Understanding User Attitudes Towards Negative Side Effects of AI Systems

SANDHYA SAISUBRAMANIAN, College of Information and Computer Sciences,
University of Massachusetts Amherst, USA
SHANNON C. ROBERTS, Department of Mechanical & Industrial Engineering,
University of Massachusetts Amherst, USA
SHLOMO ZILBERSTEIN, College of Information and Computer Sciences,
University of Massachusetts Amherst, USA

Artificial Intelligence (AI) systems deployed in the open world may produce negative side effects—which are unanticipated, undesirable outcomes that occur in addition to the intended outcomes of the system's actions. These negative side effects affect users directly or indirectly, by violating their preferences or altering their environment in an undesirable, potentially harmful, manner. While the existing literature has started to explore techniques to overcome the impacts of negative side effects in deployed systems, there has been no prior efforts to determine how users perceive and respond to negative side effects. We surveyed 183 participants to develop an understanding of user attitudes towards side effects and how side effects impact user trust in the system. The surveys targeted two domains: an autonomous vacuum cleaner and an autonomous vehicle, each with 183 respondents. The results indicate that users are willing to tolerate side effects that are not safety-critical but prefer to minimize them as much as possible. Furthermore, users are willing to assist the system in mitigating negative side effects by providing feedback and reconfiguring the environment. Trust in the system diminishes if it fails to minimize the impacts of negative side effects over time. These results support key fundamental assumptions in existing techniques and facilitate the development of new methods to overcome negative side effects of AI systems.

Additional Key Words and Phrases: Artificial intelligence systems, Negative side effects, Case study

ACM Reference Format:
Sandhya Saisubramanian, Shannon C. Roberts, and Shlomo Zilberstein. 2021. Understanding User Attitudes Towards Negative Side Effects of AI Systems. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3411763.3451654

1 INTRODUCTION

Artificial Intelligence (AI) systems are increasingly deployed in complex real-world settings. These systems are rarely perfect and may cause *negative side effects* (NSE) during their operation. Negative side effects are the unanticipated, undesirable effects that occur because the AI system's objective focuses on one aspect of the environment but its operation impacts additional aspects of the environment. For example, an autonomous vehicle that optimizes travel time may not slow down when driving through potholes. This may result in a bumpy ride for the user, which is an undesirable side effect. Another example of a side effect is an autonomous vacuum cleaner spraying water on the walls when cleaning the floor. The severity of such side effects range from mild events to safety-critical failures. The severity of negative side effects depends on factors such as the system capabilities, its assigned task, and the setting in which the

Manuscript submitted to ACM

system is deployed. The side effects may violate user preferences or alter the environment in a manner that affects safety. It is inherently challenging to identify all negative side effects during the AI system development cycle, when the system is deployed in diverse settings. As a result, NSE are often identified after the system is deployed. Unanticipated domain characteristics, cultural differences among the target users and development teams, or unanticipated consequences of system or software upgrade are common causes of NSE [18]. In the autonomous vehicle example, details such as the undesirability of a bumpy ride may be overlooked during the system design as human drivers naturally slow down when driving through potholes, even when optimizing travel time. However, unless explicitly specified, knowledge about such NSE is generally unavailable to the AI system.

Overcoming NSE is an emerging area that is attracting increased attention among AI researchers [1, 6, 7, 11, 12, 15–20, 22]. Recent works on techniques to overcome NSE [6, 16, 17, 19, 22] make various assumptions about user preferences and their (in)tolerance of NSE in order to develop practical solutions to the problem. To the best of our knowledge, however, there are no published reports on how users respond to NSE, their willingness to tolerate NSE, and how NSE affects their trust in the AI system. These factors are critical in evaluating existing solutions and developing new approaches that are realistic and deployable in the real-world. User tolerance of NSE depends on many factors such as their individual preferences and the severity of the side effect. When the NSE are safety-critical, it is clear that users will not tolerate them and system's operation needs to be suspended to address the NSE and reevaluate the system performance. In many deployed systems, however, the impacts of NSE are significant but not catastrophic, and such side effects are sometimes overlooked in discussions of reliable and trustworthy AI.

