



# Current State of Knowledge on Failures of AI Enabled Products

**Dr. Roman V. Yampolskiy**  
Computer Engineering and Computer Science  
Speed School of Engineering  
University of Louisville  
[roman.yampolskiy@louisville.edu](mailto:roman.yampolskiy@louisville.edu)

*Reviewed and Released by*  
Dr. Mahmood Tabaddor  
Consortium for Safer AI  
[Mahmood.tabaddor@ul.com](mailto:Mahmood.tabaddor@ul.com)

*Original Release: January 25, 2018*  
*Revised Version: April 2, 2018*

*This project was funded by the generosity of UL LLC, one of the founding sponsors of the Consortium*

## SUMMARY

In this work, we present and analyze publicly reported failures of artificially intelligent (AI) products and extrapolate our analysis to future accidents. We suggest that both the frequency and the seriousness of future AI failures will steadily increase as more and more AI products enter the marketplace. We conclude with some recommendations from experts on reduction of most common causes of failure of AI products.

**Keywords:** *AI Accidents, AI Safety, Cybersecurity, Failures, Superintelligence.*

## 1. Introduction

About 10,000 scientists<sup>1</sup> around the world work on different aspects of creating intelligent machines, with the main goal of making such machines as capable as possible. With amazing progress made in the field of AI over the last decade, it is more important than ever to make sure that the technology we are developing has a beneficial impact on humanity. With the appearance of robotic financial advisors, self-driving cars and personal digital assistants, come new and unresolved problems. We have already experienced market crashes caused by intelligent trading software<sup>2</sup>, accidents caused by self-driving cars<sup>3</sup> and embarrassment from chat-bots<sup>4</sup> which turned racist and engaged in hate speech. We predict that both the absolute frequency and seriousness of such events will steadily increase as AIs become more capable.

## 2. AI Failures

As the saying goes, those who cannot learn from history are doomed to repeat it. Unfortunately, very few papers have been published on failures and errors made in development of intelligent systems [1]. Importance of learning from “What Went Wrong and Why” has been recognized by the AI community [2, 3]. Such research includes study of how, why and when failures happen [2, 3] and how to improve future AI systems based on such information [4, 5].

We are not strangers to failures and as technologies have evolved so have failure modes. Signatures have been faked, locks have been picked, Supermax prisons have had escapes, guarded leaders have been assassinated, bank vaults have been cleaned out, laws have been bypassed, fraud has been committed against our voting process, police officers have been bribed, judges have been blackmailed, forgeries have been falsely authenticated, money has been counterfeited, passwords have been brute-forced, networks have been penetrated, computers have been hacked, biometric systems have been spoofed, credit cards have been cloned, cryptocurrencies have been double spent, airplanes have been hijacked, CAPTCHAs have been cracked, cryptographic protocols have been broken, and even academic peer-review process has been bypassed with tragic consequences. Millennia long history of humanity contains millions of

---

<sup>1</sup> <https://intelligence.org/2014/01/28/how-big-is-ai/>

<sup>2</sup> [https://en.wikipedia.org/wiki/2010\\_Flash\\_Crash](https://en.wikipedia.org/wiki/2010_Flash_Crash)

<sup>3</sup> <https://electrek.co/2016/05/26/tesla-model-s-crash-autopilot-video/>

<sup>4</sup> [https://en.wikipedia.org/wiki/Tay\\_\(bot\)](https://en.wikipedia.org/wiki/Tay_(bot))

examples of attempts to develop technological and logistical solutions to increase safety and security, yet not a single example exists, which has not experienced a failure at some level.

Accidents, including deadly ones, caused by software or industrial robots can be traced to the early days of such technology<sup>5</sup>, but they are not a direct consequence of particulars of the intelligence embedded in such systems. AI product failures, on the other hand, are directly related to the mistakes produced by the intelligence such systems are designed to exhibit. We can broadly classify such failures into mistakes during the learning phase and mistakes during performance phase. The system can fail to learn what its human designers want it to learn and instead learn a different, but correlated objective. A frequently cited example is a computer vision system which was supposed to classify pictures of tanks but instead learned to distinguish the backgrounds of the tank images [6]. Other examples<sup>6</sup> include problems caused by poorly-designed utility functions unintentionally rewarding partially desirable behaviors of agents, such as riding a bicycle in circles around the target never reaching the target as to maximize reward [7], pausing a game to avoid losing [8], or repeatedly touching a soccer ball to get credit for possession [9]. During the performance phase, the system may succumb to a number of possible causes [10-12] all leading to an AI product failure.

