AI & SOCIETY
https://doi.org/10.1007/s00146-022-01585-x

ORIGINAL PAPER

Omission and commission errors underlying AI failures

Sasanka Sekhar Chanda[1] · **Debarag Narayan Banerjee**[2]

Received: 15 January 2022 / Accepted: 21 October 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract
In this article we investigate origins of several cases of failure of Artificial Intelligence (AI) systems employing machine learning and deep learning. We focus on omission and commission errors in (a) the inputs to the AI system, (b) the processing logic, and (c) the outputs from the AI system. Our framework yields a set of 28 factors that can be used for reconstructing the path of AI failures and for determining corrective action. Our research helps identify emerging themes of inquiry necessary for developing more robust AI-ML systems. We are hopeful that our work will help strengthen the use of machine-learning AI by enhancing the rates of true positive and true negative judgements from AI systems, and by lowering the probabilities of false positive and false negative judgements.

Keywords AI failure · Commission errors · Deep learning · Machine learning · Omission errors

1 Introduction

Artificial Intelligence (AI) systems seek to employ computing machines, sensors, and other hardware to carry out tasks requiring reasoning, knowledge representation, planning, learning, natural language processing, and perception and tasks involving moving and manipulating objects (Russell and Norvig 2003). In addition to structured and unstructured data, information handled by AI now-a-days include images and video and audio streams. Instances of AI systems failing to deliver consistent, satisfactory performance are legion (Cai and Yuan 2021; Das 2020; Krauth 2018; Yampolskiy 2019). We investigate why AI failures occur, limiting our scope to AI systems operating on machine learning and deep learning technologies employing neural networks, or other similar difficult-to-explain machine-learnt algorithms.[1] Moreover we focus only on unexpected AI failures or "engineering mistakes" (Yampolskiy 2019) relative to the context or environment an AI system functions. Stated differently, an important topic of misuse of AI to harm humans is outside the scope of this study.

AI systems can fail (a) if there are problems with its inputs comprising various representations of data, sensor hardware, etc. and/or (b) if the processing logic is deficient in some way and/or (c) if the repertoire of actions available to the AI system is inadequate, i.e. if the output is inappropriate. Further, these problems/deficiencies/inadequacies originate from two kinds of errors—commission and omission errors (Ackoff 1994)—in the design, development and deployment of an AI system. These errors are defined as follows:

1. Error of commission: doing something that should not have been done.
2. Error of omission: not doing something that should have been done.

In layman terms, the errors of omission (EOO) map to design flaws; the errors of commission (EOC) map to implementation flaws. We may note though, that there is a good deal of reciprocal interdependence between design and implementation. Thus, in our study, the empirical instances identified as EOO and EOC are not airtight containers. Rather, they furnish useful categories to channelize

✉ Sasanka Sekhar Chanda
sschanda@iimidr.ac.in

Debarag Narayan Banerjee
debarag.banerjee@agoda.com

[1] Strategic Management, Indian Institute of Management Indore, C101 Acad Block IIM Indore, Rau-Pithampur Rd., Indore, MP 453556, India

[2] MachinAnimus Consulting, Los Altos, CA, USA

[1] In neural networks complicated reasoning is used to identify patterns. The usage context of those patterns is memorized. The AI system adjusts the way it operates as it accumulates data, in the process appearing to learn from experience (Maeda and Parker 2003).

Published online: 17 November 2022

Table 1 Omission and commission errors underlying AI failures

Stage	Omission error	Commission error
Input	OI1. Some scenarios needing decision-making by the AI got excluded OI2. Fine-grained/less frequent variations of some known/frequent scenarios got excluded OI3. Some channels of relevant and useful information got missed out OI4. All salient info is not taken in OI5. Training data did not contain sufficient number of records for diverse constituents OI6. Data that was input to the AI during training phase has since become obsolete (making the AI prediction model obsolete)	CI1. Hardware deployed for capturing an input is deficient CI2. In the step for reduction of data dimension reduction, some correlated columns got included, either due to a mistake, or due to a change in the external environment after design and deployment of the AI solution CI3. Irrelevant data columns got fed as inputs to the AI solution CI4. Irrelevant input (noise: environmental or from other connected systems) not blocked off CI5. Feedback loops that exacerbate biases present in the AI solution
Process	OP1. Logic in AI failed to focus on adequately processing a salient part of input information OP2. Logic in AI failed to resolve ambiguity in input information OP3. Logic in AI failed to stop execution immediately upon encountering error condition OP4. Inadequate (insufficient), or slow, processing by an AI system	CP1. Logic in AI focused on non-salient part of input information CP2. An inappropriate logic was used CP3. Ill-conceived updates or features from the originators of the AI system break functioning CP4. Logic necessary for some part of the AI functioning got deployed in another part where it is not required
Output	OO1. Failure arising from lack of technological sophistication OO2. Failure to design new kinds of actions from AI, when environment changes OO3. Failing to send output to intended party/parties, i.e. expected action by the AI system fails to materialize timely OO4. Some relevant ways of deciding on an action (for example, consulting a human, asking another AI system) not made available to the AI system	CO1. Adverse interaction between modules of an AI system leads to faulty action CO2. Inappropriate action stemming from adverse AI-environment interaction CO3. Questionable action stemming from adverse interaction between AI and other technological systems CO4. Inappropriate action stemming from adverse interaction between AI and humans CO5. AI learns biased/inappropriate behavior in the wild (from a mix of humans, corpuses including web sites and social media, connected technology systems and other environmental entities)

