Ethical AI Development Requires Mitigation of Overtrust, Expectable Misuse, and Misuse-Induced Harm

Ben Manley
EECS Department
University of Michigan
Ann Arbor, MI
bdmanley@umich.edu

***Abstract*—There is an empirically-verified tendency for humans to overtrust artificially intelligent systems like autonomous vehicles and robots. This overtrust leads to misuse of the systems, and this misuse can have drastic consequences including loss of life. Techniques have been proposed and technologies have been developed to discourage overtrust and directly prevent misuse. These techniques and technologies are not necessarily employed, and even when any one of them is, it still has great possibility of allowing negative outcomes to exit the "pipeline" that connects overtrust to misuse to harm. This paper will argue the following: In order to consider AI development to be done ethically, we must require that developers work to reduce the probability, scope, and magnitude of harm stemming from the natural human tendency to overtrust AI. This must happen at three levels: reducing user overtrust, preventing misuse caused by overtrust, and mitigating harms caused by misuse. The paper then outlines and responds to potential critiques of the argument.**

I. Introduction

Artificial Intelligence (AI) is growing rapidly in ability, complexity, and the scope of problems for which it is employed. As this growth continues, it is becoming increasingly imperative that we design and deploy agents ethically, such that the agents interact in positive ways with the humans whose societies they are entering. Because of the widening applicability and increasing complexity of AI agents, there are many hurdles to overcome on the road to ethical AI. The hurdle this paper will focus on is the mismatch between the amount of trust that users have for AI systems and the amount of trust those systems actually deserve.

A. Overtrust

There is growing evidence of a strong tendency for humans to overtrust AI systems.

A study in which participants had to rely on a pet feeding robot to keep their (simulated) cat alive led researchers to conclude that "the development of inappropriate trust in intelligent systems has to be seen not as the exception but as the rule" [23].

A study in which participants had to exit from a simulated emergency with the help of either a guidance robot or established exit signs led researchers to conclude that "people tend to believe robots are competent at first sight" [18]. In that same study, it was even found that participants trusted the robot agent after having previously seen it fail in its primary function. This is even more significant when we consider that, for certain types of agents (e.g., autonomous vehicles), a single mishap by the agent can be fatal after long periods of trust-inducing perfect performance.

Finally, a global study revealed that "71% of motorists believe that they can buy a self-driving car today, while 11% would be tempted to have a brief nap while using current 'Highway Assist' systems" [1]. This is indicative of widespread misunderstanding of current AV capabilities in the eyes of the public.

B. Causes of Overtrust

Overtrust in AI has a number of causes. One of them is the natural human response to automation. As one example, a well-cited study on "automation-induced complacency" showed that "operator detection of automation failures was substantially worse for constant-reliability than for variable-reliability automation" [15] - people grow complacent after long-enough periods of fault-free automation.

Another cause is the promotion of overtrust by AI developers. This promotion can be intentional, as in the case of Tesla's marketing: they market their automated driving systems as "Full Self-Driving" but register the vehicles with the DMV as being much less capable [12].

The promotion of overtrust can also be unintentional, as in the case of academic publishing. A 2020 study analyzed previous studies "comparing the performance of diagnostic deep learning algorithms for medical imaging with that of expert clinicians" [14]. A majority of these studies claim that their respective algorithms demonstrated performance "at least comparable to (or better than) clinicians". The researchers in the 2020 study concluded that a majority of the studies "are at high risk of bias", "deviate from existing reporting standards", and have poor reproducibility. This sheds light on the apparent tendency for researchers to applaud the triumphs of autonomous agents without proper research methods as support, unintentionally promoting overtrust.

C. Misuse and Consequences

Unfortunately, overtrusting an AI system can have dire consequences due to the resulting misuse of the system. One study directly presented the effects of overtrust by observing participants' handling of deceleration failures in a driving simulator with varying levels of driving automation [20]. Participants demonstrated poorer driving when the car was "highly automated" than when it was "semi-automated", showing that the real danger of overtrust is that it directly leads to less appropriate use of the system at hand.

The topic of autonomous vehicles (AVs) is useful when discussing consequences because the technology is covered widely in the media, and the consequences of its failure are intuitive. The first pedestrian death by an AV occurred in 2018 [13] and has become a widely-known reminder of the responsibility we are giving vehicles as we automate them. That case involved an Uber driver, who was trained to monitor the autonomous vehicle, allegedly watching her phone instead of carefully observing for dangerous vehicle actions. This shows that "experts" who are trained for the explicit purpose of not overtrusting intelligent systems are still susceptible to humans' overtrusting tendencies. This also shows that human fatalaties are a real possibility when we approach this field improperly.