In this work, we present results from initial user studies in two domains to understand user attitudes and preferences to NSE that are undesirable but not safety-critical. We aim to answer the following questions through these user studies: (1) are users willing to tolerate negative side effects that are not safety-critical? (2) how do negative side effects affect the user's trust in the system? (3) are users willing to assist the system in mitigating the impacts of the side effects—by providing feedback, applying minor changes to the environment, or specifying regions where the system can operate? and (4) are users willing to tolerate a sub-optimal behavior of the AI system (such as taking a longer route) in order to avoid negative side effects? Answering these questions will deepen our understanding of the side effects problem, validate key assumptions used by existing techniques, and shape future research directions on this topic.

2 RELATED WORK

As we see accelerated deployment of AI systems, addressing their negative side effects is emerging as an important research area in AI [11, 12, 16–18, 20, 22]. Inconsistent and unpredictable system behavior, some of which may be unsafe, affects user trust in the system's capabilities and operation. If the impacts of the undesirable behaviors are significant, it can also lead users to abandon the system. In fact, studies show that users may stop trusting a system after witnessing a mistake, even if the system outperforms humans in the task [2]. Hence, mitigating NSE is critical in shaping how users view, interact, collaborate, and trust AI systems. Existing works on this topic have focused on developing techniques to efficiently recognize and overcome the impacts of NSE by updating the system behavior. However, there has been no prior efforts to understand user attitudes towards NSE.

Some user surveys have been conducted to understand how users interact with self-driving cars [8, 13] and autonomous vacuum cleaners [5]. These studies highlight the concerns and promise of these technologies, and how they are perceived by users from different backgrounds. Recently, researchers have investigated the effect of accuracy on user expectations and trust in machine learning models [9, 21]. These results show that user trust in the system diminishes when the observed accuracy is lower, regardless of its stated accuracy. While these studies provide a broad overview of

user expectations and trust in AI systems, they do not provide specific insights on the negative side effects problem. Since this is an emerging topic, a survey conducted specifically to identify general user attitudes towards NSE is critical to develop effective solutions to this practical problem.

3 METHODS

Domains. We conducted two IRB-approved surveys that focused on NSE in two domains: An autonomous vacuum cleaner (Roomba) and an autonomous vehicle (AV). We considered NSE such as the Roomba spraying water on the wall when cleaning the floor, the AV driving fast through potholes which results in a bumpy ride for the users, and the AV slamming the brakes to halt at stop signs which results in sudden jerks for the passengers. Roomba domain represents a setting where the NSE is relatively mild, the users do not directly experience the NSE, and the system does not require constant supervision when it is performing its task. The AV domain represents a setting in which the NSE have moderate impact, the users experience the NSE directly (bumpiness or sudden jerk), and users generally supervise the AV performance and can take control when issues related to safety arise.

Participants. We recruited 500 participants on Amazon mechanical turk to complete a pre-survey questionnaire to assess their familiarity with AI systems and fluency in English. This questionnaire has six questions and takes less than 30 seconds to complete. All participants were informed about the purpose of the study. Based on the pre-survey responses, we invited 300 participants aged above 30 to complete each survey (Roomba and AV). We selected based on the age criteria since study shows that participants aged above 30 are less likely to game the survey conducted on the Mturk platform [4]. The surveys generally take less than ten minutes to complete. Responses that were incomplete or with a survey completion time of less than one minute were discarded. We received a total of 204 valid responses for the Roomba domain and 183 valid responses for the AV domain. To facilitate a direct comparison between the responses in both the domains, we randomly sampled 183 responses for the Roomba domain.