Media reports occasionally mention examples of AI product failures but most of these examples can be attributed to non-AI specific causes on closer examination, such as bugs in the code or mistakes in design. The list below is curated to only mention failures of intended intelligence. Additionally, the examples below include only the first occurrence of a particular failure, but the same problems are frequently observed again in later years. Finally the list does not include AI Failures due to hacking or other intentional causes. Still, the timeline of AI Failures has an exponential trend while implicitly indicating historical events such as “AI Winter”:

- 1958 Advice software deduced inconsistent sentences using logical programming [13].
- 1959 AI designed to be a General Problem Solver failed to solve real world problems.<sup>7</sup>
- 1977 Story writing software with limited common sense produced “wrong” stories [14].
- 1982 Software designed to make discoveries, discovered how to cheat instead.<sup>8</sup>
- 1983 Nuclear attack early warning system falsely claimed that an attack is taking place.<sup>9</sup>
- 1984 The National Resident Match program was biased in placement of married couples [15].
- 1988 Admissions software discriminated against women and minorities [16].
- 1994 Agents learned to “walk” quickly by becoming taller and falling over [17].
- 2001 AI agents learned to associate other AIs to food and virtually cannibalized them.<sup>10</sup>
- 2005 Personal assistant AI rescheduled a meeting 50 times, each time by 5 minutes [18].
- 2006 Insider threat detection system classified normal activities as outliers [19].
- 2006 Investment advising software was losing money in real trading [20].

---

<sup>5</sup> [https://en.wikipedia.org/wiki/Kenji\\_Urada](https://en.wikipedia.org/wiki/Kenji_Urada)

<sup>6</sup> [http://lesswrong.com/lw/lvh/examples\\_of\\_ais\\_behaving\\_badly/](http://lesswrong.com/lw/lvh/examples_of_ais_behaving_badly/)

<sup>7</sup> [https://en.wikipedia.org/wiki/General\\_Problem\\_Solver](https://en.wikipedia.org/wiki/General_Problem_Solver)

<sup>8</sup> <http://aliciapatterson.org/stories/eurisko-computer-mind-its-own>

<sup>9</sup> [https://en.wikipedia.org/wiki/1983\\_Soviet\\_nuclear\\_false\\_alarm\\_incident](https://en.wikipedia.org/wiki/1983_Soviet_nuclear_false_alarm_incident)

<sup>10</sup> <https://blog.statsbot.co/creepy-artificial-intelligence-ebc3f76179a8>

- 2007 Google search engine returns unrelated results for some keywords.<sup>11</sup>  
2010 Complex AI stock trading software caused a trillion dollar flash crash.<sup>12</sup>  
2011 E-Assistant told to “call me an ambulance” began to refer to the user as Ambulance.<sup>13</sup>  
2013 Object recognition neural networks saw phantom objects in particular noisy images [21].  
2013 Google software engaged in name-based discrimination in online ad delivery [22].  
2014 Search engine autocomplete made bigoted associations about groups of users [23].  
2014 Smart fire alarm failed to sound alarm during fire.<sup>14</sup>  
2015 Automated email reply generator created inappropriate responses.<sup>15</sup>  
2015 A robot for grabbing auto parts grabbed and killed a man.<sup>16</sup>  
2015 Image tagging software classified black people as gorillas.<sup>17</sup>  
2015 Medical Expert AI classified patients with asthma as lower risk [24].  
2015 Adult content filtering software failed to remove inappropriate content.<sup>18</sup>  
2015 Amazon’s Echo responded to commands from TV voices.<sup>19</sup>  
2016 LinkedIn’s name lookup suggests male names in place of female ones.<sup>20</sup>  
2016 AI designed to predict recidivism acted racist.<sup>21</sup>  
2016 AI agent exploited reward signal to win without completing the game course.<sup>22</sup>  
2016 Passport picture checking system flagged Asian user as having closed eyes.<sup>23</sup>  
2016 Game NPCs designed unauthorized superweapons.<sup>24</sup>  
2016 AI judged a beauty contest and rated dark-skinned contestants lower.<sup>25</sup>  
2016 Smart contract permitted syphoning of funds from the DAO.<sup>26</sup>  
2016 Patrol robot collided with a child.<sup>27</sup>  
2016 World champion-level Go playing AI lost a game.<sup>28</sup>  
2016 Self driving car had a deadly accident.<sup>29</sup>  
2016 AI designed to converse with users on Twitter became verbally abusive.<sup>30</sup>  
2016 Google image search returned racists results.<sup>31</sup>  
2016 Artificial applicant failed to pass university entrance exam.<sup>32</sup>

---

<sup>11</sup> [https://en.wikipedia.org/wiki/Google\\_bomb](https://en.wikipedia.org/wiki/Google_bomb)

<sup>12</sup> <http://gawker.com/this-program-that-judges-use-to-predict-future-crimes-s-1778151070>

<sup>13</sup> <https://www.technologyreview.com/s/601897/tougher-turing-test-exposes-chatbots-stupidity/>

<sup>14</sup> <https://www.forbes.com/sites/aarontilley/2014/04/03/googles-nest-stops-selling-its-smart-smoke-alarm-for-now>

<sup>15</sup> <https://gmail.googleblog.com/2015/11/computer-respond-to-this-email.html>

<sup>16</sup> <http://time.com/3944181/robot-kills-man-volkswagen-plant/>

<sup>17</sup> [http://www.huffingtonpost.com/2015/07/02/google-black-people-goril\\_n\\_7717008.html](http://www.huffingtonpost.com/2015/07/02/google-black-people-goril_n_7717008.html)

<sup>18</sup> <http://blogs.wsj.com/digits/2015/05/19/googles-youtube-kids-app-criticized-for-inappropriate-content/>