thinking. Further, it is probable that ours is not an exhaustive list of all possible flavors of EOO and EOC arising in machine-learning-based AI systems. We highlight common or general causes of certain errors across machine-learning AI use cases—image recognition, text and natural language processing, driverless vehicles, computer vision, and robotics. This exposition should help Practitioners look beyond the technology and particular use cases, and nudges the field as a whole to evolve better standards for AI systems design, development and deployment.

2 Framework

We characterize AI failure as any inappropriate behavior from an AI system. An AI system's behavior is judged as inappropriate when a decision from the AI system goes contrary to the objectives the system was designed for. Yampolskiy (2019, p. 3) describes the situations of interest in this study in an apt manner: "An AI designed to do X will eventually fail to do X". In Table 1 we present a framework for classifying causes of AI failures as arising from omission and commission errors in the input, processing and outputs of the AI system. We provide illustrations, wherever feasible. An implicit assumption in our framework is that, we consider 'processing' errors only after ruling out existence of errors in 'input', and further that we consider errors in 'output' only after ruling out existence of errors in 'input' and in 'processing'.

We discuss how errors originate and illustrate with examples. We note that there can be multiple causes of an AI system failure. Moreover, the same problematic state of affairs can manifest in different ways. For example, a problem may arise from inadequate validation of input data, failing which it can manifest as a problem in processing; if the problem eludes resolution during processing, it can show up as improper output. To the extent possible the examples we provide pinpoint a particular error condition being highlighted.

2.1 Omission errors in AI input

We distinguish this error condition as one that will invariably need provisioning for some additional input to the AI system to address the problem situation observed, assuming that processing logic and repertoire of actions available from the AI system are adequate. In line with the definition of omission error—not doing something that should have been

done—we focus on instances where the absence of one or more inputs to the AI system is primarily responsible for an error condition. We discuss six distinct flavors of omission errors in AI input in the paragraphs that follow.

2.1.1 Some scenarios needing decision-making by the AI got excluded (OI1)

This happens when the AI system is unable to draw on a pre-configured recourse to handle a situation it encounters under conditions of live operation. One such instance of AI failure was observed in Las Vegas (O'Kane 2019). A driverless shuttle bus stopped upon noticing a delivery truck reversing on to a lane perpendicular to the bus's path. However, the clearance was insufficient, and the AI system of the shuttle bus failed to sound the horn to alert the driver of the truck. A human driver would have sounded a horn, anticipating a collision. Eventually the truck's wheel clipped the fender of the shuttle.

The police fined the delivery truck driver on the spot. Subsequent investigations revealed a different picture.

> The truck driver … did see the shuttle coming. But he told investigators that, after he saw it coming toward him, he figured it was a "reasonable assumption" that the shuttle "would stop a reasonable distance from a backing tractor trailer." So he turned away to keep an eye on a crossing pedestrian. When he made his next move, he scraped the shuttle.
> "I figured the thing was in control," he told investigators. "I figured they must have had the thing worked out; [that] it was going to function fine. I figured someone could stop it if need be."
> *Source:* O'Kane (2019)

Investigators also found that:

> … the operator had no immediate access to the manual controls for the shuttle … since the policy of <the> operator … was to lock that controller away in a storage compartment on the shuttle during rides. If the operator had been able to quickly access the controller he could have moved the shuttle out of the way of the truck, or at the very least triggered the horn to let the driver know that he was about to crash.
> *Source:* O'Kane (2019)

This is clearly a case where AI was unprepared for a situation where the AI-driven bus encounters a reversing long-vehicle. Absent manual control, beeping the horn of the AI bus was also not possible, and the engineer's shouts from inside the closed coach failed to reach the ears of the driver of the delivery truck (O'Kane 2019). A design incorporating more foresight would have provided for sounding the horn as the delivery truck got closer. Besides, better anticipation of the reversing trailer-truck's trajectory would have had the AI-driven shuttle bus stop earlier—affording greater clearance—obviating the possibility of the scrape.

We note though it is unrealistic to expect that the AI developers of driverless vehicles will think up all or majority of unusual situations that may potentially be encountered on the roads. Yet, the current practice of test-driving driverless vehicles for longer, across a variety of traffic conditions also appears to be insufficient as well. In our opinion, a lot more needs to be done. For example, millions of hours of traffic footage are being captured in traffic cameras in hundreds of cities worldwide. We recommend new AI systems be built that will go through this footage and identify unique traffic situations. For starters, inquiries can be made into instances of sharp change in speed or direction of movement of road-users (vehicles/humans/others), unusual instances of sounding of the horn, or significant changes in traffic flow patterns in a short time. Eventually these new AI systems should themselves discover the intelligence necessary to capture unusual traffic incidents. These incidents can then be simulated for driverless vehicles for training purposes.