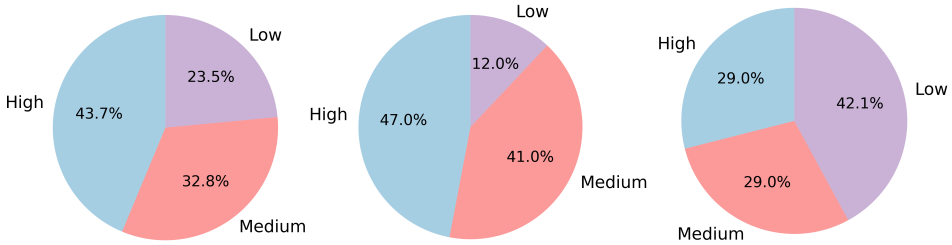
4 SURVEY DESIGN

The survey questionnaires contained similar questions for the two domains, with ten questions for the Roomba domain and eleven questions for the AV domain. The questions focused on user tolerance, trust, willingness to tolerate sub-optimal behavior so as to mitigate NSE, and various forms of human assistance. The questions included a description of NSE and participants were required to select an option that best describes their attitude. We study user tolerance of two forms of NSE in the AV domain: bumpiness and sudden jerks. This is to understand the effect of severity of NSE on user tolerance. All other survey questions on the AV domain focused *only* on the bumpy ride side effect.

User Tolerance. For each domain, the participants were required to indicate their level of tolerance of NSE: *low*—indicating their unwillingness to use the AI system due to its NSE; *medium*—indicating the system will be used less frequently due to its NSE; and *high*—indicating their willingness to continue using the system, despite NSE.

Trust. To determine if NSE affected user trust in the system's capabilities, we asked participants to select an option that best describes their trust level: *low*—do not trust the system to be capable of completing its task; *medium*—trust is affected if the system does not learn to avoid NSE over time; and *high*—trust is unaffected by NSE. We consider this simple categorization to understand how NSE may affect the system usability.

Slack Preferences. In many instances, NSE can be avoided if the system is allowed to act sub-optimally with respect to its assigned task. For example, the bumpy ride—which occurs when the AV drives fast through potholes as a result of optimizing travel time—may be avoidable if the AV takes a longer route or navigates at a lower speed. For the AV



(a) Roomba (b) AV mild (bumpy) (c) AV moderate (jerks)

Fig. 1. User tolerance of negative side effects.

domain, we considered a fixed 25% slack based on results in [16]. This slack allows the AV to drive slow or pick an alternate route, which takes up to 25% longer to reach the destination. Similarly, the slack for the Roomba domain allowed it to skip cleaning the area within five inches from the wall. Slack for the Roomba domain can be considered as allowing the system to not complete its task fully, while the slack for the AV allows it to take longer to complete the task. Participants were required to select yes or no, to indicate their willingness of allowing for a slack.

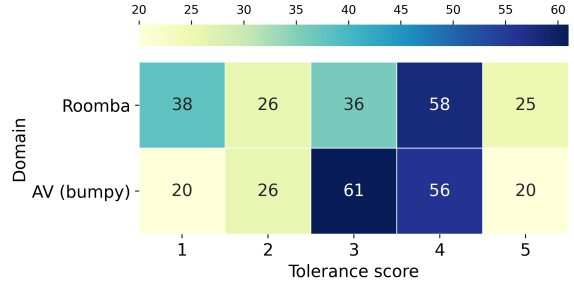
Human Assistance. AI systems often operate in environments that are configurable, which can be leveraged to mitigate NSE. By applying simple modifications to the current environment, significant improvement in performance may be observed. We surveyed the participants to determine their willingness to reconfigure the environment in order to mitigate the impacts of NSE. Reconfigurations for the Roomba domain involved installing a protective sheet on the surface to overcome the negative side effects of spraying water on the walls. For the AV domain, reconfiguration involved installing a pothole-detection sensor that detects potholes and limits the velocity of the vehicle. Participants were asked to select an option that best describes their attitude: purchase and install the sheet or sensor, install the sheet or sensor if it is provided by the manufacturer, and not willing to reconfigure.

Recent research in AI indicates that feedback, particularly from users, can be used to improve the performance of the AI system [7, 14, 16]. We surveyed participants to elicit their preferences over providing feedback and how often they are willing to provide feedback by pressing a button when they notice NSE.