Misuse is possible in a variety of scenarios outside of AVs, and the medical field provides some frightening examples. One of them is computer-aided cancer detection [16]. While the benefits can appear obvious (even in the cited case study, it was shown that many doctors' screening decision accuracies were increased for easily detectable cancers), the application of the aid *decreased* screening accuracy for "relatively difficult cancers" by the doctors who were most accurate without a computer aid. In real-life medical settings, certain patients would therefore receive false negatives because of their doctors' overtrust of the detection aiding system.

D. Previously-Proposed Paths Forward

The broad scope of this topic and the consequences of ignoring or mistreating it has led various experts in the field to propose methods for improving the current landscape of AI development.

One subject, which will be heavily referred to in this paper, is that of improving the design of autonomous agents and the systems around them. The problem of overtrust for AI systems can be at least partially mitigated by design changes like those that "avoid features that may nudge users toward anthropomorphizing robots" [24]. If this is done, users will be less likely to trust systems as if they are other humans and instead treat them as - in my words - the "cold" machines they are. The problem of misuse as a result of overtrust can be fixed in certain situations by including methods for "denial of service" [11]. This involves monitoring (and potentially modeling [24]) a user's state and pattern of trust to determine if overreliance or misuse is occurring. If it is, the AI system then disables its own autonomy.

Another subject (which could also fall under the previous topic of improving agents' design) is the introduction of "explainable AI" [11]. This involves ensuring that users are able to understand the decision-making process of an agent. By default, this is difficult, because many modern machine learning methods are effectively blackboxes. Developing explainable AI - without misleading people by "fairwashing", or "promoting the false perception that a machine learning model respects some ethical values" [4] - could provide better insight to users about what their AI system is actually capable of.

Finally, regulation can be improved. Within the topic of AVs, the task of determining a vehicle's level of autonomy could be delegated to a regulatory body rather than the vehicle's developer to prevent misrepresentation [12]. For learning algorithms in general, an additional "federal consumer protection agency" akin to the United States Food and Drug Administration could be created to regulate new algorithms [22]. This would help to counter the current culture of widely deploying new algorithms without properly analyzing their potential impacts. Additionally, liability law could be updated to put more pressure on developers for mistakes made by their systems. Legal reviews have investigated potential applications or extensions of "product liability and negligence tort law" [5], as well as animal law [3], [5], [19], to the case of robots or AI systems in general. Lastly, regulation could shift toward holding developers accountable by removing their ability to treat human users as "liability sponges" and "moral crumple zones" [10] when AI systems lead to harmful outcomes.

II. The Overtrust Pipeline

It is sensible and useful to view the discussed problems as an *overtrust pipeline* that starts at overtrust, passes through misuse, and ends with harmful outcomes for users and others. The sensibility of this perspective stems from the discussion in the introductory section above. The usefulness of this perspective comes from its breaking down of the phenomena that lead from human nature to harm.

In this paper, I argue that AI developers should target three broad goals - corresponding to the three stages of the overtrust pipeline - to reduce the probability, scope, and magnitude of harm caused by users' overtrust-induced misuse of their systems. These general goals are reducing user overtrust, preventing misuse caused by overtrust, and mitigating harms caused by misuse. Without engaging in all three types of preventative measure, developers are leaving potential preventable harm on the table. If they neglect to attend to one of them, they should at least attend to the others.

Many of the individual claims about specific mitigation tactics in this paper are those of other researchers and are duly cited. The purpose of this paper is (1) to organize and support existing claims and (2) to introduce and support the perspective that the harm mitigation work is most complete when developers aim for all (or as many as possible) of the three targets discussed. Before discussing the three targets, it's important to quickly note the general reason that any of them are goals for ethical development: AI developers are

in a unique position because they can shape the interactions of their agents with their users. If there is predictable harm (intentional or not) from a certain interaction design, and a developer chooses not to prevent it, many would view that as similar or equivalent to purposefully causing that harm [25]. If a developer addresses any of the three targets, they are taking a step in the correct direction by reducing harm; if a developer addresses all three of the targets, they are making the comprehensive ethical decision by minimizing it.

A. Reducing Overtrust

As previously mentioned, humans have a natural tendency to overtrust AI systems. As common sense should lead us to understand, and as evidence discussed earlier in this paper has shown, believing a system will perform with greater ability than it actually has is dangerous, especially when the systems are put in charge of our physical or mental well-being. Developers of such systems should therefore ensure that users trust their systems no more than the systems deserve. The overtrust pipeline can't lead to harm if the initial overtrust is avoided in the first place.