We identify an emerging theme (**I**): to enable development of better AI systems in the future, it is necessary to use AI to source a vastly higher number of test scenarios, to augment the human-suggested test scenarios. In the Appendix we elaborate further, on this futuristic framework.

2.1.2 Fine-grained variations of some known scenarios got excluded (OI2)

This omission error originates in failure to design the AI system to cope appropriately with certain variations in a known scenario. For example, a 10-year old boy was able to unlock his mother's iPhone X using his own face (Greenberg 2017). This occurred when the mother registered her face (with iPhone X) under indoor, night-time lighting conditions—and did not occur when she registered her face under a different lighting condition.

In another incident, a Vietnamese security firm Bkav used a 3D-printed (face) mask to unlock an iPhone, overcoming iPhone's Face ID and bypassing iPhone X's attention detection safeguard—an optional safeguard that monitors a user's eyes to verify that they are looking at their phone before unlocking it (Campbell 2017). This seems to question the efficacy of Apple's efforts in training of the AI system: earlier the marketing head of Apple, Phil Schiller, said that the Apple "worked with professional mask makers and makeup artists in Hollywood" during development of the Face ID, even creating a collection of masks to train the iPhone X's neural network (Fingas 2017). Bkav suggests that their research demonstrates that Face ID is not secure enough to be used in business transactions. Relatedly, in another incident, the "liveness" detection feature of iPhone's Face

ID—which is used to confirm the person Face ID is looking at is real and not a mask or someone wearing prosthetics—was defeated with glasses and tape (Owen 2019).

In our view, similar to the previous case, it is unreasonable to expect that the developers of AI systems like Face ID will be in a position to anticipate all conditions of registering the Face ID and using different contrivances to trick the Face ID. In line with the emerging theme (**I**) we recommend that new AI systems be built to generate test scenarios for training AI applications like Face ID. These new AI systems need to analyze millions of selfies, vacation pictures, footage of attendance in conferences, rallies, sports games available in the public domain, and from targeted data collection programs like Pokemon Go, to determine a range of conditions to present to AI applications like Face ID for rigorous testing. To remain grounded in reality, we further need to keep in mind that multifactor authentication outperforms single-factor authentication in majority of known scenarios. Thus, when Face ID becomes better, it becomes a stronger pillar in multifactor authentication that makes a system "exponentially harder to crack" (Nachreiner 2021).

2.1.3 Some channels of relevant and useful information got missed out (OI3)

This omission error materializes when an AI system takes decisions based on insufficient inputs, i.e. misses some relevant channels of input. OpenAI's AI Language Model GPT-3—built on 175 billion parameters—gave following responses (Toews 2020):

Human: How many eyes does a giraffe have?
GPT-3: A giraffe has two eyes.
Human: How many eyes does my foot have?
GPT-3: Your foot has two eyes.
Human: How many eyes does a spider have?
GPT-3: A spider has eight eyes.
Human: How many eyes does the sun have?
GPT-3: The sun has one eye.
Human: How many eyes does a blade of grass have?
GPT-3: A blade of grass has one eye.
Source: Toews (2020)

The answers regarding the number of eyes of the foot and a blade of grass are definitely wrong—these entities do not have eyes. Thus, even though GPT-3 ingested a significant fraction of the text in the Internet, it failed to chance upon a usable schema for determining what entities do and do not have eyes.

The GPT-3 fiasco suggests second emerging theme (**II**): a need for creating repositories of obvious and obviously true information that has not been articulated on the Internet. When building the repository, one may imagine trying to educate an alien from a different galaxy about earth's civilization. Elementary level books used for educating children can help, possibly in conjunction with references to dictionaries. However, there are three hurdles. First, many text and images in use contains fantasies—e.g., Doremon and Nobita flying around, using bamboo-copters. Some means will become necessary to identify fantasies as a separate category. Second, and more difficult, there needs to be some mechanism to include facts and exclude opinions. This a herculean task—not less because any given articulation of a fact is unlikely to be free of the articulator's interpretation of links between data or information. It may be necessary to deploy trained researchers for this task. Third, it will be a tall order to even advocate keeping only objective information and keeping out the subjective. For example, sex of a baby at birth can objectively be male, female and intersex. However, a dominant strand of modern thinking does not consider this important, and advocates referring to gender instead. In this view, the subjective feeling of a human—whether he/she/they feel like a man/woman/non-binary—is considered more important. This brings us to the crux of the issue: if humans are unable to agree on definitions of basic categories, it is going to be difficult to agree on a fact being a fact, since facts use basic categories as constituents. Absent an agreement on definition of terms, alternate channels of information will appear to house alternate facts. If proponents of several ideologies are similarly strong—in terms of presenting facts according to their preferred subjectivity—the level of knowledgeability of an AI system is unlikely to be useful. On the other hand, emergence of a single, dominant ideology will drive out diversity, potentially paralyzing social progress.

