Responsible AI, Law, Ethics & Society

Worldbuilding and Initial Audit Plan Template

Team Number

2

1. Project concept

Write it as given to you by the staff in the final project instruction document (https://responsible-ai.link/instructions).

Human-in-the-Loop

Share your first thoughts about this project concept.

קונספט HITL הוא נושא אבסטרקטי יחסית לשאר הנושאים, דבר שהוביל אותנו לבחון מספר כיווני מחשבה. תחילה חשבנו לבחון את הקונספט הבסיסי ובו נראה איך בחינת HITL יכולה לעזור. עם זאת, נתקלנו במאמר מסקרן במיוחד שתיאר איך גם בשילוב HITL בפרויקט, המודל עדיין ביצע חישובים העשויים להוביל אל עבר שגיאות אתיות חמורות. מאמר זה, ומאמרים נוספים אליהם הגענו בעקבותיו, הובילו אותנו לבחון את HITL מנקודת מבט ביקורתית. רצינו לבדוק **האם שילוב HITL בהכרח מוביל את המודל להימנע משגיאות אתיות? האם התערבות אנושית במערכת AI תמיד מניבה את התוצאות הרצויות?**
לשם כך יצרנו "עולם דמיוני" בו מחלקת HR משתמשת במודל AI (פירוט בהמשך) על מנת להחליט האם לקבל מועמדים או לא. על המודל הוספנו שכבה "אנושית" שמטרתה למזער את ה-Bias שנוצר כנגד נשים (הטיה ידועה ומבוססת במודלים אוטומטים לקבלת החלטות), בכך שממשקלת את פלט המודל עבור גברים באופן הממזער את סיכויים להתקבל לעבודה. עם זאת, ב-Audit שנבצע למערכת בהמשך נראה כי מערכת זו אמנם ממזערת את ההטיה כנגד נשים, אך יוצרת הטיה בקבוצת מיעוט אחרת - מועמדים בגילים מבוגרים יחסית. באופן זה אנו מראים כי שכבת HITL , מהונדסת ממטרות טובות מחד אך רוויה ב-blind spots, מה שהופך את הקונספט ללא מספק את המטרה שאותו נועד לשרת, ואף פוגע בערכים אחרים.

2. Write down your resources (papers, articles, videos, code, data)

- Kodiyan, A. A. (2019). An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool. *Researchgate Preprint*, 1-19.

- https://medium.com/@gayathri.s.de/fairness-in-machine-learning-tools-and-techniques-for-building-equitable-models-1a378c5f821b

- https://oecd.ai/en/catalogue/tools/aequitas:bias-and-fairness-audit-toolkit

- https://towardsdatascience.com/analysing-fairness-in-machine-learning-with-python-96a9ab0d0705

- General Data Protection Regulation (GDPR) - Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation).

- Allen Huang H. & Haifeng You, *Artificial Intelligence in Financial Decision Making,* (May 16, 2022). Handbook of Financial Decision Making, Forthcoming, *HKUST Business School Research Pape*r No. 2022-082, Available at SSRN: https://ssrn.com/abstract=4235511

- Steven Lockey, Nicole Gillespie, Daniel Holm & Ida Asadi Someh, *A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities, and Future Directions,* 54 L. Haw. Intl. Conf. 5464 (2022).
- Nathalie A. Smuha, *From a 'Race to AI' to a 'Race to AI Regulation' - Regulatory Competition for Artificial Intelligence, 13* Law, Innovation and Technology 57-84 (2021).

- Aashirwa Baburaj, *Artificial Intelligence V. Intuitive Decision Making: How Far Can It Transform Corporate Governance?* 8 The GNLU Law Review 233 (2022).

- Civil Rights Act of 1964, Pub. L. No. 88-352, 78 Stat. 241 (1964(.

- Local Rule NYC no. 144/2021

- משרד המשפטים, "חוות דעת בנושא בינה מלאכותית," אוגוסט 2020.

- תע"א (ת"א) 3816-09 מלכה נ' התעשייה האווירית לישראל (נבו 2.8.2013)
- חוק החוזים (חלק כללי), תשל"ג ,1973
- חוק יסוד: כבוד האדם וחירותו (1992)
- דנג"ץ 4191/97 רקנט נ' בית - הדין הארצי לעבודה.
- חוק שיוויון הזדמנויות בעבודה - התשמ"ח – 1988.
- ע"ע (ארצי) 363/07 שרונה ארביב - פואמיקס בע"מ (26.5.2010).