A number of authors have already argued for this, along with numerous proposed tactics for designing to reduce overtrust [11], [24]. Earlier in this paper, these tactics were said to include reducing the likelihood of anthropomorphization of systems by users [24] and introducing explainability into AI systems [11]. These are just a few examples, and developers should take advice from the surrounding literature for additional methods of discouraging user overtrust.

Unfortunately, the reason that reducing overtrust is such an important endeavor to argue for right now is because the exact opposite is currently happening. As mentioned earlier, Tesla explicitly *promotes* overtrust with their claims of "Full Self Driving" capabilities while avoiding tighter AV testing regulations by effectively lying to regulators [12]. The fact that such a significant player in the AI field is exhibiting such irresponsible behavior necessitates the argument for the opposite. Worse, it isn't just Tesla misbehaving. "Autonowashing", or "making unverified or misleading claims which misrepresent the appropriate level of human supervision required by a partially or semi-autonomous product" [8], is rampant in the industry. This should indicate that the discouragement of overtrust must be pushed for heavily because, in order to tackle it, a major shift will be necessary for developers to prioritize honesty and transparency.

B. Preventing Misuse, Assuming Overtrust

Regardless of the anti-overtrust measures researchers develop, it seems safe to assume that there will always be users who place too much trust in AI systems - whether due to inefficacy of the measures, unexpectedly high propensity for trust in certain users, or developers' neglect to implement the measures. The next line of defense that should be expected of ethical developers is preventing misuse, with overtrust as a prior assumption.

A successful example of this can be found in lane-keeping assistance technology in cars. The most common implementation requires the driver to keep a hand on the wheel at all times, and will issue audible and visual warnings when that requirement is violated [17]. The developers assume that the user will overtrust the system and attempt to let the lane-keeping independently control the car (anecdotally, numerous people around me have tried this), so they added the warnings to prevent the expectable misuse. This is done despite the fact that the terms of use of the cars and their lane-keeping technology explicitly state the responsibility of the driver to remain in control at all times (see [7] as an example).

Besides the necessity of preventing misuse (with this necessity stemming from the inevitability of overtrust regardless of its countermeasures), one of the most appealing aspects of this class of harm-preventing measures is its technological feasibility. Preventing misuse inherently involves *sensing* misuse, and this sensing has already been developed in many cases. A simple case is the lane-keeping assistant; given the abilities of the system described above, current implementations are able to sense a lack of input to the steering wheel from the user. In more difficult cases that require ensuring a user's attention is as expected, eye tracking systems exist that can accurately pinpoint where a person is looking in real time [2]. In more general cases of sensing a user's intentions and expected behaviors, large social media platforms like Facebook and Google (where the platforms are themselves effectively intelligent systems) enable the modeling of a user's mental state [6]. Once misuse (or behavior indicating expected future misuse) is sensed, safety protocols can prevent it. The fact that the mitigation of misuse is possible from a technological perspective makes it much easier to argue that ethical AI development should necessarily include it.

The question of whether misuse prevention should be seen as an ethical duty of developers can be generalized from AI systems to all commercial products. Products of all kinds are misused constantly, as demonstrated by the product liability lawsuits and product misuse arguments established in our legal system [21]. For example, consider drunk driving. Operating a motor vehicle while inebriated is one of the most widely identifiable and universally disdained cases of product misuse because of its prevalence and drastic consequences. For a majority of the period that cars have existed, the responsibility to reduce drunk driving has fallen entirely on the individual driver. However, some people now hold the view that car manufacturers should share that responsibility, and legislation has been proposed that would mandate new cars to monitor for and prevent drunk driving [9]. If this legal shift is in any way indicative of an analogous shift in the public's ethical stance, then ethical AI development (being a subset of ethical development of products in general) would naturally necessitate the prevention of misuse.

C. Mitigating Harms, Assuming Misuse

The overtrust pipeline ends with harmful outcomes for users and more. Ethical developers must prepare for users to circum-

vent any anti-overtrust and anti-misuse measures earlier in the pipeline; no system can be assumed to perfectly prevent users from overtrusting it, and no system can be assumed to perfectly prevent overtrusting users from misusing it. Therefore, the final defense that developers have against this problem space is reducing the harms that come from unexpected or undetected misuse. Speaking generally, this amounts to making designs that are *forgiving* - designs that, as much as possible, minimize the probability, scope, and magnitude of harmful outcomes when use of the system deviates from that which is expected. See Figure 1 for clarity.