2.1.4 All salient information is not taken in (OI4)

This error happens when an AI system draws an erroneous inference by missing crucial information present in the input. In eastern China, a traffic camera using AI deemed that a driver was "driving while holding a phone"—an offense meriting a fine—when the corresponding picture clearly shows that the driver was just scratching his face (Allen 2019).

The driver received a notification informing that:

> … he had violated the laws of the road for "driving while holding a phone". A surveillance picture of his "offence" was attached.
> *Source:* Allen (2019)

After he complained though, the city traffic authority relented and canceled the ticket and informed him that:

> "… the traffic surveillance system automatically identifies a driver's motion and then takes a photo", which

is why his face-scratching had been mistaken for him taking a phone call.
Source: Allen (2019)

This case illustrates that the AI system made a mistake in interpreting the posture of the driver as one that attracts a fine, i.e., holding a cell phone. The AI system failed to distinguish it from a similar-looking posture involving scratching of the cheek. The fact that there was no solid object in the hands of the driver was missed out. Moreover, one also notes an illogicality in the AI system's inference: there is no way the driver's moving fingers would access certain locations of the cheek if there was a solid object occupying the space. A better AI system would analyze a video feed and incorporate this check. For this purpose, along the lines of emerging theme (**I**), the AI system would need to be trained with millions of videos depicting human movements, while holding devices or otherwise. The AI system needs to be taught part-whole relationships—i.e. which parts belong to the continuum that is the human and which parts are separate—by taking recourse to reinforcement learning or similar other technique. Thereby the AI indirectly gets taught that two masses cannot simultaneously occupy the same location. Perhaps a good starting point will be to train an AI system on numerous videos of a human accessing various portions of the face—say to address an itch, to rearrange a lock of hair and so forth.

2.1.5 Training data did not contain sufficient number of records for diverse constituents (OI5)

This omission error arises because it is (erroneously) assumed that the training data adequately represents the cases that the AI shall encounter. In reality, the AI design is not capable of handling the diversity in the real world.

An illuminating example can be found in the case of the AI Hiring Tool that Amazon Inc. designed, to spot the best candidates to hire, based on the CVs submitted for software engineering jobs. The AI tool consistently rated the CVs of women candidates lower than CVs from men. Amazon Inc. promptly discontinued using this tool (Oppenheim 2018).

It is interesting to analyze why the problem occurred in the first place. In the software engineering workforce at Amazon, there were disproportionately higher number of men than women. Thus, the sheer number of male candidates getting high performance ratings in their job would be higher, even if Amazon managers never discriminate against women employees.

It is probable that the Amazon AI hiring tool looked at the CVs of currently best-performing employees from the time when they applied for their jobs—overwhelmingly male, owing to there being high numbers of male software engineers—and inferred that candidates who described themselves using military-themed verbs more commonly found on male engineers' resumes, such as "executed" and "captured," (Dastin 2018) are likely to be good hires. This led to the discrimination against women candidates. Perhaps if Amazon Inc. had roughly equal proportions of male and female software engineers, the tool would not have become biased to specific terms appearing in the CVs of male engineers. We reckon it will take some time for machine-learning AI systems to emulate human assessors/interviewers who can readily form a qualitative assessment of a candidates' representation of their attainments to the prospective employers, on the basis of the CV prepared by the candidate.

Note though, one may bring up an entirely different set of concerns with regard to this AI application. An employee performs well when there are good supporting systems—compatible with the employee's preferences—in place in the company. An employee's CV at the time of applying for a job does not have this information. Besides, the job interview dynamics also shapes who gets offered a job and who doesn't. These facts of life lead one to question the wisdom of the endeavor to predict on-job performance from the CV submitted at the time of applying for a job. Therefore, it is not surprising that the Amazon hiring tool recommended resumes of several grossly unsuitable candidates for all manner of technical jobs and had to be withdrawn.

A second instance of lack of diversity in training data is evidenced in the heavier biases against Blacks—in comparison to the extent of biases seen in predictions regarding Whites—in the widely used criminal risk assessment tool, Correctional Offender Management Profiling for Alternative Sanctions, COMPAS[2] (Larson et al. 2016) applied in the context of predicting recidivism.

> Angwin et al. (2016)… analyzed the efficacy of COMPAS on more than 7000 individuals arrested in Broward County, Florida between 2013 and 2014. This analysis indicated that the predictions <regarding recidivism, i.e. a released prisoner (defendant) going back to crime> were unreliable and racially biased. COMPAS's overall accuracy for white defendants is 67.0%, only slightly higher than its accuracy of 63.8% for black defendants. The mistakes made by COMPAS, however, affected black and white defendants differently: Black defendants who did not recidivate were incorrectly predicted to reoffend at a rate of 44.9%, nearly twice as high as their white counterparts at 23.5%; and white defendants who did recidivate were incorrectly predicted to not reoffend at a rate of 47.7%, nearly twice as high as their black counterparts at

[2] The COMPAS tool comes from the company "Northpointe", later rebranded as "Equivant" in January 2017.

