Parental Trust in Automated Detection of Cyberpredators

L. Brandimarte[*]
[*] Eller College of Management, University of Arizona, Tucson AZ, USA
lbrandimarte@arizona.edu

Abstract - **Online cyberpredators are a serious threat against children who are increasingly using social media and messaging systems to interact with strangers. At the same time, monitoring children's online activity is challenging for parents, due to the numerous platforms and ways a child can access them. Automated approaches could help detect dangerous conversations, but their adoption may prove difficult due to algorithm aversion – the tendency of humans to place less trust in recommendations by automated systems as compared to the judgment of other humans. Three online experiments investigate whether and under what conditions parents are willing to adopt automated systems for detection of cyberpredators, and when they are willing to trade potentially sensitive information about their children's online interaction, as well as individual control and agency. Factors tested for an effect on parental trust include accuracy of predictions by humans versus the algorithm, and storage and management of data in the cloud or local client. Implications for researchers and developers of automated systems of cybercrime detection are presented and discussed.**

Keywords – algorithmic aversion; cyberpredators; experiments.

I. Introduction

Online predators (cyberpredators, or criminals that target minors in the online world) are a serious threat against children who increasingly use social networking and messaging services to interact with strangers. For instance, one study found that one in nine teens received unwanted online solicitations [1]. Parents are advised to monitor their children's use of social media, but this is extremely difficult in practice, given the variety of networking services and access methods that a child can choose from.

Automated approaches which employ, for instance, natural language processing methods could be of tremendous benefit, but the adoption of such approaches might be hindered by a general lack of faith in algorithmic decision making. On the one hand, an automation system would relieve the user of a tedious task, analyzing a child's online communications and detecting cyberpredation – a task of critical importance. On the other hand, adoption of an automation tool to substitute the work traditionally performed manually by parents requires significant trust from the users.

Computer automation is well used and trusted for many applications. A user would never question a computer's ability in a domain such as arithmetic or spell checking, in which functionality is precisely defined and correctness is clear. Instead, trust can be a problem in the use of systems whose correctness criteria are more complicated, such as adaptive agents [2], recommender systems [3], and autonomous driving [4]. In these domains, there are many possible solutions to a problem and some form of *judgement* must be used to choose the best solution. For example, a self-driving car may need to choose how best to avoid a slow pedestrian by either speeding up to clear the intersection before the pedestrian enters, or by slowing down and wait for the pedestrian to cross. Both solutions represent a trade-off in travel time versus safety, and judgement is required to make the decision.

The tendency of humans to place less trust in the judgement of automated systems is often referred to as *algorithm aversion* [5, 6]. In addition to a generalized distrust of automated decision-making, several other factors may impact the ability of a user to trust a tool implementing a cyberpredator detection approach. For instance, algorithmic bias against minority groups, based on identifiable speech patterns, could result in false negatives, which would leave children vulnerable, as well as false positives, which would lead to unfair accusations against innocent individuals. Therefore, perceived accuracy of the algorithm is crucial. Data privacy is also critical, because a cyberpredator detection approach requires access to private, potentially sensitive communications, and a parent must rely on the system to keep those communications in confidence.

In this paper, we explore the effect of algorithm aversion on parents' willingness to adopt an automated detection tool for cyberpredators, the role played by accuracy of the tool, as well as some of the mechanisms which, according to existing work, may lie behind algorithm aversion. We run three experimental studies and gather evidence that algorithm aversion may not be as widespread as existing literature suggests, and that, instead, it may be a function of the study design and the specific context of interest.

A. Related work

Much recent literature has focused on how people perceive algorithmic decision making, and the factors influencing the propensity of people to follow or reject suggestions coming from an algorithm. There seems to be much heterogeneity in people's attitudes towards such suggestions, depending on the specific context under consideration. In journalism, there seems to be a bias against humans. An experiment in South Korea [7] showed that, when an article's author is described as a human, both the general public and journalists rate it of lower quality as compared to the same exact article but described as authored by an algorithm. One wonders if this result would replicate in other countries and contexts, or if, on the other hand, it is specific to a particular culture, historical moment, or specific context.