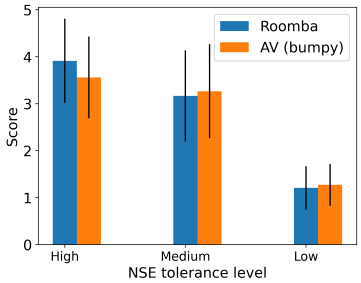
We also surveyed participants to gather information about what type of tools will encourage the users to continue using the system. The participants were asked to select *all* the tools they would be willing to utilize to mitigate NSE. They were presented with three tools: providing feedback by pressing a button every time the system produces NSE; tools to reconfigure the environment; and specifying areas where the system is most prone to NSE and is therefore not allowed to operate, such as a Roomba near the wall.

5 RESULTS

User Tolerance of Negative Side Effects. Responses for the Roomba setting show that 76.50% of the participants were willing to tolerate the negative side effects. For the driving domain, 87.97% of the respondents expressed willingness to tolerate milder NSE such as bumpiness when the AV drives fast through potholes and 57.92% were willing to tolerate relatively severe NSE such as hard braking at a stop sign. These results are shown in Figure 1. Participants were also required to enter a tolerance score on a scale of 1 to 5, with 5 indicating the highest level of tolerance. Figure 2(a) shows the distribution of the user tolerance score. For the Roomba domain, 65.02% voted a score of 3 or more. Similarly for the AV (bumpy) domain, 74.86% voted a score of 3 or more. The mean tolerance score, along with the 95% confidence

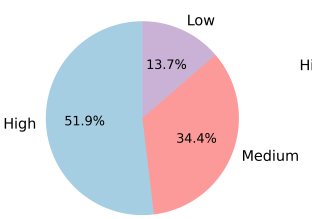


(a) Score distribution

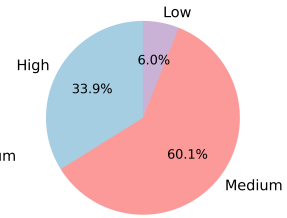


(b) Average score and standard deviation per tolerance level

Fig. 2. Tolerance score.



(a) Roomba



(b) AV

Domain	NSE tolerance	Trust Level (% responses)		
		Low	Medium	High
Roomba	High	0%	21.25%	78.75%
	Medium	3.33%	50.00%	46.67%
	Low	53.49%	37.21%	9.30%
AV	High	0%	47.67%	52.32%
	Medium	5.33%	76.00%	18.67%
	Low	31.82%	54.54%	13.64%

(c) User trust in the system, corresponding to NSE tolerance level.

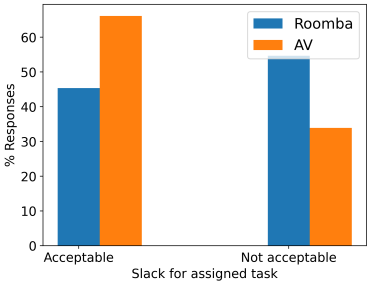
Fig. 3. Effect of negative side effects on trust.

interval, is 3.03 ± 0.20 for the Roomba domain and 3.18 ± 0.16 for the AV domain. Figure 2(b) shows the average score in each NSE tolerance category, along with the standard deviation.

Effect on Trust. For the Roomba domain, 51.91% respondents selected high trust and 34.43% selected medium trust. Similarly for the AV domain, 34.43% selected high trust and 60.10% selected medium trust. The remaining participants indicated that they do not trust the system to be capable of completing its assigned task, when it produces NSE. These results are plotted in Figure 3. Table 3(c) shows the relationship between trust and tolerance of NSE. We also measured the correlation between user tolerance of NSE and their trust in the system's capabilities when it produces NSE. The correlation coefficient in our survey results is 0.65 for the Roomba domain and 0.47 for the AV (bumpy) domain.

Slack Preferences. Among the 183 responses, 66.12% were willing to allow for a slack to avoid NSE of the AV. Similarly, 45.36% were willing to allow the Roomba to skip cleaning areas near the wall so as to avoid the negative side effects. These results are plotted in Figure 4(a). In Table 4(b), we report the relationship between user tolerance of NSE and their slack preferences. We also measured the correlation between user tolerance of NSE and their slack preferences. The correlation coefficient is 0.4 for the Roomba domain and 0.07 for the AV (bumpy) domain.

