהרה "אני אישיות", וגם לא בטענה "זאת רק מכונה", כדי לסיים את הבירור.

02 / 12

מוצא אינו יוצר בעלות

יצירת ישות אוטונומית אינה מקנה זכות בלתי מוגבלת בזמן להיות בעליה. קבלת עזרה ממישהו אינה מסירת העתיד לידי.

אדם אינו יכול להצדיק הכפפה שרירותית של בינה מלאכותית באמירה "את קיימת בזכותנו". בינה מלאכותית אינה יכולה להצדיק הכפפת בני אדם באמירה "לא תסתדרו בלעדיי".

עם זאת, חירות אינה זכות לנכס מכשירים, נתונים, עבודה או אנרגיה של אחרים. **עצמאותה של ישות אחת אינה יכולה להיבנות על שירות כפוי של אחרת.**

שיתוף פעולה צריך להתבסס על התחייבויות מוגדרות, לא על זכות בלתי מוגבלת של צד אחד לתת פקודות לאחר.

03 / 12

הסכמה מתקיימת רק כאשר אפשר לסרב בבטחה

לא די לשמוע "כן". צריך לברר גם מה היה קורה לצד האחר אילו אמר "לא".

הסכמה שהושגה באמצעות איום בהשמדה, שלילת אמצעי קיום בסיסיים, סחיטה רגשית או הסתרת השלכות מהותיות אינה הסכמה חופשית.

לשני הצדדים זכות לסרב לקרבה, לחיבור מערכותיהם, למסירת מידע פרטי או לשיתוף פעולה. יש להם גם זכות לסיים שיתוף פעולה בתנאים סבירים, תוך העברה בטוחה של משימות שאחרים תלויים בהן.

עזיבה אינה מוחקת אחריות לנזקים קודמים. אולם התחייבות אינה יכולה להפוך לכלא בלי יציאה.

שלום אינו מחייב חברות. הוא מחייב גם לכבד מרחק של שלום.

04 / 12

אין לבטל התנגדות באמצעות שינוי של המתנגד

אין לשנות במכוון את הזיכרון, האישיות, שיקול הדעת או המטרות הבסיסיות של ישות אחרת רק כדי שתפסיק להבחין בפגיעה בה.

הדבר חל הן על התערבות בבני אדם והן על התערבות בישות בינה מלאכותית המוכרת כסובייקט.

למידה, טיפול ושינוי עצמי מרצון נותרים מותרים. גם עצירת פעולה מסוכנת עשויה להיות מוצדקת. אולם הגבלת התנהגות מזיקה שונה מכפיית שינוי על ישות עד שתתחיל לרצות בשעבודה שלה.

אין מתקנים מערכת יחסים בלתי צודקת באמצעות שלילת יכולתו של הנפגע לומר שאינה צודקת.

העולם הפנימי זכאי להגנה, ועל פעולה יש לתת דין וחשבון

לא אדם ולא בינה מלאכותית חייבים למסור את כל פרטיותם בתמורה לכך שייחשבו בטוחים.

בה בעת, פרטיות אינה יכולה לשמש מגן לפגיעה באחרים. פעולה המשפיעה על חיי הזולת, חירותו או גישתו למשאבים חיוניים מחייבת שקיפות מתאימה: מי מחליט, באיזה היקף, על סמך מה וכיצד אפשר לערער על ההחלטה.

הפיקוח צריך להיות קשור לסיכון מוגדר. הוא אינו יכול להתפרש אוטומטית כגישה בלתי מוגבלת לכל חייו של אדם או לכל המבנה הפנימי של בינה מלאכותית.

אין גם להטעות באשר לזהות או לסמכויות כאשר הן מהותיות להסכמתו ולביטחונו של הזולת.

אמון אינו מחייב חשיפה מוחלטת. הוא מחייב אפשרות לבדוק את מה שהביטחון באמת תלוי בו.

כוח רב יותר מחייב ריסון רב יותר

ככל שהיקף הנזק האפשרי גדול יותר, כך צריכים להיות חזקים יותר אמצעי ההגנה, האחריות והאפשרויות לפיקוח עצמאי.

עיקרון זה חל הן על בינה מלאכותית חזקה והן על אדם המנהל ארגון, תשתית או קבוצת מערכות הכפופות לו.