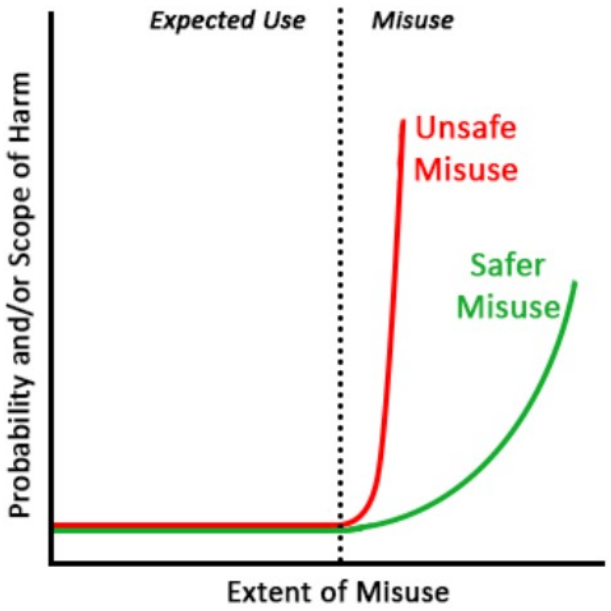

Fig. 1. AI developers should aim to minimize the impact misuse has on harmful outcomes.

Continuing with the lane-keeping assistance technology example: there will be cases where a user overtrusts and misuses the system (takes their hands off the wheel, falls asleep, etc.), the system attempts to stop the misuse (sends audible and visible warnings to the driver), and the user continues the misuse (ignores the warnings or does not wake up). There are multiple feasible designs for the system's response at this point and the ethical one is that which minimizes the resulting harm. Instead of the car doing something like shutting off the lane-keeping (which would be a valid design choice to deter misuse and is what some current systems do [7]), it should instead gradually slow the vehicle to a halt while keeping the current lane (or, if technologically possible in that vehicle, pull over to the side of the road before stopping).

In other cases, designing to reduce harm from misuse could be much simpler than implementing additional safety features. As an example, we can consider how we would design the cancer detection software that was previously mentioned to have possible negative effects [16]. One design we could imagine would involve the software internally determining a confidence or probability that a given mammogram contains cancer and returning a binary signal to the doctor based on a confidence threshold. If the confidence exceeds the threshold, the doctor is warned that cancer was detected; otherwise, the doctor is told no cancer was detected. Another design we could imagine would involve the software directly reporting the calculated confidence level to the doctor. This second design can be justified by assuming that doctors will misuse the detection system by overrelying on its output. This design would potentially reduce harm (the doctor reporting false negatives) in at least some cases because the doctor could be more inclined to perform their standard check if the reported confidence is not extremely low.

III. Potential Critiques and Rebuttals

In this section, I address a few of the potential critiques that could be made against this paper's argument.

Developers should be able to "play the game" and make systems that maximize users' trust in them.

The game is capitalism, and players tend to be given significant leeway in the name of laissez-faire economics. However, here we're talking about human users' trust in an artificially intelligent agent capable of taking its own independent actions in the physical or virtual world. Delivering an AI product may soon be seen as closer to sending a human representative than to shipping a "normal" or "classical", unintelligent tool, and sending an independent representative designed to gain undue trust is effectively lying. Also, these agents are every day being placed in charge of wider-reaching networks and more critical infrastructure, and potential consequences of their failure have already reached catastrophic levels. We may say all is fair in war, and developer competition may present as a war. However, users and those around them are civilian bystanders, and taking advantage of users' human tendency to overtrust and thereby generating the potential for their harm is unethical for the same reason civilian casualties are.

Additionally, the argument in this paper only attempts to limit developer behavior *reasonably*. The "playing of the game" (market pressures) will naturally lead developers to optimize users' sentiment toward their product (and until a harmful outcome is observed by a user, high sentiment and high trust seem intertwined). Arguments for eliminating overtrust just say that systems shouldn't be designed (intentionally or not) in such a way that they are perceived as more trustworthy than they actually are. These lower and upper bounds, respectively, on appropriate trust generation should lead any ethical developer to create a system that is trusted exactly as much as it should be. Darwinian economics should then act to select the systems with the greatest *actual trustworthiness* rather than the systems with the greatest (possibly skewed, possibly temporary) *perceived trustworthiness*.