Willingness to Assist the System. Results on the Roomba domain show that 73.22% respondents were willing to install the sheet to mitigate NSE. If the sheet is not provided by the manufacturer, 64.18% were willing to purchase the sheet ($10). In the AV domain, 91.80% respondents indicated willingness to install the sensor. If the sensor is not provided by the manufacturer, 57.38% were willing to purchase the sensor ($50). These results are reported in Figure 5(a). Table 5(b)

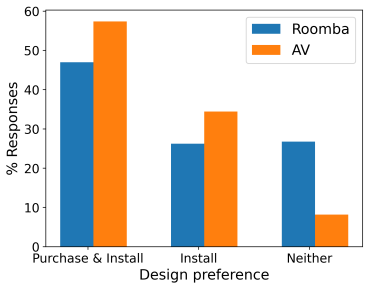


(a) Slack preferences.

Domain	NSE tolerance	Slack preference (% responses)	
		Acceptable	Not Acceptable
Roomba	High	66.25%	38.33%
	Medium	38.33%	61.66%
	Low	16.28%	83.72%
AV (bumpy)	High	66.28%	33.72%
	Medium	70.67%	29.33%
	Low	50.00%	50.00%

(b) Slack preferences of users, corresponding to NSE tolerance.

Fig. 4. Results on user preferences to slack.



(a) Willingness to reconfigure the environment.

Domain	NSE tolerance	User Preference (% responses)		
		Purchase & Install	Install	Neither
Roomba	High	71.25%	20.00%	8.75%
	Medium	40.00%	40.00%	20.00%
	Low	11.63%	18.60%	69.77%
AV	High	66.28%	30.23%	3.49%
	Medium	52.00%	38.67%	9.33%
	Low	40.91%	36.36%	22.73%

(b) User preferences to reconfigure the environment, corresponding to NSE tolerance.

Fig. 5. Willingness to reconfigure the environment to mitigate NSE.

reports user willingness to apply minor modifications to the environment, corresponding to their NSE tolerance. Users with low tolerance of NSE are less willing to perform reconfigurations.

Figure 6 plots user willingness to provide feedback. In the Roomba domain, 43.71% participants were willing to provide feedback until the Roomba learns to overcome its undesirable behavior, 53.00% were willing to provide feedback a few times and when they are around the system to supervise it, and 3.29% were not interested in providing feedback. We observed a similar trend for AV (bumpy) domain—60.11% were willing to provide feedback until the AV learns to overcome the NSE, 36.61% were willing to provide feedback a few times, and 3.29% were not willing to provide feedback. Table 1 reports user interests in utilizing the available tools to mitigate the impacts of NSE. As participants could select more than one tool they prefer to use, we report the number of responses corresponding to each tool.

6 DISCUSSION

User Tolerance. The relation between tolerance level and score (Figure 2) cross-validates the responses to survey questions on user tolerance, as users with a higher tolerance of NSE consistently assigned a higher tolerance score. The results on user tolerance suggest that (1) individual preferences and tolerance of NSE varies and depends on the severity of NSE; and (2) users are generally willing to tolerate NSE that are not severe or safety-critical, but prefer to reduce them as much as possible.

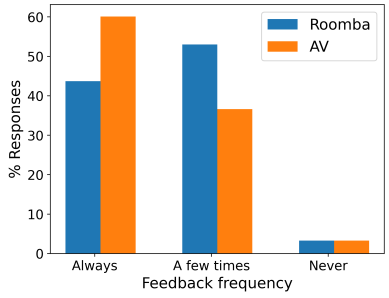


Fig. 6. User willingness to provide feedback to the AI system.

Tool	AV (bumpy)	Roomba
Feedback	30	154
Reconfigure environment	25	5
Specify operation regions	22	18
Feedback + Reconfigure environment	18	4
Feedback + Specify operation regions	45	0
Reconfigure environment + Specify operation regions	9	0
Feedback + Reconfigure environment + Specify operation regions	65	0

Table 1. # Responses corresponding to tools that encourage users to continue using the system, when NSE occur.

