סיכום המאמר – Exploring the Impact of Fault Justification in Human-Robot Trust - Authors: Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S. Melo, Ana Paiva.

שם הכנס: AAMAS 2018
קישור: http://ifaamas.org/Proceedings/aamas2018/pdfs/p507.pdf

הבעיה: רובוטים אוטונומיים הם דבר המועד לכישלונות ולעיתים בני אדם מאבדים אמון ברובוטים לאחר השגיאה הראשונה שנראית בעבודתו של הרובוט בעבודה משותפת.
מאמר זה בא לבדוק האם ניתן לפתור בעיה זו ע"י זה שהרובוט ייתן משוב לבן אדם על כישלונו ו"יצדיק" את הסיבה לכך.
בעיה זו הינה בעיה קריטית מכיוון שכאשר מפתחים שותפים רובוטיים אנחנו רוצים שאנשים ייתנו אמון בהם ושיהיה להם נוח לשתף פעולה איתם, בנוסף לכך אם הייתה תקלה כלשהי אנחנו מעוניינים שאותו אדם לא יפסיק להיעזר ברובוט בשל חוסר אמון.

תחום המחקר של מצבי שגיאה באינטראקציה בין אדם לרובוט עדיין חדש. ולכן נכון לעכשיו קיימות 3 שאלות רחבות שמקבלות את מלוא תשומת לב החוקרים שחקרו זאת בעבר:

1. איך רובוט יכול באופן אוטומטי לזהות מצבי שגיאה?
2. כיצד מצבי השגיאה משפיעים על האינטראקציה עם האדם ועל התפיסה של האדם?
3. אילו אסטרטגיות ניתן לנקוט כדי להקל על ההשפעות של כישלון?

לגבי אסטרטגיות אפשריות שרובוטים יכולים להשתמש בהם לאחר מצבי שגיאה ישנם 2 מחקרים מקוונים שמדווחים שאסטרטגיות התאוששות אכן יכולות למתן את ההשפעות השליליות של כשלים רובוטיים. למשל לי ועמיתיה הראו שאסטרטגיית ההתנצלות הייתה יעילה ביותר כדי להקל על תפיסות של יכולת, קרבה וחביבות אצל רובוט שירות. עם זאת המחברים הראו גם שהאוריינטציה של אנשים לשירותים עשויה להוביל להשפעות שונות לאסטרטגיות ההתאוששות שנצפו.

בסקר מקוון דומה, ברוקס ועמיתיו חקרו את תגובתם של אנשים לכישלונות ברובוטים אוטונומיים כלומר שואב אבק ומונית שנוהגת עצמאית, ע"י מניפולציה של 4 משתנים: סיכון הקשר, חומרת הכישלון, תמיכה במשימות ותמיכה אנושית. תפיסת המשתתפים על הרובוט הפגום נעשתה פחות שלילית כאשר הוא השתמש באסטרטגיות מיתון ע"י תמיכה במשימות, תמיכה אנושית או שתיהן. עם זאת המחברים דיווחו על נטייה מעניינת אך לא משמעותית, המעידה על העדפה של תמיכה במשימות ותמיכה אנושית במצבים חמורים, וכן

העדפה לתמיכה במשימות רק במקרים חמורים פחות. ברוקס ועמיתיו תרמו לתוצאות הקודמות של לי ועמיתיה עם הרעיון שכמות ההשפעה של האסטרטגיה על תגובת האנשים תלויה בסוג המשימה, בחומרת הכישלון והסיכון לכישלון.

אחד החסרונות של עריכת סקר מקוון כדי להבין מהי התפיסה על רובוט פגום הוא שהמשתתפים מתפקדים רק על בסיס משקיפים. ככאלה הם אינם מושפעים ישירות מכישלונות הרובוט. מאמר זה נמנע מבעיה זו בכך שהמשתתפים במחקר זה מדרגים את תפיסתם על הרובוט הפגום לאחר אינטראקציה עם הרובוט כשותף ומושפעים ישירות מהתנהגותו הפגומה. היבט נוסף שנבחן במאמר זה הוא השפעות ההקלה של אסטרטגיות התאוששות שלא נחקרו, כלומר, הרובוט "מצדיק" את התקלה. לבסוף הרובוט במאמר פועל באופן אוטונומי ומדמה התאוששות אוטונומית של כישלון טכני, אשר נבדל מכישלון להשיג מטרה או הצגת התנהגויות שגויות במהלך האינטראקציה.

הניסוי שהתבצע במאמר זה כלל 3 שלבים:
<u>שלב ראשון:</u> לקחו מספר משתתפים שלא מכירים אחד את השני ונתנו להם למלא שאלון התחלתי לגבי הצפיות שלהם מהרובוט לפני האינטראקציה שלהם איתו.
<u>שלב שני :</u> נתנו למשתתפים לשחק ב - 3 משחקי "טנגרם" עם NAO רובוט על מסך מגע. לרובוט הכניסו כשלים בכוונה, לחלק מהמשתתפים נתנו הצדקה לכישלון ולחלק מהם לא, חלק מהכישלונות גרמו לעצירת המשימה ולעומת זאת בחלק מהם היה ניתן עוד להמשיך את המשימה.
<u>שלב שלישי:</u> נתנו למשתתפים לענות על אותו שאלון כמו בהתחלה לאחר שסיימו לשחק.

כעבודה לעתיד הם מתכננים לפתח רובוט שבאמצעות מודלי הסתגלות הוא ישנה את אסטרטגיית ההתאוששות שלו בהתאם לחומרת הכישלון. בעניין סוג הכישלון, הם שואפים לחקור את השימוש של כשלים הקשורים למשימה במקום רק כישלון טכני. גם יעניין אותם לחקור מה תהיה ההשפעה על האמון באמצעות אסטרטגיות התאוששות בעלות תוכן רגשי.

מאמר זה עניין אותי מאוד מכיוון שהתפתחות התחום תלויה ברמת האמון שאנו נותנים ברובוטים, ככל שהזמן עובר אנחנו משתפים יותר פעולה עם רובוטים למשל בתחום חקר החלל מי שלמעשה חוקר את הכוכבים סביבנו אלו הרובוטים ולכן ככל שנצליח לפתח רובוטים שהשפעתם עלינו היא פחות שלילית נגרום לכך שפיתוח הרובוטים יגבר.
מאמר זה העניק תוצאה ממשית אך ורק למקרה הספציפי שבו אסטרטגיית ההתאוששות של הצדקת הכישלון עובדת רק כאשר תוצאת הכישלון הייתה פחות חמורה ובגלל שתחום מחקר זה עדיין חדש יש עוד הרבה היבטים לבדיקה.