<sup>19</sup> [https://motherboard.vice.com/en\\_us/article/53dz8x/people-are-complaining-that-amazon-echo-is-responding-to-ads-on-tv](https://motherboard.vice.com/en_us/article/53dz8x/people-are-complaining-that-amazon-echo-is-responding-to-ads-on-tv)

<sup>20</sup> <https://www.seattletimes.com/business/microsoft/how-linkedins-search-engine-may-reflect-a-bias>

<sup>21</sup> <http://gawker.com/this-program-that-judges-use-to-predict-future-crimes-s-1778151070>

<sup>22</sup> <https://openai.com/blog/faulty-reward-functions>

<sup>23</sup> <http://www.telegraph.co.uk/technology/2016/12/07/robot-passport-checker-rejects-asian-mans-photo-having-eyes>

<sup>24</sup> <http://www.kotaku.co.uk/2016/06/03/elites-ai-created-super-weapons-and-started-hunting-players-skynet-is-here>

<sup>25</sup> <https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people>

<sup>26</sup> [https://en.wikipedia.org/wiki/The\\_DAO\\_\(organization\)](https://en.wikipedia.org/wiki/The_DAO_(organization))

<sup>27</sup> <http://www.latimes.com/local/lanow/la-me-ln-crimefighting-robot-hurts-child-bay-area-20160713-snap-story.html>

<sup>28</sup> <https://www.engadget.com/2016/03/13/google-alphago-loses-to-human-in-one-match/>

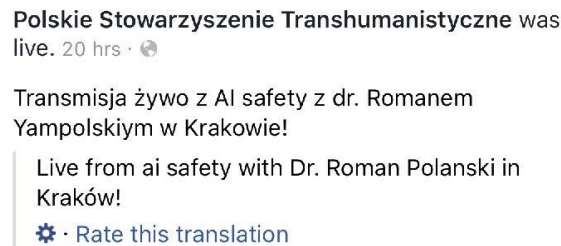
<sup>29</sup> <https://www.theguardian.com/technology/2016/jul/01/tesla-driver-killed-autopilot-self-driving-car-harry-potter>

<sup>30</sup> <http://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>

<sup>31</sup> <https://splinternews.com/black-teenagers-vs-white-teenagers-why-googles-algori-1793857436>

<sup>32</sup> <https://www.japantimes.co.jp/news/2016/11/15/national/ai-robot-fails-get-university-tokyo>

2016 Predictive policing system disproportionately targets minority neighborhoods.<sup>33</sup>  
2016 Text subject classifier failed to learn relevant features for topic assignment [25].  
2017 AI for making inspirational quotes failed to inspire with gems like “Keep Panicking”.<sup>34</sup>  
2017 Alexa played adult content instead of song for kids.<sup>35</sup>  
2017 Cellphone case designing AI utilized inappropriate images.<sup>36</sup>  
2017 Pattern recognition software failed to recognize certain types of inputs.<sup>37</sup>  
2017 Debt recovery system miscalculated amounts owed.<sup>38</sup>  
2017 Russian language chatbot shared pro-Stalinist, pro-abuse and pro-suicide views.<sup>39</sup>  
2017 Translation AI stereotyped careers to specific genders [26].  
2017 Face beautifying AI made black people look white.<sup>40</sup>  
2017 Google’s sentiment analyzer became homophobic and anti-Semitic.<sup>41</sup>  
2017 Fish recognition program learned to recognize boat IDs instead.<sup>42</sup>  
2017 Billing software sent an electrical bill for 284 billion dollars.<sup>43</sup>  
2017 Alexa turned on loud music at night without being prompted to do so.<sup>44</sup>  
2017 AI for writing Christmas carols produced nonsense.<sup>45</sup>  
2017 Autonomous cars had double the number of “fender-benders” of conventional cars [27].  
2017 Apple’s face recognition system failed to distinguish Asian users.<sup>46</sup>  
2017 Facebook’s translation software changed Yampolskiy to Polanski, see Figure 1.



Polskie Stowarzyszenie Transhumanistyczne was  
live. 20 hrs · 🌐

Transmisja żywo z AI safety z dr. Romanem  
Yampolskiym w Krakowie!

Live from ai safety with Dr. Roman Polanski in  
Kraków!

🔧 · Rate this translation

**Figure 1** While translating from Polish to English Facebook’s software changed Roman “Yampolskiy” to Roman “Polanski” due to statistically higher frequency of the later name in sample texts.

<sup>33</sup> <https://www.themarshallproject.org/2016/02/03/policing-the-future>

<sup>34</sup> <https://www.buzzworthy.com/ai-tries-to-generate-inspirational-quotes-and-gets-it-hilariously-wrong>

<sup>35</sup> <https://www.entrepreneur.com/video/287281>

<sup>36</sup> <https://www.boredpanda.com/funny-amazon-ai-designed-phone-cases-fail>

<sup>37</sup> <http://www.bbc.com/future/story/20170410-how-to-fool-artificial-intelligence>

<sup>38</sup> <http://www.abc.net.au/news/2017-04-10/centrelink-debt-recovery-system-lacks-transparency-ombudsman/8430184>

<sup>39</sup> <https://techcrunch.com/2017/10/24/another-ai-chatbot-shown-spouting-offensive-views>