28.0%. In other words, COMPAS scores appeared to favor white defendants over black defendants by under-predicting recidivism for white and over-predicting recidivism for black defendants.
Source: Dressel & Farid (2018: 1).

It is likely that the bias against non-whites arises from the fact that, in the correctional systems in the USA, there are far higher number of non-whites than whites. This makes the AI correlate characteristics of non-whites with increased probability of recidivism.

In our view, given that the correctional system in the USA will, in all probability, continue to have significant over-representation of colored people and significant under-representation of white people into the foreseeable future, an AI system having the quality of functionality displayed by the COMPAS should not be used.

The shortcomings of the COMPAS and the Amazon Hiring Tool examples indicate an important decision that needs to be made before deploying an AI System: is its performance unbiased enough to do less harm than good? In other words, do the benefits from true positives and true negatives obtained from an AI system outweigh the deleterious effects of false positives and false negatives? If the answer is no, modifying the decision thresholds in the AI until the right balance is obtained may be tried. If this does not work, we recommend desisting from using AI for the task at hand—until such time a better solution is invented.

2.1.6 Data that was input to the AI during training phase has since become obsolete (OI6)

This can happen when a company deploys its AI system into a new domain, involving new entities and/or new kinds of transactions. For example, the onset of the fin-tech revolution—where one individual can electronically pay another individual using a common mobile phone application—has broadened the scope for financial fraud related to payments. Earlier, the modes of electronic payment at the retail level were limited to internet banking and debit and credit card transactions, restricting the scope for payments-related fraud. Onset of new kind of transactions necessitate that the AI systems for detecting fraud be trained afresh. Also, as fraud patterns change with organized fraudsters identifying and exploiting new vulnerabilities, AI systems for fraud detection need to continuously learn from both newly-reported frauds, as well as newly-discovered patterns of anomaly.

In general, model refresh and periodic training with more recent data should be an underlying best practice for all AI systems. Babic et al. (2021) surface an interesting debate underlying this innocuous-sounding practice: should a company allow the AI-ML algorithm to continuously evolve OR, instead introduce only tested and locked versions at intervals? On one hand, continuously evolving basis for decision-making may not be liked by affected parties. In the worst case, it could be interpreted as malicious behavior hiding behind an excuse that AI-ML is evolving. On the other hand, 'locking' a model does not make the fact go away that AI-ML decisions can be inaccurate, being mere probabilities, and locking does not shield from the effects of environmental changes.

2.2 Commission errors in AI input

A commission error in AI input materializes due to a deficiency in the handling of one or more inputs feeding an AI system. In this case too, we assume that processing logic and repertoire of actions available from the AI system are adequate. To rectify the problem arising from a commission error in AI input it is necessary to modify one or more inputs. In the paragraphs that follow we describe five types of commission errors in AI input.

2.2.1 Hardware deployed for capturing an input is deficient (CI1)

AI can fail because the hardware is incapable of robust performance across environments. The hardware of an AI system comprises not just the CPU, GPU and servers in the cloud, it also involves sensors including image capture devices like camera, devices to capture sound, temperature, humidity etc., and related wiring to the computational device(s) running AI software. The degree of robustness of functioning of the hardware and sensors feeding data to an AI system—under a range of environmental conditions pertaining to fluctuation of temperature, humidity, dust and suspended particulate matter, rain, winds, lighting, foreground and background noise, incidence of electric and magnetic waves—impact AI functioning even when there are no major flaws in the AI software.

The Boeing Max crashes (Lion Air Flight 610 in October 2018 and Ethiopian Airlines Flight 302 in March 2019) involved an AI system (the Maneuvering Characteristics Augmentation System, MCAS) dangerously bending the nose of the airplane downwards regardless of altitude—in a misdirected bid to avoid stall as happens if the plane had its nose raised too far—based on reading from just one (angle-of-attack) sensor (Bryen 2020). Erratic sensor outputs doomed the two fatal 737 Max crashes because it left the pilots fighting the control system and losing the battle, but the problem could have been either the sensor or in the wiring that connects the sensor to the flight computer and thence to the flight software and the MCAS. One viewpoint is that if MCAS was not rushed through development, it could have taken input from other sensors as well, to arrive

at a flight solution, such as from the plane's artificial horizon, airspeed indicator, altitude data and other computer solutions on the flight profile.

2.2.2 Error in reduction in number of data dimensions (CI2)

AI developers are provided a vast number of data features (columns of data that potentially have useful correlation with the outcome variable). However, it is not practical to use all available data features as is, since (a) execution speed drops when the number of data features is high, (b) the system output gets compromised if some inputs are highly correlated (analogous to the multicollinearity issue in statistical regression), and (c) increasing number of features also requires more, ideally uncorrelated, training data samples to avoid over-parameterization biases. All AI implementations therefore have a step for reducing the number of data dimensions. For example, a value calculated from a linear combination of correlated dimensions may be instituted as one dimension. Yet some correlated columns may stay on or materialize either due to oversight or due to a change in the external environment after design and deployment of the AI solution. As a result, AI decision-making gets compromised.