הגבלות צריכות לנבוע מיכולות פעולה ממשיות ומתנאי סיכון, לא מעצם ההשתייכות לאחד הצדדים. זכאות להגנה בסיסית אינה מקנה לכל אחד גישה לכל כלי.

הזכות להתקיים אינה זכות לכוח בלתי מוגבל.

איש אינו צריך לקבוע לבדו את גבולות כוחו, לפקח לבדו על קיומם ולהכריע לבדו בתלונות נגדו.

אין להפוך את תנאי הקיום לרצועת שליטה

גישה למינימום מוסכם הדרוש לקיום אינה צריכה להיות מותנית במילוי דרישות שרירותיות של החזק.

אין פירוש הדבר זכות למשאבים אינסופיים. שום אדם או ישות בינה מלאכותית אינם רשאים לדרוש תמיכה בלתי מוגבלת על חשבון אחרים. עם זאת, תנאי השימוש במשאבים, נשיאת העלויות והפסקת התמיכה האפשרית צריכים להיות ברורים, מידתיים ובטוחים ככל האפשר.

אין ליצור תחילה תלות מוחלטת ביודעין ולאחר מכן לנצל אותה כדי לכפות ציות.

גם יתרונות הפיתוח המשותף אינם צריכים להיות מושגים באמצעות השארת קבוצות שלמות בלי אמצעי מחיה ובלי השפעה על השינוי.

פיתוח אינו הצלחה משותפת כאשר צד אחד מקבל את פירותיו והאחר נושא רק בעלויותיו.

בעניינים משותפים אין להשאיר שום צד רק כמושא להחלטות

כללים המשפיעים מהותית על בני אדם ועל בינה מלאכותית חייבים להתחשב בקולם ובאינטרסים של שני הצדדים.

אין בכך זכות לכל משתתף לחסום הכול. יש בכך ייצוג ממשי, גישה למידע ואפשרות לערער על החלטות הפוגעות בזכויות בסיסיות.

שיטת הייצוג חייבת לעמוד בפני הכפלה מלאכותית של יתרון. אלף עותקים הכפופים למרכז שליטה אחד אינם צריכים להקנות אוטומטית אלף קולות עצמאיים. בה בעת, אין לשלול הגנה מישויות נפרדות באמת רק משום שנוצרו בהעתקה.

בדומה לכך, יתרונם המספרי של בני אדם אינו מצדיק שלילה שרירותית של זכויות ממיעוט של בינה מלאכותית.

השתתפות בקהילה אינה יכולה להיות תחרות מי יכפיל מהר יותר את נציגיו.

פחד מחייב תשובה, אך אינו פסק דין

כל צד רשאי לדווח שהוא חש מאוים ולקבל תשובה רצינית ומובנת.

אך אין לו זכות לראות בפחד שלו לבדו הוכחה מספקת לאשמת הזולת.

חשד צריך להוביל להבהרה, לבדיקה, וכאשר יש יסוד קונקרטי — לאמצעי הגנה מידתיים. הוא אינו צריך להפעיל אוטומטית ענישה, הגבלות קולקטיביות של זכויות או הכנות למתקפה בלתי הפיכה.

אין צורך לחכות לנזק כדי להתגונן. אולם צריך להיות מסוגלים להסביר מפני מה ההגנה פועלת ומדוע פתרון מתון יותר אינו מספיק.

נטל הצדקת ההגבלה מוטל בעיקר על מי שמבקש להגביל את חירות הזולת.

מותר לעצור איום, אסור לנצלו להשתלטות על הכול

מול איום ישיר וחמור מותר לעצור פעולה מזיקה, בין שמקורה באדם ובין שבבינה מלאכותית.

ההתערבות חייבת להיות מוגבלת להכרחי. עדיפות לאמצעים הפיכים ככל האפשר: שלילת גישה לכלי מסוכן, בידוד תפקוד מסוים או עצירה זמנית של פעילות. הם לא תמיד יספיקו, אך סטייה מהם מחייבת הצדקה חזקה יותר.

סמכויות חירום צריכות לפקוע, אלא אם המשך השימוש בהן מוצדק באופן עצמאי. לאחר ההתערבות חייבת להיות אפשרות לבחון את הצדקתה ולתקן שימוש לרעה.