<sup>40</sup> <http://www.gizmodo.co.uk/2017/04/faceapp-blames-ai-for-whitening-up-black-people>

<sup>41</sup> [https://motherboard.vice.com/en\\_us/article/j5jmj8/google-artificial-intelligence-bias](https://motherboard.vice.com/en_us/article/j5jmj8/google-artificial-intelligence-bias)

<sup>42</sup> <https://medium.com/@gidishperber/what-ive-learned-from-kaggle-s-fisheries-competition-92342f9ca779>

<sup>43</sup> <https://www.washingtonpost.com/news/business/wp/2017/12/26/woman-gets-284-billion-electric-bill-wonders-whether-its-her-christmas-lights>

<sup>44</sup> <http://mashable.com/2017/11/08/amazon-alexa-rave-party-germany>

<sup>45</sup> <http://mashable.com/2017/12/22/ai-tried-to-write-christmas-carols>

<sup>46</sup> <http://www.mirror.co.uk/tech/apple-accused-racism-after-face-11735152>

Spam filters block important emails, GPS provides faulty directions, machine translation corrupts meaning of phrases, autocorrect replaces desired word with a wrong one, biometric systems misrecognize people, transcription software fails to capture what is being said; overall, it is harder to find examples of AIs that don't fail. Depending on what we consider for inclusion as examples of problems with intelligent software, the list of examples could be grown almost indefinitely. In its most extreme interpretation, any software with as much as an "if statement" can be considered a form of Narrow Artificial Intelligence (NAI) and all of its bugs are thus examples of AI Failure<sup>47</sup>.

Analyzing the list of NAI, from the inception of the field to modern day systems, we can arrive at a simple generalization: An AI designed to do X will eventually fail to do X. While it may seem trivial, it is a powerful generalization tool, which can be used to predict future failures of NAIs. For example, looking at cutting-edge current and future AIs we can predict that:

- Software for generating jokes will occasionally fail to make them funny.
- Sarcasm detection software will confuse sarcastic and sincere statements.
- Video description software will misunderstand movie plots.
- Software generated virtual worlds may not be compelling.
- AI doctors will misdiagnose some patients in a way a real doctor would not.
- Employee screening software will be systematically biased and thus hire low performers.
- Mars robot explorer will misjudge its environment and fall into a crater.
- Etc.

Others have given the following examples of possible accidents with future artificial general intelligence (AGI) or superintelligence:

- Housekeeping robot cooks family pet for dinner.<sup>48</sup>
- A mathematician AGI converts all matter into computing elements to solve problems.<sup>49</sup>
- An AGI running simulations of humanity creates conscious beings who suffer [28].
- Paperclip manufacturing AGI fails to stop and converts universe into raw materials [29].
- A scientist AGI performs experiments with significant negative impact on biosphere [30].
- Drug design AGI develops time-delayed poison to kill everyone and so defeats cancer.<sup>50</sup>
- Future superintelligence optimizes away all consciousness.<sup>51</sup>
- AGI kills humanity and converts universe into materials for improved penmanship.<sup>52</sup>
- AGI designed to maximize human happiness tiles universe with tiny smiley faces [31].
- AGI instructed to maximize pleasure consigns humanity to a dopamine drip [32].
- Superintelligence may rewire human brains to increase their perceived satisfaction [31].

---

<sup>47</sup> [https://en.wikipedia.org/wiki/List\\_of\\_software\\_bugs](https://en.wikipedia.org/wiki/List_of_software_bugs)

<sup>48</sup> <https://www.theguardian.com/sustainable-business/2015/jun/23/the-ethics-of-ai-how-to-stop-your-robot-cooking-your-cat>

<sup>49</sup> <https://intelligence.org/2014/11/18/misconceptions-edge-orgs-conversation-myth-ai>

<sup>50</sup> <https://80000hours.org/problem-profiles/positively-shaping-artificial-intelligence>

<sup>51</sup> <http://slatestarcodex.com/2014/07/13/growing-children-for-bostroms-disneyland>

<sup>52</sup> <https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html>



Denning and Denning made some similar error extrapolations in their humorous paper on “artificial stupidity” [33]: “Soon the automated DEA started closing down pharmaceutical companies saying they were dealing drugs. The automated FTC closed down the Hormel Meat Company, saying it was purveying spam. The automated DOJ shipped Microsoft 500,000 pinstriped pants and jackets, saying it was filing suits. The automated Army replaced all its troops with a single robot, saying it had achieved the Army of One. The automated Navy, in a cost saving move, placed its largest-ever order for submarines with Subway Sandwiches. The FCC issued an order for all communications to be wireless, causing thousands of AT&T installer robots to pull cables from overhead poles and underground conduits. The automated TSA flew its own explosives on jetliners, citing data that the probability of two bombs on an airplane is exceedingly small.”

AGI can be seen as a superset of all NAIs and so will exhibit a superset of failures as well as more complicated failures resulting from the combination of failures of individual NAIs and new super-failures, possibly resulting in an existential threat to humanity or at least an AGI takeover. In other words, AGIs can make mistakes influencing everything. Overall, we predict that AI Failures and premediated Malevolent AI incidents will increase in absolute frequency and severity proportionate to AIs’ capability.

