On the Influence of Autonomy and Transparency on Blame and Credit in Flawed Human-Robot Collaboration

Alexander Arntz
alexander.arntz@hs-ruhrwest.de
University of Applied Sciences Ruhr West
Bottrop, North-Rhine Westphalia, Germany

Sabrina C. Eimler
University of Applied Sciences Ruhr West
Bottrop, Germany
sabrina.eimler@hs-ruhrwest.de

Carolin Straßmann
University of Applied Sciences Ruhr West
Bottrop, Germany
carolin.strassmann@hs-ruhrwest.de

H. Ulrich Hoppe
University of Duisburg-Essen
Duisburg, Germany
hoppe@collide.info

ABSTRACT
The collaboration between humans and autonomous AI-driven robots in industrial contexts is a promising vision that will have an impact on the sociotechnical system. Taking research from the field of human teamwork as guiding principles as well as results from human robot collaboration studies this study addresses open questions regarding the design and impact of communicative transparency and behavioral autonomy in a human robot collaboration. In an experimental approach, we tested whether an AI-narrative and communication panels of a robot-arm trigger the attribution of more human like traits and expectations going along with a changed attribution of blame and failure in a flawed collaboration.

CCS CONCEPTS
• Human-centered computing → Empirical studies in collaborative and social computing.

KEYWORDS
human-robot collaboration; attribution of blame; perception of intelligence; online study

ACM Reference Format:
Alexander Arntz, Sabrina C. Eimler, Carolin Straßmann, and H. Ulrich Hoppe. 2021. On the Influence of Autonomy and Transparency on Blame and Credit in Flawed Human-Robot Collaboration. In Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI ’21 Companion), March 8–11, 2021, Boulder, CO, USA. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3434074.3447196

1 INTRODUCTION
Industrial usage of Human-Robot Collaboration (HRC) is expected to be enhanced by advanced artificial intelligence (AI) technology in the future, enabling robots to operate either partially or fully autonomously in conjunction with the employees [22]. Whereas current implementations of HRC in industrial settings often use non-humanoid robots (e.g. robot-arms) controlled by an operator or following predefined routines, an autonomous AI-driven robot, capable of adapting and reacting in the required task, would embrace the full potential of the HRC concept [2]: Human personnel could be relieved by delegating repetitive or heavy work onto the robot, while contributing through intuition and experienced-based decision-making, thus combining the advantages of both parties [11]. However, besides the technical and safety challenges to be solved, there are reservations against AI and robots in large parts of the general population [29] that need careful investigation. Prior research in the realm of e.g. the Media Equation Theory [27] and CASA [26] has shown that individuals tend to project human characteristics onto robots while interacting with them and that this even holds true for technologies with non-humanoid appearance [9, 28]. Therefore, it can be assumed that principles and results obtained in research on group collaboration might be applicable for HRC as well.

Comparable to collaboration among human personnel, the collaboration with an autonomous robot to achieve a common goal creates an interdependency in the work relation [8]. Errors in such a context can hamper the successful outcome of the procedure. Considering they might lead to costly or hazardous ramifications for the human collaboration partner, it is of interest to investigate if an autonomous robot will be held accountable for an error and which characteristics along with behavior can influence people’s attribution of blame. Research that examined the accusation of errors in non-work-related, casual scenarios [24], revealed that more autonomy displayed by a robot-arm results in an increased attribution of errors whereas comprehensibility/ transparency of the robot’s actions leads to a decrease in blame [17]. However, the attribution of accountability and credit made in HRC-workplace environments involving autonomous robots remains an open question. Also, reservations towards AI need further exploration to design successful and accepted HRC scenarios, especially attributions relevant in collaborative tasks (e.g. manufacturing procedures) are worth being researched.

To explore this, we used the virtual reality sandbox application by [5] that is capable of simulating a variety of industrial HRC scenarios difficult to realize under experimental conditions [20]. The environment contained a robot-arm with autonomous behavior serving as a basis for an online study testing using non-interactive videos, in a collaborative task, the influence of a) the robot-arms
behavioral autonomy and b) its transparency in communication on the attribution of blame and credit.

2 RELATED WORK

Research addressing the attribution of blame and credit is well established in the field of human group collaboration. The self-serving bias in attribution has been identified as the main contributing factor in people’s assessment of outcomes [16]. Two types can be distinguished: the internal attribution, that includes the own characteristics of an individual, and the external attribution, containing outside influences [12]. Internal attribution is often associated with successful outcomes whereas people incline to apply external attribution to poor outcome [23]. Studies involving Human-Robot Interaction showed that this behavior also occurs when people engage with robots [10, 13, 17]. This misattribution can negatively affect trust in the robot’s capability to accomplish a task, thus mitigating the collaboration process [15]. While the self-serving bias in attribution provides a strong foundation, contrary to prior studies, Lei and colleagues’ participants attributed more credit and less blame to the robot [19]. The studies used divergent representations of robots with different levels of autonomy and communication, which might have contributed to the inconsistent results. Accordingly, the design of the robot has an effect on the described attribution process (compare [17]). To design robots to be the best possible collaboration partners and bypass distrust [15], research is needed to understand which characteristics drive the attribution of blame in the collaboration with robots.