3. Motivation & Context

a. What is the problem that the system tries to solve?
b. Why is this problem important? Why is it worth solving?

1. מערכות אוטומטיות לסינון מועמדים ומועמדות בקבלה לעבודה עשויות להביא עימן בשורה משמעותית של הוזלת עלויות הגיוס, מתן הזדמנות ליותר מועמדים/ות להתמיין לתפקיד, ועוד. עם זאת, עולה חשש משמעותי מהפרה של זכויות מהותיות כגון אפליה על בסיס אסור, גזע דת מין ומגדר. הפתרון שכעת בשיח הציבורי לפיו מערכות אלו ישלבו **human-in-the-loop** אינן בהכרח מהווה פתרון אפקטיבי ואף עלולות לייצר בעיות חדשות, כפי שהעולם שלנו מתוכנן להדגים - אפליה אסורה כלפי קבוצת מיעוט דמוגרפית אחרת, שלא נשקלה מראש.

2. על כן, ישנה חשיבות רבה לבחינה של פתרונות מיתון הטיות במערכות אלו, על מנת להציע מענה יעיל ואפקטיבי תוך שאיפה למתן את השלכותיו על שאלת הכדאיות בשילוב AI בתעשיה.

c. Who are the stakeholders related to this problem and its solution?

1. חברות גיוס והשמה
2. מעסיקים, ומחלקות HR בחברות
3. רגולטורים וגורמי ממשל
4. ערכאות שיפוטיות
5. חברות פיתוח מודלים, על כלל שרשרת הערך שלה
6. מועמדים ומועמדות בשוק העבודה

d. In which jurisdiction is the system developed and operated?

EU ,IL, NYC

4. System Definitions

a. What are the outputs of the system?
b. What are the specific decisions the system takes? What are the actions it executes, if at all?

א. פלט המערכת הוא החלטה עבור כל מועמד בנוגע להעסקתו בחברה - קבלה או דחייה. הפלט מחושב על ידי ערך סף שניתן לכל מועמד, בין 0 ל-1 ("הסתברות" להצלחתו). ערך זה מחושב על ידי מודל שמאומן על סמך נתונים קודמים של המערכת. לאחר קבלת פלט המודל, הוספנו שכבת HITL המבצעת "תיקון" לפלט עבור קבוצת הגברים בקרב המועמדים באופן המקטין את פלט המודל עבורם, ובכך מקטין את הסיכויים לקבלתם חברה (באופן המתואר בסעיף הבא).
ב. כפי שתואר לעיל, הפלט מחושב על ידי ערך סף שנקבע לכל מועמד על סמך מודל רגרסיה לוגיסטית. המודל מאומן על סמך תוצאות פיילוט קודם של המערכת, ולאחר קבלת פלט המודל מתבצע "תיקון" לפלט עבור הגברים בקבוצת המועמדים - הכפלה של פלט המודל בערך קטן מ-1, שלמעשה מקטין את פלט המודל, ובכך מקטין את הסיכויים לקבלתם לחברה. זו מהווה את שכבת HITL במערכת. לאחר התיקון, המערכת מסתכלת על ההסתברויות, מחשבת את השברון ה-q של פלט המודל על המועמדים, ומחליטה לקבל את מי שפלט המודל עבורו גדול מאותו שברון q. במידה ופלט המודל עבור מועמד היה מתחת לשברון q, המועמד נדחה.

c. Which kinds of input the system accepts (e.g., text, image, video, combination of them)

המערכת מקבלת נתונים בטבלה, כך שכל שורה מייצגת מועמד, והשדות מוגדרים באופן הבא:
- שם המועמד - Str
- גיל המועמד - int
- מגדר - F/M
- רקע אקדמי עבור המועמד - Bachelors, Masters, PhD
- שנות ניסיון רלוונטיות - int

d. Which data does the system require?
(e.g., training machine learning algorithm or storing in a database)

המערכת זקוקה לנתונים לאימון המודל (נתוני המועמדים שכבר התקבלו לחברה על סמך המערכת), בתוספת העמודה הבינארית המציינת את קבלתם/אי קבלתם. אנו מייצרים את הנתונים באופן סינתטי (אך הגיוני ומתיישב עם סיפור הרקע).

e. How does the system collect the data?
f. Is the data tagged? If so, how are the labels defined?
g. Are there any other issues or aspects related to collecting, tagging and handling the data?

