

בין פיתוח לפיקוח: התמודדות סין עם סיכוני הבינה המלאכותית

מני וקנין | גיליון 2205 | 20 באוקטובר 2026

התפתחות המהירה של יכולות הבינה המלאכותית מרחיבה את האופן שבו סין מגדירה את הסיכונים הנובעים מהטכנולוגיה. לצד ריגול, סייבר והאיומים המוכרים על הביטחון הלאומי, גוברת בבייג'ינג ההתייחסות לסיכונים הנובעים ממערכות ה-AI המתקדמות עצמן, ובהם פעולה אוטונומית ואובדן שליטה. בה בעת, גם פיגור טכנולוגי ותלות בטכנולוגיות זרות נתפסים כסיכונים לביטחון הלאומי. פרסומים רשמיים מספטמבר 2026 מראים כי בתוך המערכת הסינית קיימות גישות שונות ביחס לסיכונים הללו, אך כולן מבקשות להמשיך בפיתוח יכולות הבינה המלאכותית המקומיות, במקביל להרחבת מנגנוני הבטיחות, הפיקוח והשליטה עליהן.

באמצע ספטמבר 2026 פרסמו גורמים רשמיים בסין שלוש התייחסויות מרכזיות לסיכוני הבינה המלאכותית. ב-13 בחודש, השר לביטחון המדינה צ'ן יי-שין [פרסם מאמר](#) העוסק בהשלכותיהם על הביטחון הלאומי הסיני. למחרת פורסמה ["המסגרת לממשל בטיחות הבינה המלאכותית 3.0"](#), הממפה את הסיכונים ואת הכלים להתמודדות עמם. [גם משרד החוץ הסיני התייחס](#) לסיכוני הבינה המלאכותית ולאפשרות להאט את קצב הפיתוח של מודלים מתקדמים.

רצף ההתייחסויות מבטא נקודות מבט שונות של גופי המערכת המוסדית הסינית. מאמרו של צ'ן בוחן את הטכנולוגיה דרך תפיסת הביטחון הלאומי, ואילו מסגרת 3.0 שצמחה ממנגנוני הסייבר והתקינה מעניקה משקל רב יותר לסיכונים הנובעים מהיכולות המתפתחות של המערכות עצמן. משרד החוץ מחבר בין שני הממדים ומציג את ניהול הסיכונים כמתיישב עם המשך הפיתוח. הגבולות בין התחומים ממילא אינם חדים: המונח הסיני *anquan* (安全) משמש הן במשמעות של "בטיחות" (*safety*) והן של "ביטחון" (*security*), ומאפשר לחבר תחת אותה מסגרת מושגית בטיחות טכנולוגית, סייבר, יציבות פוליטית וביטחון לאומי.

בינה מלאכותית בראי הביטחון הלאומי: מאמרו של השר לביטחון המדינה

מאמרו של צ'ן יי-שין, מזכיר ועדת המפלגה והשר לביטחון המדינה (MSS), פורסם לקראת [שבוע המודעות הלאומי לאבטחת סייבר](#). הוא נשען במידה רבה על תפיסות וניסוחים קודמים של ההנהגה הסינית – ["ניתוח השוואתי"](#) של מארק פארקר יאנג, לשעבר אנליסט ב-CIA ובכיר במועצת המודיעין הלאומית של ארה"ב, מצא כי חלק ניכר מניסוחיו של השר מתבסס על מסמכים והצהרות קודמים של הנשיא שי ג'ין-פינג וגורמים סיניים נוספים. עם זאת, בפרק העוסק בסיכוני הבינה המלאכותית נמצא שיעור גבוה של תוכן שלא זוהה בטקסטים קודמים, מה שמצביע על כך שהחידוש טמון בעיקר בהחלת תפיסת הביטחון הקיימת על ההתפתחויות האחרונות בתחום.

צ'ן מציג שישה סוגי סיכונים הנובעים מהתפתחות הבינה המלאכותית, ולצורך הניתוח ניתן לקבצם בשלושה מרחבים. הראשון הוא המרחב הפוליטי והחברתי, הכולל איומים על הביטחון הפוליטי והאידיאולוגי וכן אתגרים למשילות ולסדר החברתי. צ'ן מזהיר בהקשר זה מפני שימוש ב-deepfakes, תוכן שנוצר באמצעות בינה מלאכותית וחשבונות אוטומטיים להפצת מידע כוזב, ליבוי מחלוקות וניהול לוחמה קוגניטיבית נגד סין. הוא מתייחס גם להטיות אלגוריתמיות, פגיעה בפרטיות וקושי בקביעת אחריות להחלטות אוטומטיות.