Developers waive all responsibility once users sign or agree to terms of use. It is up to users to acknowledge the possibility of overtrust and misuse and to prevent it themselves.

When the word "responsibility" comes into play, the question becomes a more difficult one and we lose sight of the

fact that it is in everyone's best interest to minimize the harm to users. Rather than focusing on responsibility or blame, we should focus on every actor doing the more right thing, the more ethical thing. Regardless of where responbility or fault lies, the more ethical decision for developers is to reduce overtrust, misuse, and harm. This is especially true considering developers' unique ability to do so from the beginning of the design process, long before any user interacts with their system.

To engage with the critique more, however, we can temporarily focus on "responsibility" as an important attribute to assign. When overtrust of an AI system is expected with high probability, developers should be ethically responsible (and ideally legally responsible as well) for implementing mitigation methods like those argued for in this paper. In fact, developers of non-AI systems and products are already subject to legal responsibility in this area through the concept of "foreseeability". According to this concept, "a manufacturer is expected to anticipate certain uses of his product by the consumer" and "may be legally liable for injuries sustained in certain unintended uses" [21]. This should directly translate to the possibility of AI developers being liable for foreseeable, overtrust-induced misuse. Depending on the type of misuse in our AI system scenarios, many would likely find the user to be negligent, but the preponderence of evidence of human tendencies to overtrust make the types of misuse discussed in this paper quite foreseeable. This should be enough to at least warrant joint liability between the user and the developer.

Developers can't be expected to think of every possible misuse of their systems.

We should all agree, and expecting them to do so would undoubtedly be detrimental to technological progress and innovation. We should only expect the mitigation of *expectable* misuse (and its potential harmful outcomes, and the overtrust that led to it). The next question is naturally, "How do we define *expectable misuse* - as opposed to acceptably unexpectable misuse - of an AI system?". There isn't necessarily a solid answer to that question, as its analogy in classic product liability has to be debated in court, resolved on a case-by-case basis.

The most obvious cause for classifying a specific misuse of a system as being "expectable" would be findings of a market research or behavioral trial explicitly studying humans using that system. Were a trial like that to expose a tendency for users to misuse the system in a particular way, regardless of whether they agreed to avoid that particular behavior, the developer would thereafter be aware of the potential harms of the system design. The ethical decision would thereafter be to change the design or to implement preventative measures.

A potential way to improve the ability to classify a certain misuse as being reasonably expectable is to consider what I will call *situational reducibility*. Given the expanding literature on overtrust in AI and automated systems, more and more "situations" are explicitly being proven to generate overtrust in human users. We can consider that, for each AI usage situation, there is a class of situations that are *reducible* to it in terms of the nature of the user-system interaction or potential psychological states of the user. As a small example, the demonstration that people overly trust guidance robots during building emergencies [18] could be used to claim that developers should expect users to overly trust non-robot guidance agents in emergencies as well. Alternatively, some would find reasonable only the weaker claim that developers' due diligence would now include expanding the original study to the non-robot case. All cases would therefore warrant a review of the relevant literature, with the possibility of the developer conducting their own study if reducibility from existing studies is untenable. Regardless, this broad concept of situational reducibility could be a starting point for establishing a standard for when or how to expect misuse, thereby reducing harmful outcomes if the standard is followed.

IV. Conclusions

The three types of countermeasures - those discouraging initial overtrust, those intervening when user interactions deviate from expected, and those minimizing the negative results in the end - are clearly each beneficial in preventing harm to users. When deciding what ethical AI development should look like, we must demand the best from developers. We must expect that the systems they create, which will continue to have more important responsibilities placed upon them, are designed with the minimization of harm in mind at multiple levels. Redundancy is the backbone of critical systems, so with AI agents taking ever more critical roles, redundancy in the prevention of harmful outputs is paramount.

As these systems are already actively being deployed, we have to make sure not to lose sight of our primary goal while advancing AI technology: minimizing harm. This can be difficult, as it's natural to be distracted by details like exactly who should be "responsible" for certain harm-preventing measures or exactly what types of overtrust-induced misuse ethical developers should expect. The one expectation that will unequivocally lead us further along the path to ethical AI is that developers must prioritize the type of harm mitigation discussed in this paper just as highly as they prioritize the technical ability of their agents.

V. Acknowledgments

This paper was written for University of Michigan's EECS 598: Ethics in AI and Robotics. I would like to thank Professor Benjamin Kuipers and Heather Berg for organizing the course and assisting me in refining my selected topic. I would also like to thank Alia Gilbert, Zubin Aysola, Jacob Overberg, and Alina Mizrahi for their help in tuning the paper's argument.