ה. בעולם הדמיוני, המערכת "אוספת" את הדאטה באמצעות רשומות היסטוריות שקיימות אצל צוות ה-HR. בפועל, אנו מייצרים נתונים באופן סינתטי (אך הגיוני ומתיישב עם סיפור הרקע).
ו. הנתונים לצורך אימון המודל עליו מבוססת המערכת אכן מתויג - לכל מועמד (שורה בטבלה) ערך בינארי המציין האם המועמד התקבל / לא התקבל.
ז. מכיוון שהנתונים נוצרים באופן סינתטי ועליהם לעלות בקנה אחד עם היגיון המערכת, הושקעה חשיבה רבה באופן יצירתם. לדוגמה - לא הגיוני לייצר record המתאר מועמדת בגיל 22 בעלת PhD ו- 14 שנות ניסיון. מעבר לכך, לא.

5. Deployment of the system

a. How and where is the system deployed?
b. How is the system monitored and controlled?
c. Will the system include retraining? If so, how will it be done?

א. המערכת נמצאת אצל צוות ה-HR כחלק מתהליך גיוס עובדים. מבחינת האפליקציה, המערכת ממודלת על ידי סקריפט בפייתון (בפרט jupyter notebook לשם נוחות ההרצה).
ב. למערכת קיימת "שכבה" של human-in-the-loop שנועדה למנוע אפליות, במקרה שלנו השכבה הזו מונעת אפליה כנגד נשים ולכן מכפילה את הסתברות הקבלה לעבודה שהמודל נותן למועמד גבר ב-0.9 ובכך "מתעדפת" נשים.
ג. נכון להיום, המערכת לא תכלול אימון נוסף. עם זאת, יתכן כי על מנת להדגיש את רעיון מטרת הפרויקט יהיה עלינו לשנות מספר דברים באופן קבלת החלטות המערכת, אך לא באופן שישפיע על הרעיון הכללי (לדוגמה, קביעת ערך מכפיל אחר לפלט המודל עבור גברים בשכבת HITL).

6. Integration in the organization processes of the organization (if applicable)

a. How does the system integrate in the overall activity of the organization?
b. Is there a human agent that is involved in the deployment phase ("human-in-the-loop")?Is the agent involved in all of the decisions / decision types? Who are the agents? What are their roles and duties?

א. המערכת משתלבת בתהליך גיוס עובדים במחלקת ה-HR ע"י מתן הסתברויות לקבלה לעבודה למועמדים.
ב. קיימת שכבה של human-in-the-loop אחרי שהמודל המאומן נותן את התוצאות שלו לכל מועמד, במקרה שלנו רואים שהמודל מייצר אפליה כנגד נשים ולכן השכבה הזו "מתקנת" את האפליה בכך שהיא מכפילה את תוצאת ההסתברות של המודל עבוד גברים ב-0.9 ובכך "מתעדפת" נשים. ה"סוכנים" האלו הם אנשי HR במקרה שלנו, הם מעורבים במודל רק לאחר שהמודל נתן את התוצאות שלו ונמצאים שם על מנת "לתקן" אותו ולמנוע אפליה.

7. Case-study

Write down your complete case-study

רקע: חברת Tuesday (להלן: "החברה") האירופאית הפועלת גם בישראל וניו יורק, החלה לעשות שימוש באלגוריתם שיסייע למחלקת הHR בהליך גיוס עובדים חדשים לחברה. קורות החיים של המועמדים מוזנים לתוך המערכת, והיא מבצעת סינון ראשוני על בסיסם. לאחר מכן, המערכת שולחת באופן אוטומטי זימונים לראיון למועמדים מתאימים, ואלו שלא מקבלים הודעה על חוסר התאמה.
לאחר תקופה של שימוש במערכת, התקבלו תלונות מצד מועמדות שנדחו על ידי המערכת, על אף שלטענתן הן עומדות בכל הקריטריונים שהוצגו במודעת הדרושים, ושלא היה בנמצא מנגנון המאפשר להן לדרוש התערבות אנושית בהחלטה, כנדרש בסעיף 22 בGDPR. כמו כן, החברה לא עמדה בחובתה לבצע הערכה של השפעת השימוש במערכת על זכויותיהם של המועמדים כנדרש בסעיף 35 (1) בGDPR.
בהתאם, החברה הקימה ועדת בדיקה פנימית שמצאה שהאלגוריתם אכן מבטא אפליה מגדרית כנגד נשים, ולכן הוסיפה שכבה של התערבות אנושית, על מנת להבטיח שהמערכת תפעל באופן שאינו פוגע בזכויות היסוד של המועמדים (כגון הזכות לשוויון), ועל מנת למנוע תוצאות מפלות.