User Trust. In both domains, higher NSE tolerance correlates with higher trust of the system, despite NSE occurrence. Users with lower NSE tolerance have low to medium trust in the system. The two key takeaways from these responses are: (1) mitigating NSE is important to improve trust in AI systems; and (2) users are generally willing to give the AI systems some time to learn to avoid the side effects and their trust is affected when the system does not adapt. This highlights the importance of developing techniques to mitigate NSE in order to design trustworthy AI systems.

Slack Preferences. The values in Table 4(b) suggest that participants with high NSE tolerance are generally willing to allow for a slack. The results in Figure 4(a) and Table 4(b) indicate that users are more willing to allow for a slack in the AV (bumpy) domain. We observe that at least 50% of the participants are willing to allow for a slack, independent of their NSE tolerance. This is likely because selecting a longer route or driving slowly to avoid a bumpy ride is common among human drivers. Since many users expressed willingness to allow for a slack in the AV, independent of their tolerance of NSE, the correlation coefficient for the AV domain has a lower value than the Roomba domain. Overall, the results indicate that users are generally willing to accept sub-optimal behavior with respect to the system's assigned task in order to mitigate NSE, as long as the system completes its assigned task.

Human Assistance. The results in Table 1 show that users prefer the direct feedback method the most. This is likely due to the simplicity of the interaction with the system, as they are required to only press a button every time they observe an undesirable behavior. Furthermore, the results in Figure 6 show a higher fraction of users willing to provide feedback to an AV *until* it learns to avoid the NSE. This is likely because the users of an AV are usually in the vehicle when it operates, making it is easier to provide feedback when they observe NSE. The results in Figure 5(a) and Table 5(b) indicate that users are generally willing to engage in environment reconfiguration to mitigate the impacts of NSE. In fact, many users expressed willingness to pay for procuring the items for reconfiguration.

Overall, our results suggest that users are willing to assist the system in mitigating the impacts of NSE. The results suggesting user willingness to provide feedback, often until the system learns to avoid NSE, backs an important assumption in current AI research. Interestingly, users are more willing to utilize all the tools available to mitigate the NSE in the AV domain, compared to the Roomba setting. This interest may be due to the direct implications of the NSE on the user's experience of the ride.

7 SUMMARY AND FUTURE WORK

In this paper, we investigate how people react to negative side effects of AI systems and whether the occurrence of side effects affects user trust in the system's capabilities, via human subjects experiments. We find that users are generally willing to tolerate mild to moderate impacts but prefer to reduce NSE as much as possible. The results also suggest that users are willing to engage with the system by providing feedback, allowing for a slack, or reconfiguring the environment to mitigate NSE. This is in accordance with a recent study that shows that users are generally willing to tolerate an imperfect AI system if they are able to make minor modifications to its performance and outcomes [3]. Furthermore, our results show that the occurrence of NSE could affect user trust, especially if the system does not adapt over time. Our results also show that preferences towards NSE vary across individuals, making the case for the design of customizable systems to improve user satisfaction. Since people prefer different tools to mitigate NSE, depending on the severity and their preferences, it is important to recognize that no one solution approach will work well for all settings.

Our study focuses on a setting where the negative side effects are (1) *known* to the user—we fully describe the side effects to the participants; (2) *deterministic*—the same type of negative side effects always occur when the system executes a certain action, such as the sudden jerk to the passengers when the AV halts suddenly; and (3) *transparent*—the users can observe the occurrence of these side effects. When a new user interacts with a system, the negative side effects may not be known, transparent, and deterministic. That is, the user may not know what types of NSE to anticipate and whether their occurrence is stochastic. The results of this study and the trends we observe may change when users are uncertain about when and why the NSE occurs or what the NSE may be. Throughout this study, we focus on NSE that are undesirable but not safety-critical. The tolerance, trust, slack preferences, and the preferred tools will likely change when NSE are severe or safety-critical. In this work, the impact of NSE on user trust is studied using trust levels (low, medium, or high). Since trust is a latent variable, users may sometimes inadvertently misreport their attitude. In the future, we aim to measure trust using more comprehensive constructs, similar to the models discussed in [10].