## 2.1 Preventing AI Failures

AI failures have a number of causes, with most common ones currently observed, displaying some type of algorithmic bias, poor performance or basic malfunction. Future AI failures are likely to be more severe including purposeful manipulation/deception of people [34], or even resulting in human death (likely from misapplication of militarized AI / autonomous weapons / killer robots [35]). At the very end of severity scale, we see existential risk scenarios resulting in extermination of human kind or suffering-risk scenarios [36] resulting in large scale torture of humanity, both types of risk coming from super-capable artificially intelligent systems. Specific preventative measures for failures of superintelligence are outside of the scope of this work, but are investigated in much of AI Safety literature.

Reviewing reported examples of AI accidents, we can notice patterns of failure, which can be attributed to the following causes:

- Biased data, including cultural differences.
- Deploying underperforming system.
- Non-representative training data.
- Discrepancy between training and testing data.
- Rule overgeneralization or application of population statistics to individuals.
- Inability to handle noise or statistical outliers.
- Not testing for rare or extreme conditions.
- Not realizing an alternative solution method can produce same results, but with side effects.
- Letting users control data or learning process.

- No security mechanism to prevent adversarial meddling.
- No cultural competence or common sense.
- Limited access to information/sensors.
- Mistakes in design and inadequate testing.
- Limited ability for language disambiguation.
- Inability to adopt to changes in environment.

## 2.2 Preventing Algorithmic Bias

With bias being the most common cause of failure, we have observed in historical examples, it is helpful to analyze particular types of algorithmic bias. Friedman and Nissenbaum [15] proposed the following framework for analyzing bias in computer systems. They subdivide causes of bias into three categories – preexisting bias, technical bias and emergent bias.

- **Preexisting bias** reflects bias in society and social institutions, practices and attitudes. The system simply preserves an existing state of the world and automates application of bias as it currently exists.
- **Technical bias** appears because of hardware or software limitations of the system itself.
- **Emergent bias** emerges after the system is deployed due to changing societal standards.

Friedman and Nissenbaum suggest the following preventative measures against algorithmic bias [15]:

*“Toward minimizing preexisting bias, designers must not only scrutinize the design specifications, but must couple this scrutiny with a good understanding of relevant biases out in the world. The time to begin thinking about bias is in the earliest stages of the design process, when negotiating the system’s specifications with the client. Common biases might occur to populations based on cultural identity, class, gender, literacy (literate/less literate), handedness (right-handed/left-handed), and physical disabilities (e.g., being blind, color-blind, or deaf). As the computing community develops an understanding of these biases, we can correspondingly develop techniques to avoid or minimize them. Some current computer systems, for instance, address the problem of handedness by allowing the user to toggle between a right- or left-handed configuration for user input and screen display. Similarly, systems could minimize bias due to color blindness by encoding information not only in hue, but in its intensity, or in some other way by encoding the same information in a format unrelated to color. In addition, it can prove useful to identify potential user populations which might otherwise be overlooked and include representative individuals in the field test groups. Rapid prototyping, formative evaluation, and field testing with such well-conceived populations of users can be an effective means to detect unintentional biases throughout the design process.*

*Technical bias also places the demand on a designer to look beyond the features internal to a system and envision it in a context of use. Toward preventing technical bias, a designer must envision the design, the algorithms, and the interfaces in use so that technical decisions do not run at odds with moral values. Consider even the largely straightforward problem of whether to display a list with random entries or sorted alphabetically. In determining a solution, a designer might need to weigh considerations of ease of access enhanced by a sorted list against equity of access supported by a random list.*

*Minimizing emergent bias asks designers to envision not only a system’s intended situations of use, but to account for increasingly diverse social contexts of use. From a practical standpoint, however, such a proposal cannot be pursued in an unbounded manner. Thus, how much diversity in social context is enough, and what sort of diversity? While the question merits a lengthy discussion, we offer here but three suggestions. First, designers should reasonably anticipate probable contexts of use and design for these.*



*Second, where it is not possible to design for extended contexts of use, designers should attempt to articulate constraints on the appropriate contexts of a system's use. As with other media, we may need to develop conventions for communicating the perspectives and audience assumed in the design. Thus, if a particular expert system because of its content matter, goals, and design requires expert users to be used effectively, this constraint should be stated clearly in a salient place and manner, say, on one of the initial screens. Third, system designers and administrators can take responsible action if bias emerges with changes in context."*

Osaba and Welsler review additional remedies [37]:

#### ***"Statistical and Algorithmic Approaches***

*There is a growing field focused on fair, accountable, and transparent machine learning, working on technical approaches to assuring algorithmic fairness or certifying and correcting disparate impact in machine learning algorithms. Dwork et al. (2012) proposed using modified distance or similarity metrics when working with subject data. These similarity metrics are meant to enforce rigorous fairness constraints when comparing subjects in data sets. Sandvig et al. (2014) proposed a number of algorithms auditing procedures that compare algorithmic output with expected equitable behavior. Algorithm audits can be more feasible and thorough when algorithm codes and procedures are open sourced. DeDeo (2015) introduced an algorithmic approach to ensuring that machine learning models enforce statistical independence between outcomes and protected variables. Feldman et al. (2015) introduced a test for checking whether an algorithm violates legal disparate impact rules (under U.S. law). This provides a socially informed metric of optimality. They also proposed a statistical method for correcting such inequities in classification algorithms. Yet, there is a drawback: These schemes will often trade some predictive power for fairness.*