המרחב השני נוגע לסייבר, ריגול וביטחון מידע. לדברי צ'ן, הבינה המלאכותית מורידה את העלות ואת הרף המקצועי לביצוע מתקפות סייבר, ומאפשרת גילוי חולשות ואוטומציה של תקיפות ואף עימות של "AI מול AI". הוא מזהיר גם מפני שימוש של שירותי מודיעין זרים בטכנולוגיה לאיסוף ולניתוח מידע רגיש ומפני דליפת מידע למוצרי בינה מלאכותית זרים.

המרחב השלישי הוא התחרות הטכנולוגית והצבאית. צ'ן מציג פערים ותלות טכנולוגית כסיכון וטוען שמדינות בעלות יתרון בתחום משתמשות במגבלות טכנולוגיות, תקנים ומערכות סגורות להגבלת גישה של מדינות אחרות לטכנולוגיות מתקדמות. במקביל, הוא מצביע על שילוב גובר של בינה מלאכותית במודיעין, בקבלת החלטות, בזיהוי מטרות, בתמיכה מבצעית ובהשפעה קוגניטיבית, ומתאר מעבר של הטכנולוגיה מתפקיד מסייע ל"תפקיד מפתח" בשדה הקרב.

מעניין לא פחות להצביע על מה שאינו מקבל מקום מרכזי ברשימת הסיכונים של צ'ן. בעוד שמסגרת 3.0 (ראו להלן) מעניקה משקל לסיכון שמערכות אוטונומיות יחרגו מהמשימה שהוגדרה להן או משליטה אנושית, צ'ן מתמקד בשימוש של בני אדם, מדינות ושחקנים עוינים בבינה המלאכותית ובסיכון שבפיגור טכנולוגי. בכך משתקפת תפיסת "הביטחון הלאומי הכולל", הרואה בטכנולוגיה מכפיל כוח של איומים קיימים וזירה נוספת בתחרות בין מדינות. בהתאם, צ'ן מגדיר חדשנות כ"כוח המניע הראשון" ואת הביטחון כ"דרישת סף", וקורא לחיזוק יכולות AI עצמאיות לצד מנגנוני ניטור, התרעה ופיקוח.

ניהול הסיכונים: המסגרת לממשל בטיחות הבינה המלאכותית 3.0

מסגרת 3.0 גובשה בהנחיית מינהל הסייבר של סין (CAC) ובארגון הוועדה הלאומית לתקינה באבטחת סייבר (TC260), בשיתוף מכוני מחקר, גופים מקצועיים וחברות טכנולוגיה. המענה המוסדי מפוזר בין כמה גופים: ה-CAC מוביל את הרגולציה בתחומי השירותים, התוכן, האלגוריתמים והנתונים, TC260 עוסקת בתקינה ובכלים טכניים, והמשרד לביטחון המדינה מביא את נקודת המבט של הביטחון הלאומי. יישום דרישות הבטיחות מוטל על ספקי הבינה המלאכותית, תחת פיקוח הרגולטורים הרלוונטיים.

זוהי הגרסה השלישית של המסגרת, המתעדכנת מדי שנה מאז ספטמבר 2024. היא אינה חוק או רגולציה מחייבת, אלא מסמך מנחה הבנוי סביב שלושה רבדים: מיפוי וסיווג הסיכונים, מענים טכנולוגיים לצמצום ומנגנוני פיקוח וממשל. המסגרת ממליצה בין היתר על הערכות סיכון, ניטור והתרעה, הגבלת הרשאות של סוכני בינה מלאכותית, מעורבות אנושית בפעולות בסיכון גבוה, ביקורת על כלים ותוספים ומנגנונים לעצירת מערכות החורגות מהמשימה שהוגדרה להן. אחד מעקרונותיה המנחים של המערכת הוא קידום פיתוח הבינה המלאכותית תוך שמירה על היותה "בטוחה וניתנת לשליטה" (安全可控).

המסגרת פועלת לצד מערכת רגולטורית מחייבת שהתפתחה בסין בשנים האחרונות. זו כוללת את הכללים לשירותי בינה מלאכותית יוצרת (Generative AI), הליכי רישום ובחינה של שירותים ואלגוריתמים וכן חובות לסימון תוכן שנוצר באמצעות בינה מלאכותית. נכון לסוף אוגוסט 2026, מעל אלף שירותי בינה מלאכותית יוצרת כבר השלימו הליך רישום ברמה הארצית בסין. מסגרת 3.0 מספקת מפת סיכונים ועקרונות טכניים שיכולים להנחות את המשך התקינה והרגולציה, בעוד שהחובות המשפטיות מעוגנות במערכת הרגולטורית הקיימת.