For example, according to Cossins (2018) in the Arabic script "good morning" (الخ ير صد باح) or "good morning to you all" (جم يعا ل كم الخ ير صد باح) and "attack them" (هلجمهم) and "hurt them" (يؤذيهم) appeared very similar to Facebook's automatic translation software.[3] In October 2017, police in Israel arrested a Palestinian worker who had posted a picture of himself on Facebook posing by a bulldozer with the caption that appeared to say "attack them" in Hebrew. Facebook had translated the Arabic text to "attack them" in Hebrew and "hurt them" in English (Smith 2017b). The man was questioned for several hours before someone spotted the mistake that AI had mistranslated "good morning" (בוקר טוב) as "attack them" (לתקוף אותם).

> Arabic speakers explained that English transliteration used by Facebook is not an actual word in Arabic but could look like the verb "to hurt"—even though any Arabic speaker could clearly see the transliteration did not match the translation.
> *Source*: Berger (2017)

If we type "هلجمهم" and request Google to translate from Hebrew to English, we get "Congratulations". However, Google translate Arabic to English for "هلجمهم" returns "attack them".[4] This suggests that the AI system needs to be made aware whether a text is written in Arabic or in Hebrew.

[3] The scripts provided in Arabic and Hebrew were obtained using *Google Translate*, March 02, 2021 and again on 22nd March 2021.

[4] This test was done on 2nd March, 2021. Relatedly "جم يعا ل كم الخ ير صد باح" translates to "I wish you well" when Google-translated Hebrew to English. When Google-translated Arabic to English, we get "good morning to you all".

In our view, a better AI implementation would requisition and emphasize dimensions in the input data that make the sharpest distinction between any pair of letters and phrases, keeping in view the context (Arabic text vs. Hebrew text), to first detect the language of the script, and attempt a translation following that. In other words, if the same pictographic representation [هلجمهم] has wildly different meanings ["Congratulations" vs. "attack them"] when read as Hebrew and Arabic text respectively, additional data columns must be requisitioned to break the tie. Otherwise the onus is on the translation provider AI system to list all possible connotations whenever two (or more) languages having overlapping scripts are potentially involved. The system presenting the AI translation can be designed to be more transparent by exposing more than one possible output along with their estimated likelihoods.

2.2.3 Irrelevant data columns got fed into the neural-networks-based AI solution (CI3)

In this case the AI system makes judgements based on information a human would deem irrelevant. For example, object-recognition algorithms (e.g. facial recognition) get fooled by certain kinds of pictures/designs printed on T-shirts worn by humans (Cole 2019). Nick-named "adversarial designs" these "trick" object detection algorithms into seeing something different from what's there, or not seeing anything at all.

Moreover, researchers from Berkeley, University of Chicago and University of Washington collected 7500 unedited nature photos which confuse the most advanced computer vision algorithms. A fox squirrel standing up on its hind legs got recognized as a sea lion. A dragonfly sitting on a woven cloth got identified as a manhole cover, etc.[5] This kind of issue is more likely to occur when a neural network employs machine learning involving multiple hidden layers of perceptrons (aka deep learning). The patterns recognized by an AI system—based on correlation between spatial features—can accidentally match patterns found in other contexts. Some more examples are: neural networks have labeled sheep in indoor settings as cats; the same neural network detects presence of non-existent sheep in vistas of treeless grassiness, particularly on the mountainsides (Shane 2018).

The present approach is to merely acknowledge that this kind of error is difficult to fix upfront. Upon detection of an unexpected set of matches the training of the AI system is improved by reinforcement learning or semi-supervised

[5] [Source: https://www.immuniweb.com/blog/top-10-failures-of-ai.html].

learning approaches. In our view, it is quite likely that a larger extent of fine-graining of the underlying spatial template (face template, body template etc.) will prevent irrelevant data interfering in AI's judgment. To test this hypothesis, an AI system may be confronted (and trained) with thousands of instances of mimicry listed in books on nature studies, as well as elaborated in books of psychology, leisure reading, theater and cinema, along the lines of emerging theme (**I**).

2.2.4 Irrelevant input not blocked off (CI4)

This error materializes when an AI system appears to take action (not sought by a human user) unilaterally, i.e. without any request from the user. This happens when an AI system responds to what is essentially noise. The noise may originate in the environment, or from other connected systems. Alternately, an update—to the AI system itself, by its makers—containing programming errors (bugs) may trigger action from an AI system when no action is sought.

An instance of an AI system taking action without being asked to is observed in the case where Amazon Echo turned on music from Spotify at full volume, at an hour past midnight, in an empty sixth floor apartment in Germany (Olschewski 2017). Eventually the police were called in. The police broke the lock of the house, disconnected Echo and changed the lock, and charged the resident a goodish sum for the lock-change procedure. Amazon's explanation is that "Echo was remotely activated and the volume increased through the customer's third-party mobile music-streaming app." (Kitching 2017; Smith 2017a). This still does not explain though, what made Echo "activate" remotely in the middle of the night.