In another experiment using forecasting tasks (predicting MBA student performance based on admission data; [5]), people showed a bias against algorithms. The mechanism behind this bias was that people observing a human and an algorithm making wrong predictions tend to lose faith in the algorithm faster than losing faith in the human. The same set of authors found a way to counter such algorithm aversion in a follow-up paper [8], in which they show that, when given some level of control (to modify the algorithm's prediction, thus simply using the algorithm as a tool or a starting point), people are more willing to accept the recommendation by the algorithm.

Another set of experiments [9] comparing medical suggestions provided by human or artificial intelligence also show a bias against the algorithm. This bias appears to be caused by the perception that algorithms are not able to take into account the uniqueness of an individual. The bias therefore disappears if the algorithmically generated suggestion is described as personalized or directed to someone other than the self.

In a series of three experiments, we tested whether automated recognition of cyberpredators is trusted by adults. There are at least two levels of trust that need to be taken into consideration. First of all, we consider **Trust in Correctness**, trust intended as belief that the system's prediction is accurate or, at least, more likely to be accurate than the adult's own prediction. Second, we consider **Trust in Privacy**, trust towards the system which needs to be given access to potentially sensitive data in order to perform the assessment. Are people willing to trade sensitive information as well as individual control and agency (for the detection of online dangers) when it comes to identifying cyberpredators? And under what circumstances?

B. *Trustworthiness of Algorithmic Detection*

What does it mean for a system to be trusted? In information security, trust by people is not necessarily a quality that one wants to achieve in a system. Many distributed systems, such as blockchain for instance, make of the lack of required trust by users toward a central authority a major strength. Similarly, one of the main advantages of end-to-end encryption in communication is that users do not need to trust the service provider in order to feel secure and protected when exchanging messages or materials. Therefore, a good rule of thumb in information security seems to be that one should not want to trust a system [10] because trust is not necessary – the system "just works."

However, not all relationships can be fully decentralized, and whenever there is a service provider that acts as a centralized authority in the relationship, there always exists a form of vulnerability [11] that makes trust a necessary condition to the acceptance and to the lasting of that relationship. The question then becomes: What makes average users (not necessarily experts) trust that a centralized system "just works?" What convinces them to believe in it and ultimately use it?

Different disciplines have provided a variety of definitions of trust. In machine learning, a model is trustworthy if it is secure and resilient to attacks by adversaries [12]. In human-computer interaction, researchers developed a definition of online trust as an evolution of its offline counterpart. In the real world, "trust is the social capital that can create cooperation and coordination." In the cyber-world, trust becomes "an attitude of confident expectation in an online situation of risk that one's vulnerabilities will not be exploited" [13]. Economists define trust in a similar fashion as "a willingness to bet that another person will reciprocate a risky move (at a cost to themselves)" [14], and often measure it via the trust game [15]. This is a game in which one player is given a sum of money and a choice to either keep it to him/herself or to pass (part of) it to the other player. The rules of the game are that as soon as the money is passed to the other player, it triples, and the other player can either keep the tripled sum all to him/herself or reward the trust of the first player and split it. A similar definition is used in the management information systems literature, where "Emotional trust is the extent to which one feels secure and psychologically comfortable about depending on the trustee" [16]. Behavioral decision scientists have looked specifically at trust towards algorithms/computers in the medical context, and have defined it simply as "trust in the quality of the recommendation" [17]. Similarly, in marketing, trust "is the perceived competence (i.e. credibility) and benevolence of a target" [18], a definition that some have also used in reference to relationships with robots [19].

Our interpretation of trust encompasses all these definitions and, ultimately, refers to the preconditions for a cyberpredator detection algorithm to be adopted by parents. As briefly noted above, we consider two such preconditions: trust in correctness and trust in privacy. Notice that these are not necessarily two disjunct concepts. Recent work on willingness to download contact tracing apps in order to reduce the spread of SARS-CoV-2 shows that people are willing to trade-off some of their privacy if they perceive the technology to be accurate [20].