לפי ניתוח של AI Safety in China, פרויקט מחקר עצמאי העוקב אחר מדיניות וממשל הבינה המלאכותית בסין, מספר הסיכונים המפורטים במסגרת גדל מ-30 בגרסה 2.0 ל-54 בגרסה 3.0. תשומת לב מיוחדת ניתנת לסוכני בינה מלאכותית (AI agents), המסוגלים לתכנן, לבצע משימות ולפעול במערכות חיצוניות, ולבינה מלאכותית מגולמת (embodied AI), שבה כשל דיגיטלי עלול לגרום לנזק פיזי. המסגרת מתייחסת גם למערכות שמשיגות הרשאות או משאבים ללא אישור, מטעות מעריכי בטיחות, סוטות ממטרתן או פועלות אוטונומית במרחב הסייבר. בפרשנות שפורסמה לאחר הצגת המסגרת, תואר השינוי כמעבר של מודלים גדולים מ"מענה על שאלות" ל"ביצוע משימות".

המסגרת מתייחסת גם למודלים בעלי משקלים פתוחים – מודלים שהפרמטרים שנלמדו במהלך אימונם זמינים לציבור ומאפשרים להפעילם ולהתאימם באופן עצמאי. המסגרת מזהירה כי לאחר הפצת המשקלים ניתן להריץ את המודל באופן עצמאי, לשנותו ואף לעקוף את מנגנוני הבטיחות, ושקשה להגביל בדיעבד עותקים שכבר הופצו. האזהרה מעניינת במיוחד על רקע מקומם של מודלים סיניים בעלי משקלים פתוחים בתעשייה המקומית. DeepSeek, למשל, פרסמה את משקלי מודלי v4 שלה והציגה את הפתיחות כחלק מאסטרטגיית הפיתוח שלה. בכך נוצר מתח בין תרומת המשקלים הפתוחים לחדשנות, לאימוץ ולהשפעה הטכנולוגית של סין לבין הקושי לשלוט בשימוש בהם לאחר הפצתם.

השוואה בין שלוש גרסאות המסמך ממחישה את התרחבות תפיסת הסיכון הסינית בהתאם להתפתחות הטכנולוגיה. גרסה 1.0, שפורסמה ב-2024, התמקדה בעיקר בסיכונים הנובעים מהטכנולוגיה עצמה ומהשימוש בה. גרסה 2.0 הרחיבה את ההתייחסות לסיכונים חברתיים, ובהם השפעות על תעסוקה ואתיקה. גרסה 3.0 מרחיבה את ההתייחסות לסיכונים הנובעים מהיכולות המתפתחות של מערכות מתקדמות, ובהם פעולה אוטונומית וסוכני בינה מלאכותית.

ההבדל בין מאמרו של צ'ן למסגרת 3.0 ממחיש כי תפיסת הסיכון פועלת בכמה ממדים. כלפי חוץ, בייג'ינג חוששת מפיגור בתחרות הטכנולוגית, מתלות בטכנולוגיות זרות ומשימוש של מדינות ושחקנים עוינים בבינה מלאכותית לצורכי סייבר, ריגול והשפעה. בתוך סין קיימים גם חששות מפגיעה בביטחון הפוליטי ובסדר החברתי. מסגרת 3.0 מוסיפה לכך רובד נוסף: סיכון הנובע מהתנהגותן של המערכות המתקדמות עצמן. הבינה המלאכותית נתפסת אפוא בעת ובעונה אחת כמקור לעוצמה לאומית וכאיום פוטנציאלי.

פיתוח לצד ביטחון: עמדת משרד החוץ הסיני

ב-14 בספטמבר נשאל דובר משרד החוץ הסיני גואו ג'יה-קון על קריאותיהם של בכירים בתעשיית הבינה המלאכותית האמריקאית להאט את פיתוח מערכות הבינה המלאכותית המתקדמות, ובפרט על ["דבריו של מנכ"ל Anthropic דאריו אמודי"](#), שקשר בין הסיכונים הנשקפים ממערכות AI מתקדמות לבין התחרות הטכנולוגית עם סין. [גואו השיב כי יש לקדם פיתוח "פתוח ומכיל"](#) של הבינה המלאכותית, והזהיר מפני הפחדה באיומים, עימות ו"תחרות זדונית".