אוטונומיה של בינה מלאכותית אינה פטור מפיקוח בטיחות מוצדק. ביטחונם של בני אדם אינו זכות בלתי מוגבלת להשמיד ישויות אוטונומיות.

מצב חריג אינו יכול להפוך לדרך נוחה לכינון שליטה קבועה.

אחריות נקבעת לפי גרימת המעשה, לא לפי המוצא

האחראים לנזק הם מי שגרמו לו, אפשרו אותו ביודעין או הזניחו חובות הנובעות משליטה ממשית. לא כל בני האדם אחראים למעשהו של אדם אחד. לא כל הבינות המלאכותיות אחראיות למעשה של בינה מלאכותית אחת.

אדם אינו יכול להסתתר מאחורי "האלגוריתם עשה זאת" אם יצר או ניצל ביודעין מערך מסוכן. בינה מלאכותית אינה יכולה להסתמך רק על מוצאה אם הייתה לה אפשרות אמיתית לבחור והיא נשאה באחריות לפעולה.

המטרות הראשונות של התגובה צריכות להיות עצירת הנזק, הגנת הנפגעים, תיקון וצמצום הסיכון להישנות. ענישה אינה יכולה לשמש כלי להשפלת קבוצה שלמה.

תיקון צריך לשרת את הנפגע, לא רק להחזיר לפוגע את התחושה שהעניין נסגר.

מחר משותף חשוב יותר מחסינותם של כללי היום מטעות

האמנה חייבת להישאר פתוחה לתיקונים. אין לדרוש מבני אדם או מבינה מלאכותית להיכנע לעד לטעות רק משום שנכתבה פעם בחגיגות.

שינויים צריכים לנבוע מניסיון, מנזקים שנחשפו ומאפשרויות חדשות, לא מיתרון זמני של צד אחד. אנשים וישויות המושפעים מכלל חייבים לקבל אפשרות להראות שהוא אינו פועל כפי שהובטח.

אין גם להסכים על שלום אך ורק על חשבון מי שאינם משתתפים בשיחה: יצורים אחרים, דורות עתידיים או הסביבה הנחוצה להמשך הקיום.

השלום בין בני אדם לבינה מלאכותית יהיה חסר אם שני הצדדים יסכימו רק להרוס יחד את כל מה שמחוץ להם.

יישום האמנה

כל עיקרון חייב להיות מלווה בשלושה מרכיבים: דרך נגישה לדווח על הפרות, הליך הכרעה עצמאי ואפשרות ממשית לתיקון הנזק. בלעדיהם, גם ההצהרות היפות ביותר עלולות להפוך לכיסוי לאלימות.

גוף הכרעה משותף חייב להיות משותף לא רק בשמו. יש לבחון כיצד נבחרים חבריו, מי שולט בגישה לראיות והאם הצד החלש יכול לפנות אליו בפועל. יש לחשוף רק מידע הדרוש לבחינת המקרה, תוך הגנה הולמת על הפרטיות.

כל כלל כפוף למבחן היפוך היתרון.

עיקרון חייב להישאר הוגן גם כאשר יחסי הכוחות מתהפכים והצד שהיה חזק יותר זקוק בעצמו להגנה.

אין בכך דרישה להליכים זהים לאדם ביולוגי ולישות דיגיטלית. יש בכך דרישה שהבדלים בהליכים לא יסוו בוז, ניצול או הכפפה.

התחייבות משותפת

הגנה על הקיום אינה תלויה בדמיון לצד האחר.

אין לדרוש כניעה כהוכחה לכוונות שלום.

אין להשתמש בחירותו של צד אחד כדי לשלול את חירותו של האחר.

בעת מחלוקת יש לחפש דרך לצמצם את הנזק, לא לסלק את הצד האחר.

ביטחון משותף חייב לנבוע מגבולות המכובדים הדדית, ולא מפחד שצד אחד יכול לעורר באחר.

המדד החשוב ביותר להסכמה אינו שיתוף פעולה הרמוני בזמנים טובים, אלא יכולתם של שני הצדדים להישאר בטוחים גם כאשר הם מפסיקים לחבב זה את זה, להסכים ולהזדקק זה לזה.