2.1 The Effect of Expected Autonomy

Kim and colleagues [17] demonstrated an influence of the robot’s autonomy on the attribution of blame. A robot that presented more autonomy was more likely blamed for an undesirable outcome. This is specifically of interest in industrial collaboration settings, where employees often have misconceptions and negative attitudes towards autonomous systems [29]. Especially the term AI is associated with negative feelings [1], since people are afraid that robots with AI-capabilities will take their jobs. As a consequence of this widespread misconceptions about AI among the general population, people tend to attribute various forms of human-like characteristics and behavior towards such systems [21]. This is also plausible against the background of CASA [26]. It is therefore assumed that the introduction of the term AI along with the autonomous behavior by a robot-arm will invite participants to project more human-like abilities and behavior onto the system [30]. As a result, people will use an external attribution and blame the robot with higher expected autonomy more for errors and negative outcomes of a collaboration process. Accordingly, the following hypotheses are assumed:

H1: Participants attribute more human-like abilities (intelligence, morality) to a robot-arm with AI-capabilities compared to one without.

H2: Participants attribute more blame and less credit to a robot-arm with AI-capabilities compared to one without.

2.2 The Effect of Communication and Transparency

Beside the perceived autonomy and intelligence of the robot, transparency was found to affect the attribution of blame [17]. A robot that explains it’s own behavior was found to evoke lower attributions of blame. This might also an explanation for the results of Lei and colleagues, since they tested the attribution of the talking humanoid robot NAO [19]. Communicative behavior that elicits transparency of the robot’s behavior, seems to prevent external attribution. In industrial settings robot-arms are often limited in their communicative abilities. As the environment is often loud, verbal outputs would not work. Studies therefore suggest to use text-panels to enrich the communicative output of the robot. Making the robot’s behavior transparent to the human collaborator by augmenting the robot with communication capabilities was found to result in various benefits, e.g. perceived stress and general, positive emotions on the side of the human collaborator [4, 5]. Accordingly, we assume that an communication panel affects the perception of the robot as collaboration partner and thereupon leads to fewer external attributions of errors. This leads to the following hypotheses:

H3: Participants perceive a robot-arm equipped with a communication panel as better collaboration partner (more cooperative and better quality of the collaboration) than one one without communication ability.

H4: Participants attribute less blame and more credit to a robot-arm equipped with a communication panel compared to one without.

As described above, studies involving communicative robots often use voice output that distinctly link the statements to the respective robot [17, 19]. However, industrial robots with non-humanoid appearances in extremely loud environments demand different communication channels. A prior study recognized text-panels in natural language as a viable means of communication in industrial HRC settings [3]. Results indicated that proximity and visual relation to the robot-arm are decisive aspects, since the external text statements are not as intuitively assignable to the robot. Only when the communication behavior (text-panel) is assigned to the robot, it is plausible that this affects the attribution process of errors and robot perception. It is therefore of interest if participants associate the text-panels to the robot-arm or whether it is perceived as an other autonomous entity. Thus, the following research question is to be answered:

RQ1: Do participants see the text-panel augmentations as part of the robot-arm or as another autonomous entity?

3 METHOD

For the study, we used a virtual reality simulation of a HRC shared-task setup as described by [3]. Since the pandemic made a VR-lab experiment impossible, an online experiment was set up, in which participants were presented a first-person perspective video of the HRC-setup. In a 2 (augmented communication vs. non-augmented condition) x 2 (AI-narrative vs. non-AI-narrative) between-subjects design participants were asked to imagine themselves in the role of the human worker assigned to the HRC working arrangement. A total of 225 participants took part in the online study. Participants
under 18, an outlier aged 99 as well as participants completing the experiment too fast or too slow (i.e. deviating 1.5 standard deviations from mean completion time) were excluded. Altogether 34 were sorted out. The average age of the remaining 191 participants was \( M = 25.12 (SD = 7.51) \). 91 were female, 97 male and 3 non-binary people participated.