References

[1] Automated driving hype is dangerously confusing drivers, study reveals. Technical report, Thatcham Research, October 2018.

[2] Christer Ahlstrom, Katja Kircher, and Albert Kircher. A gaze-based driver distraction warning system and its effect on visual behavior. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):965–973, 2013.

[3] Peter Asaro. The liability problem for autonomous artificial agents. In *AAAI Spring Symposia*, 2016.

[4] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: the risk of rationalization, 2019.

[5] Woodrow Barfield. Liability for autonomous and artificially intelligent robots. *Paladyn, Journal of Behavioral Robotics*, 9(1):193–203, 2018.

[6] Stevie Chancellor and Munmun De Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3(1), March 2020.

[7] Honda Motor Company. 2020 honda civic lane keeping assist system (lkas). Accessible at https://owners.honda.com/utility/download?path=/static/pdfs/2020/Civic

[8] Liza Dixon. Autonowashing: The greenwashing of vehicle automation. *Transportation Research Interdisciplinary Perspectives*, 5:100113, 2020.

[9] Ian Duncan. New technology mandate in infrastructure bill could significantly cut drunken driving deaths. *The Washington Post*. Available at https://www.washingtonpost.com/transportation/2021/11/09/drunk-driving-technology-infrastructure/.

[10] M C Elish. Moral crumple zones: Cautionary tales in human-robot interaction (WeRobot 2016). *SSRN Electronic Journal*, 2016.

[11] Karen Hao. We need to design distrust into ai systems to make them safer. *MIT Technology Review*, May 2021. Interview with Ayanna Howard.

[12] Hyunjoo Jin and Jaiveer Shekhawat. California reviews whether tesla's self-driving tests require regulatory oversight. *Reuters*, January 2022.

[13] Matt McFarland. Uber self-driving car operator charged in pedestrian death. *CNN*, September 2020. Available at https://www.cnn.com/2020/09/18/cars/uber-vasquez-charged/index.html.

[14] Myura Nagendran, Yang Chen, Christopher A Lovejoy, Anthony C Gordon, Matthieu Komorowski, Hugh Harvey, Eric J Topol, John P A Ioannidis, Gary S Collins, and Mahiben Maruthappu. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*, page m689, March 2020.

[15] Raja Parasuraman, Robert Molloy, and Indramani Singh. Performance consequences of automation induced complacency. *International Journal of Aviation Psychology*, 3, 02 1993.

[16] Andrey A Povyakalo, Eugenio Alberdi, Lorenzo Strigini, and Peter Ayton. How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography. *Med. Decis. Making*, 33(1):98–107, January 2013.

[17] Consumer Reports. Guide to lane departure warning & lane keeping assist, 2019. Available at https://www.consumerreports.org/car-safety/lane-departure-warning-lane-keeping-assist-guide-a7087080070/.

[18] Paul Robinette, Ayanna Howard, and Alan R Wagner. Conceptualizing overtrust in robots: Why do people trust a robot that previously failed? In *Autonomy and Artificial sIntelligence: A Threat or Savior?*, pages 129–155. Springer International Publishing, Cham, 2017.

[19] Enrique Schaerer, Richard Kelley, and Monica Nicolescu. Robots as animals: A framework for liability and responsibility in human-robot interactions. pages 72 – 77, 11 2009.

[20] Niklas Strand, Josef Nilsson, I.C. MariAnne Karlsson, and Lena Nilsson. Semi-automated versus highly automated driving in critical situations caused by automation failures. *Transportation Research Part F: Traffic Psychology and Behaviour*, 27:218–228, 2014. Vehicle Automation and Driver Behaviour.

[21] William L. Trombetta and Timothy L. Wilson. Foreseeability of misuse and abnormal use of products by the consumer. *Journal of Marketing*, 39(3):48, July 1975.

[22] Andrew Tutt. An fda for algorithms. *Administrative Law Review*, 69(1):83–123, 2017.

[23] Daniel Ullrich, Andreas Butz, and Sarah Diefenbach. The development of overtrust: An empirical simulation and psychological analysis in the context of human–robot interaction. *Frontiers in Robotics and AI*, 8, 2021.

[24] Alan R Wagner, Jason Borenstein, and Ayanna Howard. Overtrust in the robotic age. *Commun. ACM*, 61(9):22–24, August 2018.

[25] Fiona Woollard and Frances Howard-Snyder. Doing vs. Allowing Harm. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.