Trust in correctness is operationalized in our studies as the described accuracy of the algorithm (experimentally manipulated in Study 2) and as perceived efficacy of the machine used to perform the necessary computations. Trust in privacy is studied by manipulating the location in which the computation occurs (either on one's machine, requiring no data transfer, or on the cloud, thus requiring data transfer via the internet; see Study 3) and through a series of measures that cover various kinds of concerns that existing literature has identified as important in privacy choices. Detailed description of the studies is reported in the following sections. Overall, the three experimental studies investigated the prevalence of algorithm aversion among parents in the context of automated detection of cyberpredators (Study 1); the role of relative accuracy of algorithmic versus human detection of cyberpredators (Study 2); the role of privacy concerns associated with control and relinquishing agency in decisions concerning one's children (all studies), and with sensitive data transfers (Study 3). Results indicate that algorithm aversion may not be as prevalent as earlier work suggested, but perhaps a result of elicitation methods used in experimental procedures or of specific contexts.

II. Study 1

A. *Methods*

Study 1 was a basic test of algorithmic aversion. Participants (all parents recruited via the Amazon Mechanical Turk platform for a 10-minute survey and paid $0.5) were told to imagine they had a chance to have their child's online conversations over the last 4 weeks assessed for free to ensure such conversations did not involve an online predator. Participants were reassured that conversations were anonymized and that the child's identity remained confidential. The study only manipulated (between-subjects) the entity performing the assessment: either an algorithm (algo condition) or a criminologist (human condition) trained to detect predatory conversations. Participants were told that the algorithm/criminologist had historically identified cyberpredators correctly 82-85% of the times, so accuracy was the same across conditions. Participants were then asked to indicate on a scale from 1 (Definitely distrust) to 7 (Definitely trust) how much they trusted or distrusted the algorithm/criminologist to make an accurate assessment. A preference for the criminologist would provide evidence of algorithmic aversion.

Participants were then asked a series of questions which were the same across all three studies. The first one measured interest in having the child's online conversations over the last 4 weeks assessed, and it was followed by an open-ended request to explain the reasons for the choice. Then, in order to measure whether algorithmic aversion was a function of personal involvement (it is possible that in general people trust algorithms, but when it comes to the safety of their own children their emotions and preference for a sense of personal agency dominate their choices), a question asked whose judgement *other parents* should trust more when it comes to detecting predatory conversations – a Human Criminologist or an Algorithm. Answers were on a scale from 1 (Definitely criminologist) to 7 (Definitely algorithm), so a lower value indicated an expectation of algorithmic aversion by other parents.

The next set of questions measured frequency of use of technology by the child, expected level of safety while the child is online, participants' familiarity with privacy/ security concepts and technology in general, and participants' use of Privacy Enhancing Technologies (PETs).

To test for algorithmic aversion in a context different from cyberpredators, participants were then told that artificial intelligence technologies are used in many different fields, including healthcare. For instance, they are being used to improve diagnoses for certain types of cancer. They were then asked to consider a scenario in which a family member – including their child – needed a potential cancer diagnosis, and were asked how likely they would be, on a scale from 1 (Extremely unlikely) to 4 (Extremely likely), to seek out a doctor that uses artificial intelligence for cancer diagnoses.

To better explore the mechanism behind algorithmic aversion, the next series of questions asked, on a scale from 1 (Strongly agree) to 4 (Strongly disagree), whether participants agreed that: AI can perform decision-making just as well as humans; Humans are more reliable for makings decisions rather than AI; A human doctor knows them better than any machine/computer ever could; they expected humans to make mistakes, but machines/ computers should be infallible; they were more willing to accept recommendations made by AI than humans so long as they can modify it, even in the slightest way; and Humans more quickly learn from their mistakes and adapt their beliefs faster than machines.

The study ended with a set of standard demographic questions (age, race, gender, country of residence, occupation, education, salary range).