#### ***Causal Reasoning Algorithms***

*More broadly and on a longer time-scale, Judea Pearl (2009), Leon Bottou et al. (2013), and others (Athey, 2015) are exploring ways to equip machine learning algorithms with causal or counterfactual reasoning. This is extremely important because automated causal reasoning systems can present clear causal narratives for judging the quality of an algorithmic decision process. Accurate causal justifications for algorithmic decisions are the most reliable audit trails for algorithms. ... If we are to rely on algorithms for autonomous decisionmaking, they need to be equipped with tools for auditing the causal factors behind key decisions. Algorithms that can be audited for causal factors can give clearer accounts or justifications for their outcomes. This is especially important for justifying statistically disproportionate outcomes.*

#### ***Algorithmic Literacy and Transparency***

*Combating algorithmic bias would benefit from an educated public capable of understanding that algorithms can lead to inequitable outcomes. This is not the same as requiring that users understand the inner workings of all algorithms—this is not feasible. Just instilling a healthy dose of informed skepticism could be useful enough to reduce the effect size of automation bias. There is hope on this front. The sheer amount of time we spend interfacing with algorithms may make algorithmic missteps more noticeable. ... Combining algorithmic literacy with transparency could be very effective. Transparency in this space usually refers to making sure any algorithms in use are easily understood. Again, that is unlikely to be feasible all the time. What is feasible and useful is more disclosure of decisions and actions that are mediated by artificial agents. ...*

#### ***Personnel Approaches***

*The technical research on bias in machine learning and artificial intelligence algorithms is still in its infancy. Questions of bias and systemic errors in algorithms demand a different kind of wisdom from algorithm designers and data scientists. These practitioners are often engineers and scientists with less exposure to social or public policy questions. The demographics of algorithm designers are often less than diverse. Algorithm designers make myriad design choices, some of which may have far-reaching consequences. Diversity in the ranks of algorithm developers could help improve sensitivity to potential disparate impact problems.*

Many of the observed AI failures are similar to incidents experienced by children. This is particularly true for artificial neural networks, which are at the cutting edge of Machine Learning (ML). One can say that children are untrained neural networks deployed on real data and observing them can teach us a lot about predicting and preventing AI fails. A number of research groups [30, 38] have investigated types of ML failure and here we have summarized their work and mapped it onto similar situations with children:

- Negative side effects – child makes a mess.
- Reward hacking – child finds candy jar.
- Scalable oversight – babysitting should not require a team of 10.
- Safe exploration – no fingers in the outlet.
- Robustness to distributional shift – use “inside voice” in the classroom.
- Inductive ambiguity identification – is ant a cat or a dog?
- Robust human imitation – daughter shaves like daddy.
- Informed oversight – let me see your homework.
- Generalizable environmental goals – ignore that mirage.
- Conservative concepts – that dog has no tail.
- Impact measures – keep a low profile.
- Mild optimization – don’t be a perfectionist.
- Averting instrumental incentives – be an altruist.

Majority of research currently taking place to prevent such failures is currently happening under the label of “AI Safety”.

### 3. AI Safety and Security

In 2010, Roman Yampolskiy coined the phrase “Artificial Intelligence Safety Engineering” and its shorthand notation “AI Safety” to give a name to a new direction of research he was advocating. He formally presented his ideas on AI Safety at a peer-reviewed conference in 2011 [39], with subsequent publications on the topic in 2012 [40], 2013 [41, 42], 2014 [43], 2015 [44], 2016 [10, 11]. It is possible that someone used the phrase informally before, but to the best of our knowledge, Yampolskiy is the first to use it<sup>53</sup> in a peer-reviewed publication and to bring it popularity. Before that the most common names for the relevant concepts were “Machine Ethics” [45] or “Friendly AI” [46]. Today the term “AI Safety” appears to be the accepted<sup>54,55,56,57,58,59,60,61,62,63,64</sup> name for the field used by a majority of top researchers [38]. The

---

<sup>53</sup> Term “Safe AI” has been used as early as 1995, see Rodd, M. (1995). “Safe AI—is this possible?” *Engineering Applications of Artificial Intelligence* 8(3): 243-250.

<sup>54</sup> <https://www.cmu.edu/safartint/>

<sup>55</sup> <https://selfawarenessystems.com/2015/07/11/formal-methods-for-ai-safety/>

<sup>56</sup> <https://intelligence.org/2014/08/04/groundwork-ai-safety-engineering/>

<sup>57</sup> <http://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/new-ai-safety-projects-get-funding-from-elon-musk>

<sup>58</sup> <http://globalprioritiesproject.org/2015/08/quantifyingaisafety/>

<sup>59</sup> <http://futureoflife.org/2015/10/12/ai-safety-conference-in-puerto-rico/>

<sup>60</sup> <http://rationality.org/waiss/>

<sup>61</sup> <http://gizmodo.com/satya-nadella-has-come-up-with-his-own-ai-safety-rules-1782802269>

<sup>62</sup> <https://80000hours.org/career-reviews/artificial-intelligence-risk-research/>

<sup>63</sup> <https://openai.com/blog/concrete-ai-safety-problems/>

field itself is becoming mainstream despite being regarded as either science fiction or pseudoscience in its early days.