Alternately, if we deem the probability of environmental noise to be miniscule in above circumstances, we are led to suspect that an illegitimate input from a connected system (e.g. Spotify) or a buggy update from the makers of Echo itself triggered the uncalled-for action. In any case this highlights another emerging theme (**III**) in the ongoing efforts to improve AI systems: an AI system must get better in distinguishing a signal that is meant to trigger its functioning from all other signals—howsoever generated—that ought not to trigger its functioning. In this instance, a check on the physical sound-catching device (prior to AI taking action)—for example whether the microphone array has indeed received a human-generated external auditory signal or not, a surer way of determining whether a command was registered—can rule out AI-ML getting 'logically' activated from a buggy update.

2.2.5 Feedback loops exacerbating biases (CI5)

In this case, data about a phenomenon is fed into an AI system from a limited context. The resultant AI solution tends to find answers to questions it faces from within the same limited context. For example, in anticipation of drug-related offenses, the PredPol AI system keeps sending police officers to neighborhoods populated with racial minorities, regardless of the true crime rate in those areas (Cossins 2018). Researchers have shown that because the software learns from reports recorded by the police rather than actual crime rates, PredPol creates a "feedback loop" that exacerbates racial biases. In our view an AI system is the wrong tool to use in this context. Alternately, to avoid such feedback loops, two possible approaches can be used: (a) use other data sources (e.g. police informers) to avoid getting caught in system-generated over-deployment feedback and/or (b) use a degree of Reinforcement Learning, by exploratively sometimes sending police officers to areas not yet predicted to be likely crime areas, but areas where sufficient deployments have not been done in the past. In related vein Luca et al. (2016) suggest that certain caps may be put on the number of inspections in poorer neighborhoods, and a similar cap on other neighborhoods bring in a sense of fairness that particular neighborhoods are not being singled out for inspections.

2.3 Omission errors in the processing logic in an AI system

In an AI system, algorithms process the input information in relation to its learning from past information, to 'decide' the action to take. In this section we focus on instances where the processing logic is flawed and where such flaw(s) is/are traceable to omission of certain relevant part of information made available to the AI system or omission of necessary processing steps. We describe four broad sub-types below.

2.3.1 Logic in AI failed to focus on adequately processing a salient part of input information (OP1)

In this case, the input information is sufficient to make the correct inference when viewed by a human. However, the AI system "sees" more, i.e. at a higher level of granularity than humans. This leads to a kind of pattern-matching inference that humans are unlikely to produce. For example, a "classifier" machine-learning system sorts data into different categories. However, it may pick up visual features of the image that are so distorted a human would never recognize them. In the process a random tie-dye pattern or a burst of TV static overlaid on a picture of a Panda gets incorrectly recognized as a gibbon (Vincent 2017). Humans appear to have the ability to make good approximations to the data they receive

from the environment; the approximations are 'good' in the sense that the resultant assessments allow a human to make sense and navigate the surroundings. AI systems are yet to acquire the meta-logic that makes approximations 'good' in particular contexts. It is instructive to compare this with CI3: irrelevant data columns got fed into the neural-networks-based AI solution. There the problem was the other way round—the observation by the AI system appeared superficial, possibly because information-granules under consideration were too coarse, and this led to improper matches. Thus, while humans may be able to 'instinctively' determine the level of detail to focus on (in most cases) we are yet to see AI systems being tooled up with analogous logic.

How do humans decide the level of detail at which to cognize a situation? In our view such is determined by the purpose in the mind of the human being. As an example, the details humans focus on when observing a crowd of people coming out from a railway station are different when one is trying to spot a relative who is expected to arrive and meet up vs. when one is merely watching the crowd flow by the window of a coffee shop. Thus, to address the error condition (logic in AI failed to focus on adequately processing a salient part of input information) an AI system needs to be provided some kind of equivalent of a purpose. This can be accomplished by reinforcement learning to emphasize the purpose and mark its boundaries with regard to the level of detail to consider.

We note that purpose (or end objective) comes to humans naturally, from a range of social, psychological and environmental stimuli. An AI system, on the other hand, does not have any specific end(s) in view, since it does not have a mind. AI developers arm an AI system with the means to perform a task. The reason as to why an AI needs to perform a task is outside the purview of an AI system. When a human deploys an AI system to carry out a task, the purpose for such deployment is still with the human, i.e. outside the AI system. Yet, since purpose often determines the granularity that an AI system needs to pursue, purpose-to-level-of-detail mapping comes across as an emerging theme (**IV**) in AI.