B. *Results*

We recruited 582 participants. After screening out those who did not complete the study or failed the attention check question (they answered affirmatively to the question of whether they were born before the year 1920), the usable sample consisted of 423 participants (59% male, M_{age} = 34.84, SD_{age} = 9.11). They showed marginal algorithmic aversion, as the average level of trust towards the criminologist (M_{human} = 5.27, SD_{human} = 1.08) was higher than trust towards the algorithm (M_{algo} = 5.13, SD_{algo} = 1.15, 1-tailed t-test $p < 0.1$). Interestingly, algorithmic aversion disappeared when participants imagined other parents were facing a similar decision: the average answer was right in the middle of the scale (M = 4.38, indicating indifference between algorithm and criminologist) and there was no difference across conditions (t-test p-value = 0.49). There was also no evidence of algorithmic aversion in the healthcare scenario, where the median response (and the mode) was 3 (Somewhat likely to seek out a doctor who uses AI for cancer diagnosis) and there was no difference across conditions.

The majority (73%) of participants expressed interest in having their child's conversations assessed, and there was no difference across conditions (z-test $p > 0.1$). Questions on potential mechanisms were not discriminant, as all distributions showed a vast majority of participants agreeing (either somewhat or strongly) with each statement.

III. Study 2

A. *Methods*

Study 2 tested whether algorithmic aversion disappears when the relative accuracy of algorithms versus humans is manipulated. Identical to Study 1, participants (again, all parents recruited via the Amazon

Experiments were funded by the MIS Department at the Eller College of Management, University of Arizona, through their Small Research Grant Program.

Mechanical Turk platform for a 10-minute survey and paid \$0.5) were told to imagine they had a chance to have their child's online conversations over the last 4 weeks assessed for free to ensure such conversations did not involve an online predator. In this Study, however, they were given two options – either an algorithm or a criminologist – and the accuracy of the two options was manipulated between subjects. In the equal condition, accuracy was identical (90%); in the algorithm condition, the algorithm (option 2) was more accurate (95% vs. 86%), and vice versa in the human condition (86% vs. 95%). Participants were then asked to indicate on a scale from 1 (Definitely Option 1, Human) to 7 (Definitely Option 2, Algorithm) which option they would trust the most to analyze and provide an expert opinion of their child's conversations. Low values of this measure, especially in the algorithm condition, would constitute evidence of algorithmic aversion. The remainder of the study was identical to Study 1.

B. *Results*

We recruited 292 participants. After screening out those who failed the attention check question (identical to Study 1), the usable sample consisted of 211 participants (58% male, $M_{age} = 36.15$, $SD_{age} = 9.30$). An analysis of the measure of choice between the algorithm and the human showed that participants valued accuracy and algorithmic aversion was not observed. A one-way ANOVA showed a significant difference across conditions ($F(2, 208) = 21.02$, $p < 0.01$), with all pairwise comparisons resulting significant after Bonferroni adjustment ($M_{equal} = 4.37$, $SD_{equal} = 1.88$, significantly lower than $M_{algo} = 5.53$, $SD_{algo} = 1.40$, and significantly higher than $M_{human} = 3.51$, $SD_{algo} = 2.17$; all p-values < 0.05).

The majority (68%) of participants expressed interest in having their child's conversations assessed, and there was no difference across conditions (43 out of 70 in the equal condition, 50 out of 68 in the algorithm condition, and 51 out of 73 in the human condition; a chi-square test reports a p-value of 0.29).

When participants imagined other parents were to decide whether they should trust the algorithm or a human more, the average answer was right in the middle of the scale ($M = 4.31$, indicating indifference between algorithm and criminologist). Differently from Study 1, and consistent with the lack of algorithm aversion, the accuracy manipulation significantly affected this variable: a one-way ANOVA shows significant differences across conditions ($F(2, 208) = 7.84$, $p < 0.001$), and Bonferroni-adjusted pairwise comparisons show a significant difference between the algorithm ($M_{algo} = 4.91$, $SD_{algo} = 1.63$) and the human condition ($M_{human} = 3.70$, $SD_{algo} = 2.05$, $p < 0.05$). There was also no evidence of algorithmic aversion in the healthcare scenario, where, just like in Study 1, the median response (and the mode) was 3 (Somewhat likely to seek out a doctor who uses AI for cancer diagnosis) and there was no difference across conditions. In particular, one should notice the null effect of the accuracy manipulation in this context: even participants exposed to an algorithm that is less accurate than a human showed no aversion to it assisting the human in the diagnosis.