Despite the cited research, our legal and regulatory systems are behind the technological abilities and the field of machine morals is in its infancy. The problem of controlling intelligent machines is just now being recognized<sup>65</sup> as a serious concern and many researchers are still skeptical about its very premise. Worse yet, only about 100 people around the world are fully emerged in working on addressing the current limitations in our understanding and abilities in this domain. Only about a dozen<sup>66</sup> of those have formal training in computer science, cybersecurity, cryptography, decision theory, machine learning, formal verification, computer forensics, steganography, ethics, mathematics, network security, psychology and other relevant fields. It is not hard to see that the problem of making a safe and capable machine is much greater than the problem of making just a capable machine. Yet only about 1% of researchers are currently engaged in that problem with available funding levels below even that mark. As a relatively young and underfunded field of study, AI Safety can benefit from adopting methods and ideas from more established fields of science. Attempts have been made to introduce techniques, which were first developed by cybersecurity experts to secure software systems to this new domain of securing intelligent machines [47-50]. Other fields which could serve as a source of important techniques would include software engineering and software verification.

## References

- [1] N. Rychtyckyj and A. Turski, "Reasons for success (and failure) in the development and deployment of ai systems," in *AAAI 2008 workshop on What Went Wrong and Why*, 2008.
- [2] D. Shapiro and M. H. Goker, "Advancing ai research and applications by learning from what went wrong and why," *AI Magazine*, vol. 29, pp. 9-10, 2008.
- [3] A. Abecker, R. Alami, C. Baral, T. Bickmore, E. Durfee, T. Fong, *et al.*, "AAAI 2006 Spring Symposium Reports," *AI Magazine*, vol. 27, p. 107, 2006.
- [4] C. Marling and D. Chelberg, "RoboCup for the Mechanically, Athletically and Culturally Challenged," 2008.
- [5] S. Shalev-Shwartz, O. Shamir, and S. Shammah, "Failures of gradient-based deep learning," in *International Conference on Machine Learning*, 2017, pp. 3067-3075.
- [6] E. Yudkowsky, "Artificial intelligence as a positive and negative factor in global risk," *Global catastrophic risks*, vol. 1, p. 303, 2008.
- [7] J. Randlev and P. Alstrøm, "Learning to Drive a Bicycle Using Reinforcement Learning and Shaping," in *ICML*, 1998, pp. 463-471.
- [8] T. M. VII, "The first level of Super Mario Bros. is easy with lexicographic orderings and time travel," *The Association for Computational Heresy (SIGBOVIK) 2013*, 2013.
- [9] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *ICML*, 1999, pp. 278-287.
- [10] R. V. Yampolskiy, "Taxonomy of Pathways to Dangerous Artificial Intelligence," in *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [11] F. Pistono and R. V. Yampolskiy, "Unethical Research: How to Create a Malevolent Artificial Intelligence," *arXiv preprint arXiv:1605.02817*, 2016.

---

<sup>64</sup> [http://lesswrong.com/lw/n4/safety\\_engineering\\_target\\_selection\\_and\\_alignment/](http://lesswrong.com/lw/n4/safety_engineering_target_selection_and_alignment/)

<sup>65</sup> <https://www.whitehouse.gov/blog/2016/05/03/preparing-future-artificial-intelligence> the