3.1 Material

Participants were exposed to one of the four conditions. In all conditions participants a virtual representation of a LBR iiwa 7 R800 CR robot-arm was displayed that used multi colored light-signals and action initiating/terminating and standby gestures [5, 18, 25]. In the high-transparency condition (= augmented condition) text-panels in natural language were used to express guidance and explanations. In the low transparency condition (= non-augmented condition), the explanatory and guiding text-panels were omitted, while everything else to be witnessed in the procedure was identical in terms of movement and actions by the robot and the human. The purpose for removing the text-panel, was to withhold the explanation of the robot-arm’s behavior provided by the text-panel but retain the other communication methods. This maintained the robot-arm’s ability to convey a detected error but obscure the system’s interpretation of the error. To manipulate different levels of autonomy, participants were either told that the robot-arm has AI-capabilities or the scene was just depicted as a collaboration between a human worker and a robot-arm. In all conditions, participants witnessed, from a first-person perspective, a simulated shared-task in which they, as the human operator, were tasked to manufacture metal buttons through a press with their robot collaboration partner [5].

During the procedure, both partners deviated from an assembling task (2 items, \( \alpha = .820 \)) and to the self for the errors during the assembling task (2 items, \( \alpha = .820 \)), as well as regarding the attribution of credit to the robot (2 items, \( \alpha = .616 \)) and the self for task completion (2 items, \( \alpha = .625 \)) [17]. The Perceived Moral Agency scale by [6] was used to assess morality (6 items, \( \alpha = .763 \)) and dependency (4 items, \( \alpha = .683 \)). Embodiment of the robot-arm was assessed through the EmCorp-Scale [14], containing the sub-scales: corporeality (3 items, \( \alpha = .685 \)), expressiveness (4 items, \( \alpha = .701 \)), tactile interaction & mobility (6 items, \( \alpha = .539 \)) and perception & interpretation (7 items, \( \alpha = .765 \)). To analyze anthropomorphism (5 items, \( \alpha = .592 \)), animacy (5 items, \( \alpha = .661 \)), likeability (5 items, \( \alpha = .815 \)) and perceived intelligence (5 items, \( \alpha = .761 \)) of the robot-arm, the questionnaire incorporated the Godspeed-scale by [7]. Moreover, the collaboration success was measured with an ad-scale consisting of 6 items (\( \alpha = .795 \)). The assessment of the components assigned to the robot was realized through screenshots of the application, where every visible item was highlighted through a bounding box. For each component participants were asked to decide whether or not it belonged to the robot. The questionnaire closed with demographics (e.g. age, gender, job position, educational background).

4 RESULTS

To test the hypotheses, multiple analyses of variance (MANOVA and ANOVA) were run including the relevant independent and dependent variables for testing.

H1: Participants attribute more human-like abilities (intelligence, morality) to a robot-arm with AI-capabilities compared to one without.

The analysis did not indicate significant differences in the attribution of intelligence and perceived morality between both AI-conditions. Thus, the AI-narrative did not lead to higher perceived intelligence or higher attribution of moral capabilities. The robot-arm was rated significantly better in the ability to perceive and interpret its surroundings (\( F(1,187) = 5.70, p = 0.018, \eta^2_p = 0.03 \)) in the AI-narrative condition (\( M = 2.37, SD = .66 \)) compared to the non-AI-narrative (\( M = 2.14, SD = .73 \)). Moreover, the capacity for cooperation is rated significantly better (\( F(1,187) = 6.47, p = 0.012, \eta^2_p = 0.03 \)) in the AI-narrative condition (\( M = 3.91, SD = .92 \)) compared to the non-AI-narrative (\( M = 3.55, SD = 1.11 \)). Furthermore, the robot-arm was rated as significantly less dependent on predefined programming (\( F(1,187) = 5.92, p = 0.016, \eta^2_p = 0.03 \)) in the AI-narrative condition (\( M = 4.22, SD = .74 \)) compared to the non-AI-narrative (\( M = 4.46, SD = .60 \)). Although the AI-narrative did not lead to more perceived intelligence and morality, these results indicate that participants associate more human-like characteristics to the robot-arm like being independent, cooperative and able to perceive and interpret. Thus, H1 is partly supported.

H2 & H4: Participants attribute more blame and less credit to a robot-arm with AI-capabilities compared to one without and participants attribute less blame and more credit to a robot-arm equipped with a communication panel compared to one without.

Analyses testing this did not show any differences between the conditions for neither the attribution of blame nor for credit attribution. Therefore, H2 and H4 could not be supported.

H3: Participants perceive a robot-arm equipped with a communication panel as better collaboration partner (more cooperative and better quality of the collaboration) than one without communication ability.