IV. Study 3

A. *Methods*

Study 3 was an exploratory study which manipulated the location of the automated analysis of children's conversations within subjects. Specifically, participants were asked to imagine they were using an application that assesses their children's online conversations over the last 4 weeks for free, to ensure such conversations did not involve a cyberpredator. Similarly to the previous studies, participants were reassured that the conversations would be anonymized and that the children's identity would remain confidential. They were also informed that storage and analysis could be local on their personal device, with no transmission of data, or transferred to the cloud service. Participants were then asked to choose their preferred option to store and analyze the conversations (Local /Cloud service/ Indifferent).

To explore the reasons behind the choice, they were asked, on a scale from 1 (Strongly disagree) to 4 (Strongly agree), whether they agreed or disagreed with several statements. The first two measured concerns towards cloud-based technologies (Cloud-concern) and asked whether participants: believed data are secure when transferred to the cloud service over the internet (reverse-coded); believed data stored in the cloud are secure (reverse-coded). The third statement asked whether participants were concerned about data privacy when it comes to cloud storage (Privacy-concern). The fourth one asked whether they believed they had more control of information stored and analyzed locally on their personal devices (Control). The last three statements measured perceived efficacy of (or, reverse-coding all responses, concern towards) their own local technology (Local-concern) and asked whether participants believed their computer/device was reliable; it could adequately handle the tasks they need it to perform; it had sufficient storage. The rest of the study was identical to the previous two.

TABLE I. Study 3 Regression Results

	Coefficient	***SE***
Cloud_concern	0.94***	0.24
Privacy_concern	0.04	0.20
Control	0.28	0.24
Local-concern	-1.10***	0.32
N = 224, Pseudo R-squared = 0.11		

*** p < 0.01

B. *Results*

We recruited 292 participants in the same way as the previous studies and, after eliminating those who failed the attention check, the usable sample included 224 parents. The majority (58%) chose to use the algorithm locally, with 34% choosing the cloud service and the remaining being indifferent between the two options.

To investigate which reasons mattered most in the choice, we estimated a logit model using the choice of the cloud service as the binary dependent variable (equal to 1 for Cloud and 0 otherwise) and the four reasons described in the previous section as predictors. The first reason was Cloud-concern and it was built averaging data from first two statements (answers were strongly correlated, with a Chronbach's alpha of 0.87). The next two reasons were Privacy-concern and Control, and the last one was Local-concern, built by averaging data from the last three statements (answers were strongly correlated, with a Chronbach's alpha of 0.74). Perhaps unsurprisingly, results (reported in Table I) indicated that Cloud- and Local-concern were significant predictors of the choice of the cloud service over the local client. However, Privacy-concern and Control were not significantly correlated with the choice.

The majority (76%) of participants expressed interest in having their child's conversations assessed and were neutral regarding the choice by other parents to choose an automated or a human evaluation (M = 4.16, SD = 1.63). There was also no evidence of algorithmic aversion in the healthcare scenario, where, similar to the previous studies, the median response (and the mode) was 3 (Somewhat likely to seek out a doctor who uses AI for cancer diagnosis).

V. Discussion

Three experimental studies investigated the propensity of parents to use an automated system of detection of cyberpredators. Study 1 found a marginally significant level of algorithm aversion: by manipulating the availability of either a human criminologist or an automated system between subjects, parents were more trusting of the human even though the accuracy of the two methods was identical. This form of algorithm aversion may be due to parents not being willing to give up agency over decisions regarding their children. In fact, when asked about what other parents should choose, participants showed no sign of algorithm aversion. Furthermore, algorithm aversion was not detected in the context of a medical doctor using an AI to diagnose a disease, which provides further evidence for the mechanism of agency: when participants think they are leaving the decision to an algorithm, they tend to distrust it, but when the algorithm is only described as supporting the human in their decision, trust is restored. From a societal perspective, this is reassuring, as AI researchers and practitioners have been converging on the idea that the highest contribution AI can provide in the economy is to support rather than replace humans [21].