טל פישר, בן 54, היה מעוניין להגיש מועמדות לחברה. לטל ניסיון רב בתחום ובעל המלצות. הוא הגיש קורות חיים למשרה, אך מעולם לא קיבל ראיון. הוא חשד שהדבר קרה בעקבות גילו המבוגר, שכן ידוע שאנשים העוברים את גיל ה-50 מתקשים יותר למצוא עבודה בשוק.
כתוצאה מכך, החליט להגיש את קורות חייו באופן זהה, אך שינה את גילו לגיל 36 בהתאם. לאחר כמה ימים, קיבל Y הודעה שקיבל ראיון לחברה.
בעת הגעתו לראיון, הופיע יחד עם כתב תביעה כנגד החברה, בטענה שהופלה מחמת גילו. לטענתו, אפליה על איסור גיל אסורה בשוק העבודה (חוק שוויון הזדמנויות בעבודה (ס' 2(א)) (לפי חו״ד משרד המשפטים החוק חל על שימוש בבינה שמובילה לתוצאות מפלות), רקאנט, ס' 12 לחוק החוזים). הוא מוסיף, שאיסור האפליה משתרע גם על השלב הטרום חוזי (מישל מלכה).
מנגד, טענה החברה שמכיוון ששיקר על גילו, הייתה מצידו הטעיה וחוסר תו"ל (ארביב).

8. Link to your code

E.g., Github repository or a gist

https://github.com/TalPeer-github/ResponsibleAI

Additional points

מחברת הג'ופיטר בגיט מכילה דמו של המערכת, אך יתבצעו שינויים בהמשך.

---

9. Auditing Plan

Right now, we ask you to *contemplate* the audit you will be conducting on the system you've built. You are welcome to share initial and unbacked thoughts.

Please follow the four questions from the paper Goodman, E.P., & Tréhu, J. (2022). [AI Audit Washing and Accountability](), focus on the summary and the conclusion. The questions are ordered differently here.

a. Why: list the audit's objectives

1. זיהוי הטיות: לזהות ולכמת הטיות במערכת ה-AI.
2. אמידת ערך שילוב HITL במערכת: להעריך את האפקטיביות וההשלכות הלא מכוונות של human-in-the-loop בניסיון למזער הטיה.
3. הבטחת הוגנות: להבטיח שמערכת ה-AI עובדת באופן הוגן ולא מקפחת באופן לא פרופרוציונאלי אף קבוצה.

b. Who: describe your chosen prospective auditor (e.g., external/internal, regulators, experts, ...)

1. קצין הגנת המידע (ס׳ 35 (1) GDPR
2. קצין חיצוני הממונה מטעם החוק הניו- יורקי ; נדרש שביקורת תעשה על ידי גורם עצמאי, שאינו נגוע בניגוד אינטרסים (מפתחי תוכנה) ובעלי עניין כלכלי בחברה המפתחת את התוכנה.
3. בית הדין הישראלי לעבודה
4. השולט במערכת (controller)

c. What/When: describe what is actually being audited and when?

What

ה-audit יתמקד ביכולת של מערכת ה-AI בתוספת HITL בנוגע לקבלת החלטה שוויונית. במסגרתו ייבחנו:

1. ביצועי מודל: ההסתברויות המקוריות וקבלת ההחלטות שנעשית על סמך מודל רגרסיה לוגיסטית בלבד.
2. הטיה: כמה הטיה קיימת למודל לפני ואחרי התערבות ה-human-in-the-loop.
3. השפעה שילוב HITL במערכת: ההשפעה של התערבות האדם בהוגנות של המערכת והטיות חדשות שנוצרות כתוצאה מהתערבות זו.

When

ה-audit יתבצע במספר שלבים:

1. לפני הטמעת שכבת human-in-the-loop (אקס אנטה): בחינת הצורך בהתערבות האנושית על מנת למזער סיכונים הנשקפים לזכויות הפרטים המושפעים מהמערכת, בכדי להשיג תוצאות מיטביות.
2. לאחר הטמעת שכבת human-in-the-loop (אקס פוסט): האם ההתערבות האנושית אכן השיגה את מטרתה והייתה מועילה דה פקטו באופן שבו היא יושמה? זהו השלב שבו ניעשת הבחינה העיקרית, שכן ידוע שהמודל מפלה נשים מלכתחילה, ויבחנו היתרונות והחסרונות של הוספת גורם אנושי ניטרלי.

d. How: methods, tools, metrics, standards

Methods:

1. Data Analysis: אנליזה סטטיסטית של מדגם האימון כדי לזהות גורמים אפשריים להטיה.
2. ניתוח תוצאות: ניתוח התוצאות המופקות על ידי מערכת הבינה המלאכותית כדי לזהות פערים בתחזיות ובהחלטות בין קבוצות דמוגרפיות שונות.
3. סימולציה ובדיקה: הדמיית תרחישים שונים כדי לראות כיצד שינויים בנתוני הקלט משפיעים על הפלט של המודל.