Results show significant differences (\( F(1,187) = 5.95, p = 0.016, \eta^2_p = 0.003 \)) between the augmented condition (\( M = 3.57, SD = .62 \)) and the non-augmented condition (\( M = 3.33, SD = .76 \)) for perceived intelligence. Significant differences occurred for dominance (\( F(1,187) = 8.67, p = 0.004, \eta^2_p = 0.04 \)). The robot-arm in the augmented condition was perceived as more dominant (\( M = 3.37, SD = .94 \)) than in the non-augmented condition (\( M = 2.98, SD = .84 \)). A significant difference was found for collaboration success (\( F(1,186) = 5.34, p = 0.002, \eta^2_p = 0.03 \)), augmented condition (\( M = 2.80, SD = .72 \)) vs. non-augmented (\( M = 3.05, SD = .77 \)). H3 is supported.
This study explored the effect of transparency (i.e. communicative association with the robot-arm. Indeed participants perceived the text-panel as belonging to the robot/ part of the robot-arm. Accordingly, the presence of text-panels as a means for the robot to increase communicative transparency lead to higher attributions of intelligence, dominance and the perception of a more successful collaboration (H3). Although participants perceived the collaboration more successful in the augmented version of the robot-arm, the text-panel did not affect the evaluation of the robot-arm’s perceived cooperativeness. In contrast, the augmented robot-arm was perceived as more dominant. This results could produce conflicts, since the perception of dominance elicits negative feelings in the human collaborator. Especially people with fears and negative attitudes might feel patronized and avoid collaborating with the robot-arm. Thus, designers have to use communication features inducing transparency with caution, since they could trigger a boomerang-effect.

While the perception of the robot-arm was affected by the induced autonomy and transparency (H1, H3), no significant difference regarding blame and credit was found (H2, H4). While other studies showed that the self-serving bias in attribution cannot always be demonstrated in interactions with robots [19], the projected ramification that is expected by the individual [16] must be considered as a limitation of this study.

Due to the restrictions of the COVID-19 pandemic we were unable to conduct an experimental study where participants could actually collaborate with robot-arm in the virtual reality scenario. While the virtual reality sandbox application provided by [5] enabled us to substitute an online study design, no direct interaction with the robot-arm was possible. The immersive effect of the environment might be able to create a sense of more direct involvement with a higher sensitivity for the outcomes of the errors happening. Employees exposed to autonomous robots in industrial HRC settings could face real consequences from errors made during the collaboration e.g. injury or career disadvantages that witnessing a video cannot fully mimic. Future work should address the used scenario using an interaction study to overcome these limitations. Also, future studies should investigate if additional communication augmentations (e.g. voice output) and inputs provided by the text panels affect attributions differently.

6 CONCLUSION
While this study could not replicate established findings from the literature regarding the attribution of blame and credit, results reveal an interesting effect regarding the attribution of human characteristics on the robot-arm caused by an AI-narrative that are worth being further explored. Future studies should look into the dynamics of people’s mental models and preconceptions brought into the collaboration scenario with AI-based robots. Communication behavior should be outbalanced in a way that it does not trigger dominance perception on the one hand, but elicits enough transparency to induce trust in the collaboration partner on the other hand.

RQ1: Do participants see the text-panel augmentations as part of the robot-arm or as another autonomous entity?

Analyzing the components that were associated with the robot, no significant differences were observed between the AI-narrative condition vs. non-AI-narrative condition. A heat map revealed that participants from all conditions identified the body of the robot-arm (Fig 1). In addition, 93.3% of the participants from the text-panel condition associated the text-panel as part of the collaboration partner. Also, 73.2% of the participants associated the information text-panel as belonging to the robot/ part of the robot-arm. As stated by [30], the wide spectrum of the term AI together with the widespread misconceptions invites people with media biased knowledge to project numerous abilities and expectations onto AI-enhanced systems [29]. Future studies should explore the content of the mental models and expectations and their effects on the collaboration process.

RQ1 investigated which components are attributed to the robot-arm to ensure the statements displayed on the text-panel is associated with the robot-arm. Employees exposed to autonomous robots in industrial HRC settings could face real consequences from errors made during the collaboration e.g. injury or career disadvantages that witnessing a video cannot fully mimic. Future work should address the used scenario using an interaction study to overcome these limitations. Also, future studies should investigate if additional communication augmentations (e.g. voice output) and inputs provided by the text panels affect attributions differently.

6 CONCLUSION
While this study could not replicate established findings from the literature regarding the attribution of blame and credit, results reveal an interesting effect regarding the attribution of human characteristics on the robot-arm caused by an AI-narrative that are worth being further explored. Future studies should look into the dynamics of people’s mental models and preconceptions brought into the collaboration scenario with AI-based robots. Communication behavior should be outbalanced in a way that it does not trigger dominance perception on the one hand, but elicits enough transparency to induce trust in the collaboration partner on the other hand.