Study 2 manipulated relative accuracy of the algorithm versus the human between subjects, and asked participants to choose one or the other. When given the option to choose (rather than being assigned to one or the other) and when provided with salient information about accuracy, again, no algorithmic aversion is observed: parents rationally choose the more accurate option.

The empirical evidence of these first two studies suggests that the results found in existing literature, indicating widespread algorithm aversion, may be somewhat inflated by the specific experimental design used to investigate the phenomenon. Specifically, between- versus within-subject manipulations and salience of accuracy information can be important factors affecting the results.

Study 3 was an exploratory study aiming at understanding whether factors that existing literature indicates to be relevant in privacy decision making also affect the propensity to use a potentially privacy-invasive algorithm. Consistent with previous work (e.g., [22]), we find that perception of data security in the cloud is positively correlated with choosing to use an automated cyberpredator detector in the cloud, and efficacy of one's local machine is negatively correlated with it (intuitively: the better one's machine performs, the higher the propensity to choose it over a cloud-based service). However, contrary to our expectations (e.g., [23]), specific data privacy concerns and control concerns were not significantly correlated with the choice of the cloud-based service. The study only allows for speculation as to why that might be the case, but it is possible that those concerns matter at the extensive margin (whether to use the service or not, which was not an option in this study), but less so at the intensive margin (a choice over the location where the computation is performed, locally or in the cloud).

VI. Limitations and Future Directions

While several studies have shown that, with proper checks, the quality of data collected via Amazon Mechanical Turk and the representativeness of the sample as compared to national surveys increases (e.g., [24]), our studies may not replicate in different populations or different contexts.

Furthermore, we only explored a few of the mechanisms that may be responsible for algorithm aversion. For instance, future work may investigate whether attitudes toward algorithms change as a function of the person's mental model and understanding of what an algorithm actually is. Reference [25] provides a good overview of experimental evidence on trust and other attitudinal measures towards algorithms. The claim of the paper is that trust increases if some, but not too much information is provided regarding how the algorithm works. However, the dependent variables used do not actually measure trust towards the algorithm, but rather, they measure trust towards peers grading an essay in the context of an online class. In another set of experiments in the context of medical decision making [26], researchers have shown that algorithm aversion is, at least in part, caused by the fact that algorithms are perceived as inscrutable black boxes, more so than recommendations made by human doctors. This better understanding of human decisions is, however, an illusion, and once patients are exposed to more information about how algorithms reached a certain recommendation, aversion towards them decreases. In the context of predictions and recommendations, research has shown that people's understanding of recommendation systems is quite limited [27]. Furthermore, while algorithms are better than humans in various contexts, such as predicting which jokes people will consider funny, people are still averse to trust the algorithm because they do not understand the automated process [28]. More generally, the evidence on the effect of transparency on trust and other related constructs is mixed, and in need of synthesis and critical review.

VII. Conclusion

Protecting children from dangerous online conversations with cyberpredators is fundamental to their safety but difficult for parents to do effectively, since children may use a wide variety of communication channels and it may be impossible for parents to constantly monitor them. The use of automated tools for detection of cyberpredators represents an effective alternative but it is important to evaluate whether parents would be willing to adopt them. The fact that such tools may be perceived by parents as taking agency away from them when it comes to protecting the safety of their children, and that they may be accessing potentially sensitive information could constitute an impediment to adoption. More generally, parents may not trust an algorithm to decide which situations may be dangerous for their children, and they may show algorithm aversion. In three experimental studies, we find that, while some degree of algorithm aversion may be detected, it is sometimes a consequence of the specific elicitation methods used, and when it is indeed present, it can be addressed by providing salient information regarding the accuracy of the algorithm and its role in support rather than in substitution of human judgment.

Acknowledgment

We thank former students at the University of Arizona's Artificial Intelligence Lab for help in the application for IRB approval and in the setup of the online studies. We also thank the Eller College of Management for funding support.

References

[1] S. Madigan, V. Villani, C. Azzopardi, D. Laut, T. Smith, J. R. Temple, D. Browne, and G. Dimitropoulos. The prevalence of unwanted online sexual exposure and solicitation among youth: A meta-analysis. *Journal of Adolescent Health*, 63(2):133 – 141, 2018.