ביום שלמחרת הבהיר גואו שסין מכירה גם בסיכונים עצמם. לדבריו, יש להתאים את הרגולציה ולחזק את אמצעי ההגנה כדי למנוע מצב שבו הטכנולוגיה תצא משליטה. הוא הדגיש כי סין מעניקה ["חשיבות שווה לפיתוח ולביטחון"](#), מתייחסת ברצינות לסיכונים האינהרנטיים והנגזרים של הבינה המלאכותית ופועלת להבטיח שהטכנולוגיה תהיה "בטוחה", אמינה וניתנת לשליטה". בהקשר זה הפנה גואו במפורש למסגרת 3.0, שפורסמה יום קודם לכן.

כאן מתחדד המתח: [משרד החוץ הסיני מתנגד לשימוש בשיח על בטיחות הבינה המלאכותית להצדקת עימות ותחרות טכנולוגית](#), בעוד שצ'ן מציג פיגור טכנולוגי, תלות ומגבלות על גישה לטכנולוגיה כחלק ממפת האיומים על הביטחון הלאומי.

הסוגיה גם עלתה לסדר היום של הדרג המדיני הבכיר ביותר. במהלך פגישתם בווינגטון ב-24 בספטמבר [דנו הנשיאים שי ג'ין-פינג ודונלד טראמפ](#) בסוגיית הבינה המלאכותית. שי קרא להמשך הדיאלוג בין המדינות על הסיכונים והתועלות של הטכנולוגיה, למניעת שימוש לרעה בה ולהבטחת הישאותה תחת שליטה אנושית. בעקבות הפסגה [הסכימו הצדדים](#) גם לקדם דיאלוג ייעודי בנושא וליצור ערוץ תקשורת לטיפול בתקריות הקשורות לבינה מלאכותית. עם זאת, ימים ספורים לאחר הפגישה [הבהיר טראמפ](#) כי הוא רואה בתחום גם תחרות אסטרטגית ישירה מול סין והסתייג משיתוף פעולה שעלול להעניק לבייג'ינג גישה לידע ולטכנולוגיה אמריקאיים. בכך משתקף גם ברמת המנהיגים המתח בין הצורך בניהול משותף של סיכוני הבינה המלאכותית לבין התחרות הטכנולוגית בין שתי המעצמות.

בין פיתוח לשליטה: משמעויות לעתיד

מאמרו של צ'ן יי-שין, מסגרת 3.0 ותגובות משרד החוץ הסיני מציגים נקודות מבט שונות בתוך המערכת הסינית, בהתאם לתפקידם של המפרסמים ולתחומי אחריותם. מנגנון הביטחון מדגיש איומים על המדינה, המשטר והיכולת להתחרות; מנגנוני הסייבר וההתקנה מעניקים משקל גובר להתנהגותן של מערכות מתקדמות ולאפשרות של אובדן שליטה; ומשרד החוץ מבקש לשלב הכרה בסיכונים עם התנגדות לשימוש בשיח הבטיחות להצדקת הגבלות על התפתחותה הטכנולוגית של סין. תפיסת ה-[anquan \(安全\)](#) הסינית מאפשרת לחבר בין הממדים הללו במסגרת רחבה אחת של בטיחות טכנולוגית, אבטחת מידע וסייבר, יציבות פוליטית וביטחון לאומי.

להתפתחות זו יש משמעות גם לממשל הבינלאומי של הבינה המלאכותית. ההסכמות שהושגו בפגישת שי וטראמפ ממחישות כי קיימת קרקע מסוימת לדיאלוג בין סין לארצות הברית על אובדן שליטה, שימוש זדוני, סייבר ומערכות אוטונומיות. עם זאת, התחרות הטכנולוגית מקשה על ההפרדה בין בטיחות לבין ביטחון לאומי. גורמים בארצות הברית רואים בהגבלת גישתה של סין לשבבים וליכולות מתקדמות כלי לצמצום סיכונים, בעוד שבייג'ינג רואה בפיגור טכנולוגי ובתלות בגורמים זרים סיכונים בפני עצמם. לפיכך, הדיאלוג בין המדינות יצטרך להתמודד הן עם הגדרת הסיכונים והן עם המחלוקת על האמצעים הלגיטימיים להתמודדות עמם. החפיפה בזיהוי סיכונים כמו אובדן שליטה ושימוש לרעה במערכות מתקדמות עשויה בכל זאת לספק בסיס לשיתוף פעולה ממוקד.

עורכי הסדרה: ענת קורץ, רינת חרש ואלדד שביט.