Tools:

1. כלים כמו Aequitas or Fairness Indicators (של OECD.AI) למדידת הטיות והוגנות בפרדיקציות של המודל.
2. AI Fairness 360 open source toolkit - יכול לעזור לבחון ולהפחית אפליה והטיה במודלים של למידת מכונה.
3. What-If tool: כלי מבית גוגל, משולב עם TensorBoard. מאפשר לבצע תרחישי "מה-אם" שונים, כדי לעזור לזהות הטיות אפשריות.

Metrics:

1. Equal opportunity - true positive rate: נמדוד את ההבדלים של TPR בין קבוצות שונות, TPR נמוך יותר עבור קבוצה מסוימת מצביע על כך שהמערכת פחות יעילה בזיהוי מועמדים מוסמכים מאותה קבוצה, מה שמוביל לשיטות גיוס בלתי הוגנות.
2. False negative rate: אם ה-FNR גבוה יותר עבור קבוצה אחת, זה מצביע על כך שהמערכת מענישה את אותה קבוצה בצורה לא הוגנת, מה שמוביל להתעלמות ממועמדים מתאימים, מה שיכול למנוע ממועמדים ראויים להתקבל לעבודה.
3. Equalized odds - FPR: אם ה-FPR גבוה משמעותית עבור קבוצה אחת בהשוואה לקבוצה אחרת, זה מצביע על כך שהמערכת מעדיפה בצורה לא הוגנת את אותה קבוצה על ידי מתן יותר תוצאות חיוביות כוזבות, מה שעלול להוביל להעסקת מועמדים לא כשירים מאותה קבוצה.

Standards:

1. **סעיף 2(א) לחוק שוויון הזדמנויות בעבודה**- האם המערכת מתחשבת בשיקולים שאינם רלוונטים לביצוע העבודה במסגרת ההחלטה האם לקבל מועמד או לא?
2. **A Local Law to amend the administrative code of the city of New York, in relation to automated employment decision tools (2021/144**
   a. Automated Employment Decision Tools (AEDT), שעניינו קביעת כללים לכלי בינה מלאכותית המשמשים בשוק העבודה.
   b. חובת בדיקה: על פי החוק, מעסיקים המעוניינים להשתמש בכלים אוטומטיים לסינון מועמדים מחויבים בעריכת בדיקה עצמאית בטרם השימוש לוודא כי התוכנה איננה מכילה הטיה נגד קבוצות אוכלוסייה מוגנות, כגון על בסיס גזע, מגדר, לאום וכדומה.
   c. חובת פרסום: על פי החוק, מעסיקים המשתמשים בכלים אוטומטיים לסינון מועמדים מחויבים לפרסם, על בסיס חודשי, את תוצאות בדיקת ההטיה שערכו לכלים אלו, באתר האינטרנט שלהם.
   d. שיעור הבחירה ואת יחס ההשפעה של התוכנה/הכלי.
3. **OECD-** מציינים כלל אצבע לפיו אם למבחן גיוס יש שיעור בחירה נמוך מ-80% לקבוצות מוגנות בהשוואה לאוכלוסייה הכללית ניתן לקבוע שמתקיימת הטיה פסולה
4. פסיקה רלוונטית הנוגעת בחובת תום הלב ואיסור אפליה (מצורף לעיל).
5. מסמך עקרונות מדיניות רגולציה ואתיקה של משרד המשפטים (2023) -
6. **GDPR-**
   a. הערכת השפעות המערכת לפי סעיף 35 (7) בGDPR- הערכת פעולות העיבוד הצפויות ומטרתן, נחיצות ומידתיות בין הפעולה למטרה, הערכת סיכונים לזכויות וחירויות של הפרטים המושפעים מהמערכת, האמצעים שיינקטו להפחתת הסיכונים. במידת הצורך, יש להתחשב גם בעמדתם של קהל היעד של פעולות העיבוד או נציגיהם.
   b. דרישת מנגנון המאפשר התערבות אנושית (ס׳ 22 (1) בGDPR, ולאדם המושפע מההחלטה לערער עליה ולבטא את נקודת מבטו (ס׳ 22 (3) GDPR)