[2] A. Glass, D. L. McGuinness, and M. Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, 2008.

[3] P. Pu and L. Chen. Trust building with explanation interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interfaces*, 2006.

[4] T. Rezvani, K. Driggs-Campbell, D. Sadigh, S. S. Sastry, S. A. Seshia, and R. Bajcsy. Towards trust- worthy automation: User interfaces that convey internal and external awareness. In *IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 682–688, 2016.

[5] B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

[6] A. Prahl and L. Van Swol. Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6):691–702, 2017.

[7] J. Jung, H. Song, Y. Kim, H. Im, and S. Oh. Intrusion of software robots into journalism: The public's and journalists' perceptions of news written by algorithms and human journalists. *Computers in Human Behavior*, 71:291–298, 2017.

[8] B. J. Dietvorst, J. P. Simmons, and C. Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170, 2018.

[9] C. Longoni, A. Bonezzi, and C. K. Morewedge. Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4):629–650, 2019.

[10] Trusted third parties are security holes. Available at https://nakamotoinstitute.org/trusted-third-parties/. Last accessed on February 18, 2023.

[11] Y. D. Wang and H. H. Emurian. An overview of online trust: Concepts, elements, and implications. *Computers in Human Behavior*, 21(1):105–125, 2005

[12] N. Papernot. What does it mean for machine learning to be trustworthy? San Francisco, CA, Jan. 2020. USENIX Association.

[13] C. L. Corritore, B. Kracher, and S. Wiedenbeck. On-line trust: concepts, evolving themes, a model. *International Journal of Human-Computer Studies*, 58(6):737–758, 2003.

[14] C. F. Camerer. *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press, 2011.

[15] J. Berg, J. Dickhaut, and K. McCabe. Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142, 1995.

[16] S. Y. Komiak and I. Benbasat. The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly*, pp. 941–960, 2006.

[17] M. Promberger and J. Baron. Do patients trust computers? *Journal of Behavioral Decision Making*, 19(5):455–468, 2006.

[18] P. M. Doney and J. P. Cannon. An examination of the nature of trust in buyer–seller relationships. *Journal of Marketing*, 61(2):35–51, 1997

[19] J. Wirtz, P. G. Patterson, W. H. Kunz, T. Gruber, V. N. Lu, S. Paluch, and A. Martins. Brave new world: Service robots in the frontline. *Journal of Service Management*, *29*(5): 907-931, 2018.

[20] G. Kaptchuk, E. Hargittai, and E. M. Redmiles. How good is good enough for covid19 apps? The influence of benefits, accuracy, and privacy on willingness to adopt. *arXiv preprint arXiv:2005.04343*, 2020.

[21] A. McAfee and E. Brynjolfsson. *Machine, platform, crowd: Harnessing our digital future*. WW Norton & Company, 2017.

[22] H. J. Smith, T. Dinev T, and H. Xu. Information privacy research: An interdisciplinary review. *MIS Quarterly*, 1:989-1015, 2011.

[23] L. Brandimarte, A. Acquisti, and G. Loewenstein. Misplaced confidences: Privacy and the control paradox. *Social psychological and personality science*, 4(3):340-7, 2013.

[24] N. Qureshi, M. Edelen, L. Hilton, A. Rodriguez, R. D. Hays, and P. M. Herman. Comparing data collected on Amazon's Mechanical Turk to national surveys. *American Journal of Health Behavior*, 46(5):497-502, 2022.

[25] R. F. Kizilcec. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2390–2395, 2016.

[26] R. Cadario, C. Longoni, and C. K. Morewedge. Understanding, explaining, and utilizing medical artificial intelligence. *Nature human behaviour*, 5(12):1636-42, 2021.

[27] O. Alvarado, H. Heuer, V. V. Abeele, A. Breiter, and K. Verbert. Middle-aged video consumers' beliefs about algorithmic recommendations on youtube. *arXiv preprint arXiv:2008.03202*, 2020.

[28] M. Yeomans, A. Shah, S. Mullainathan, and J. Kleinberg. Making sense of recommendations. *Journal of Behavioral Decision Making*, 32(4):403–414, 2019.