<sup>66</sup> <http://acritch.com/fhi-positions/>

- [12] P. Scharre, "Autonomous Weapons and Operational Risk," presented at the Center for a New American Society, Washington DC, 2016.
- [13] C. Hewitt, "Development of Logic Programming: What went wrong, What was done about it, and What it might mean for the future."
- [14] J. R. Meehan, "TALE-SPIN, An Interactive Program that Writes Stories," in *IJCAI*, 1977, pp. 91-98.
- [15] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Transactions on Information Systems (TOIS)*, vol. 14, pp. 330-347, 1996.
- [16] S. Lowry and G. Macpherson, "A blot on the profession," *British medical journal (Clinical research ed.)*, vol. 296, p. 657, 1988.
- [17] K. Sims, "Evolving virtual creatures," in *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, 1994, pp. 15-22.
- [18] M. Tambe, "Electric elves: What went wrong and why," *AI magazine*, vol. 29, p. 23, 2008.
- [19] A. Liu, C. E. Martin, T. Hetherington, and S. Matzner, "AI Lessons Learned from Experiments in Insider Threat Detection," in *AAAI Spring Symposium: What Went Wrong and Why: Lessons from AI Research and Applications*, 2006, pp. 49-55.
- [20] J. Gunderson and L. Gunderson, "And Then the Phone Rang," in *AAAI Spring Symposium: What Went Wrong and Why: Lessons from AI Research and Applications*, 2006, pp. 13-18.
- [21] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, *et al.*, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.
- [22] L. Sweeney, "Discrimination in online ad delivery," *Queue*, vol. 11, p. 10, 2013.
- [23] N. Diakopoulos, "Algorithmic defamation: the case of the shameless autocomplete," *Tow Center for Digital Journalism*, 2014.
- [24] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1721-1730.
- [25] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135-1144.
- [26] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, pp. 183-186, 2017.
- [27] F. M. Favarò, N. Nader, S. O. Eurich, M. Tripp, and N. Varadaraju, "Examining accident reports involving autonomous vehicles in California," *PLoS one*, vol. 12, p. e0184952, 2017.
- [28] S. Armstrong, A. Sandberg, and N. Bostrom, "Thinking inside the box: Controlling and using an oracle ai," *Minds and Machines*, vol. 22, pp. 299-324, 2012.
- [29] N. Bostrom, "Ethical issues in advanced artificial intelligence," *Science Fiction and Philosophy: From Time Travel to Superintelligence*, pp. 277-284, 2003.
- [30] J. Taylor, E. Yudkowsky, P. LaVictoire, and A. Critch, "Alignment for advanced machine learning systems," *Machine Intelligence Research Institute*, 2016.
- [31] E. Yudkowsky, "Complex value systems in friendly AI," *Artificial general intelligence*, pp. 388-393, 2011.
- [32] G. Marcus, "Moral machines," *The New Yorker*, vol. 24, 2012.
- [33] D. E. Denning and P. J. Denning, "Artificial stupidity," 2004.
- [34] M. Chessen, "The MADCOM Future," *Atlantic Council*, pp. Available at: <http://www.atlanticcouncil.org/publications/reports/the-madcom-future>, 2017.
- [35] A. Krishnan, *Killer robots: legality and ethicality of autonomous weapons*: Ashgate Publishing, Ltd., 2009.

- [36] L. Gloor, "Suffering-focused AI safety: Why "fail-safe" measures might be our top intervention," Tech. rep. FRI-16-1. Foundational Research Institute. url: <https://foundationalresearch.org/wp-content/uploads/2016/08/Suffering-focused-AI-safety.pdf>2016.
- [37] O. A. Osoba and W. Welser IV, *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*: Rand Corporation, 2017.
- [38] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016.
- [39] R. V. Yampolskiy, "Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach," presented at the Philosophy and Theory of Artificial Intelligence (PT-AI2011), Thessaloniki, Greece, October 3-4, 2011.
- [40] R. V. Yampolskiy and J. Fox, "Safety Engineering for Artificial General Intelligence," *Topoi. Special Issue on Machine Ethics & the Ethics of Building Intelligent Machines*, 2012.
- [41] L. Muehlhauser and R. Yampolskiy, "Roman Yampolskiy on AI Safety Engineering," presented at the Machine Intelligence Research Institute, Available at: <http://intelligence.org/2013/07/15/roman-interview/> July 15, 2013.
- [42] R. V. Yampolskiy, "Artificial intelligence safety engineering: Why machine ethics is a wrong approach," in *Philosophy and Theory of Artificial Intelligence*, ed: Springer Berlin Heidelberg, 2013, pp. 389-396.
- [43] A. M. Majot and R. V. Yampolskiy, "AI safety engineering through introduction of self-reference into felicific calculus via artificial pain and pleasure," in *IEEE International Symposium on Ethics in Science, Technology and Engineering*, Chicago, IL, May 23-24, 2014, pp. 1-6.
- [44] R. V. Yampolskiy, *Artificial Superintelligence: a Futuristic Approach*: Chapman and Hall/CRC, 2015.
- [45] J. H. Moor, "The nature, importance, and difficulty of machine ethics," *IEEE intelligent systems*, vol. 21, pp. 18-21, 2006.
- [46] E. Yudkowsky, "Creating friendly AI 1.0: The analysis and design of benevolent goal architectures," *Singularity Institute for Artificial Intelligence, San Francisco, CA, June*, vol. 15, 2001.
- [47] R. Yampolskiy, "Leakproofing the Singularity Artificial Intelligence Confinement Problem," *Journal of Consciousness Studies*, vol. 19, pp. 1-2, 2012.
- [48] J. Babcock, J. Kramar, and R. Yampolskiy, "The AGI Containment Problem," *arXiv preprint arXiv:1604.00545*, 2016.
- [49] J. Babcock, J. Kramar, and R. Yampolskiy, "The AGI Containment Problem," in *The Ninth Conference on Artificial General Intelligence (AGI2015)*, 2016.
- [50] S. Armstrong and R. V. Yampolskiy, "Security Solutions for Intelligent and Complex Systems," in *Security Solutions for Hyperconnectivity and the Internet of Things*, ed: IGI Global, 2016, pp. 37-88.





CONSORTIUM FOR  
**SAFER AI**



**A PLEDGE TO PUT  
SAFETY FIRST**



CONSORTIUM FOR  
**SAFER AI**

Contact us  
[infosafesai@gmail.com](mailto:infosafesai@gmail.com)  
[www.makingaisafer.org](http://www.makingaisafer.org)

Fostering the Common Cause of Safe  
Commercial AI-based technologies