We investigate NSE in AI systems that make people's lives easier but are not an essential tool. The results may vary when the role of the system varies, along with the resulting NSE. For example, users may be more willing to tolerate the NSE when the system is an essential tool and the only product on the market, and may not be willing to use the product when NSE are severe such as compromising on the user's privacy. Understanding the relationship between user tolerance of NSE, the severity of the impact, and the purpose and cost of the product is an interesting direction for future work. Furthermore, we considered survey participants aged above 30 since they are less likely to game the Mturk platform. It is likely that the results trend will be slightly different with a younger population who may be more willing to tolerate certain types of negative side effects in the interest of adopting new technologies early. Additional studies are required to investigate the relationship between the user's age and their tolerance of different types of NSE.

Overall, this study encourages the development of effective mechanisms to identify and mitigate negative side effects of deployed AI systems as a way to increase their usability, trustworthiness, and cost effectiveness.

ACKNOWLEDGMENTS

We thank the Amazon mechanical turk participants for their contribution to this study. Support for this work was provided in part by the Semiconductor Research Corporation under grant #2906.001.

REFERENCES

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. *CoRR* abs/1606.06565 (2016).

[2] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[3] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2018. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* 64, 3 (2018), 1155–1170.

[4] Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are Your Participants Gaming the System? Screening Mechanical Turk Workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2399–2402.

[5] Jodi Forlizzi and Carl DiSalvo. 2006. Service Robots in the Domestic Environment: A Study of the Roomba Vacuum in the Home. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-Robot Interaction*.

[6] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J. Russell, and Anca Dragan. 2017. Inverse Reward Design. In *Advances in Neural Information Processing Systems*.

[7] Bill Hibbard. 2012. Avoiding Unintended AI Behaviors. In *International Conference on Artificial General Intelligence*. Springer, 107–116.

[8] Lynn M. Hulse, Hui Xie, and Edwin R. Galea. 2018. Perceptions of Autonomous Vehicles: Relationships with Road Users, Risk, Gender and Age. *Safety Science* 102 (2018), 1–13.

[9] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

[10] Moritz Körber. 2018. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In *Congress of the International Ergonomics Association*. Springer, 13–30.

[11] Victoria Krakovna, Laurent Orseau, Miljan Martic, and Shane Legg. 2019. Penalizing Side Effects using Stepwise Relative Reachability. In *AI Safety Workshop, IJCAI*.

[12] Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and Shane Legg. 2020. Avoiding Side Effects By Considering Future Tasks. In *Proceedings of the 20th Conference on Neural Information Processing Systems*.

[13] Miltos Kyriakidis, Riender Happee, and Joost C.F. de Winter. 2015. Public Opinion on Automated Driving: Results of an International Questionnaire Among 5000 Respondents. *Transportation research part F: traffic psychology and behaviour* 32 (2015), 127–140.

[14] Ramya Ramakrishnan, Ece Kamar, Debadeepta Dey, Julie Shah, and Eric Horvitz. 2018. Discovering Blind Spots in Reinforcement Learning. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*.

[15] Stuart Russell. 2017. Provably Beneficial Artificial Intelligence. *Exponential Life, The Next Step* (2017).

[16] Sandhya Saisubramanian, Ece Kamar, and Shlomo Zilberstein. 2020. A Multi-Objective Approach to Mitigate Negative Side Effects. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*.

[17] Sandhya Saisubramanian and Shlomo Zilberstein. 2021. Mitigating Negative Side Effects via Environment Shaping. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*.

[18] Sandhya Saisubramanian, Shlomo Zilberstein, and Ece Kamar. 2020. Avoiding Negative Side Effects due to Incomplete Knowledge of AI Systems. *CoRR* abs/2008.12146 (2020).

[19] Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, Pieter Abbeel, and Anca Dragan. 2019. Preferences Implicit in the State of the World. In *Proceedings of the 7th International Conference on Learning Representations*.

[20] Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. 2020. Conservative Agency via Attainable Utility Preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

[21] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

[22] Shun Zhang, Edmund H. Durfee, and Satinder P. Singh. 2018. Minimax-Regret Querying on Side Effects for Safe Optimality in Factored Markov Decision Processes. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*.