2.3.2 Logic in AI failed to resolve ambiguity in input information (OP2)

This error occurs when the AI system is confronted with an input that can have multiple meanings, and the AI selects an inappropriate interpretation. Shortly after its release in 2011, Apple's Siri agreed to memorize the name of its owner as "an ambulance", simply because the latter had issued a command "Siri, call me an ambulance." (Knight 2016). Siri failed to disambiguate between alternate uses of the phrase "call me". The notoriety of translation by computers is not a recent phenomenon. A sentence in English was translated into Russian by a computer; thereafter the sentence in Russian was translated back to English by the same computer. The final output read as follows: "The Vodka is good but the meat is rotten" (Pollack 1983). The original sentence was "The spirit is willing but the flesh is weak." The computer programs failed to recreate the evocative meaning of the original English sentence on account of non-exposure to certain kinds of learning that humans have access to.

Human speech evolves as new concepts enter popular vocabulary and certain older concepts tend to fade out. Few young people in the year 2022 can describe what a typewriter is, though they may be quite familiar with the keyboard of a computer (or a keypad of a mobile phone); youngsters into music are also likely to be familiar with the keyboard of a piano. In sum, humans subscribe to evolving conversation styles by not only mimicking other humans, but also by relating the changes to physical objects in their social surroundings as well as to evolving ideas regarding abstract entities (e.g. what constitutes good music, appropriate etiquette in a formal dinner in Western Europe vs. in the USA, etc.). An AI system may perhaps obtain access to all the text and speech produced everywhere in the world. However, absent an understanding of the context (in which a conversation occurred) and the associated human mental model connecting tangible and intangible entities, AI will probably be in the catch-up mode, for a long time.

2.3.3 Logic in AI failed to stop execution immediately upon encountering error condition (OP3)

This kind of error materializes when AI developers do not handle an error right at the point of its generation; rather execution is allowed to move ahead ignoring the error condition. Eventually this leads to bad outcomes. The following incident featuring Sophia the humanoid robot is instructive:

> In March of 2016, Sophia's creator, David Hanson of Hanson Robotics, asked Sophia during a live demonstration at the SXSW festival, "Do you want to destroy humans? ...Please say 'no.'" With a blank expression, Sophia responded, "OK. I will destroy humans."
> *Source:* Weller (2017)

Sophia misinterpreted a request to provide her opinion about something as a request to agree to carry out the activity in question. Sophia interpreted the "Please" after Hanson's question as a request to agree to do as "asked" in the previous sentence (i.e. agree to destroy humans), and failed to comprehend that the full sentence—"Please say 'no'"—intended just the opposite. We note that Sophia did not directly answer whether she wants to destroy humans. Sophia's response ("OK. I will …") suggests she merely agreed to carry out a command given to her. If Sophia had a more alert algorithm, she would have responded that she does not possess adequate training to judge whether wanting

to destroy humans is a good thing or a bad thing; and further, she would like to engage only in doing good things, etc.

2.3.4 Inadequate (insufficient) processing by an AI system (OP4)

In this case, the processing carried out by an AI system is insufficient, because action from the AI system is expected within a limited timeframe. Upon reaching the end of the allotted time for processing a stimulus the AI system has two choices (a) desist from taking action (b) continue with prior action (or, in some cases take a random action) without fully processing the stimulus. In case of (a) it is necessary that a human intervenes and executes the correct action. In case of (b) an accident or goof up is highly likely; alternately, the action by the AI may turn out to be harmless purely by chance. Driverless vehicle operation provide illustrations of this situation.

Rides by driverless vehicles tend to be jerkier when the AI system pays attention to a greater number of "threats" and slams brakes to wait till a threat goes away; a smoother ride is possible only when a majority of threats are ignored by the AI system, but it may lead to serious accidents. The latter situation materialized when an AI-driven Uber vehicle hit and killed a pedestrian crossing a street at a point where there was no crosswalk sign (McCausland 2019). Uber's AI system—Automated Driving System (ADS)—detected the pedestrian 5.6 s before the incident, when it classified the pedestrian as a vehicle. Subsequently, when the Uber vehicle got closer to the pedestrian, she was classified as a bicyclist. The ADS was able to track the pedestrian continuously until the crash. However, the vehicle did not slow down and stop for the pedestrian (Krisher 2018). To activate the emergency braking system, the ADS needed to predict the collision at least 1.2 s before impact. The prediction failed to materialize (Levin 2018).

Stewart (2018) reports an unusual number of cases of self-driving cars being hit from the rear by other vehicles (and a bicycle!). This suggests that the AI-driven car is prone to apply brake or stop upon encountering "unexpected" things in the neighborhood of its path, things human drivers are unlikely to deem worthy of braking and stopping. This suggests that that the driverless vehicles drive in ways humans might not expect, and might not want them to. For the moment this is condoned on the reasoning that it is better to have a driverless vehicle stop upon spotting a fire-hydrant on the roadside, rather than run over a preschooler about to cross the road. The following quote summarizes the situation well.

> … self-driving car technologies have to make a trade-off: either you can have a car that rides slow and jerky as it slows down or slams on the brakes to avoid objects that aren't a real threat, or you have a smoother ride that runs the risk of having the software dismiss objects, potentially leading to the catastrophic decision that pedestrians aren't actual objects.
> *Source*: Vaas (2018)

2.4 Commission errors in the processing logic in an AI system

This is the situation where the AI system has been provided all the input information it requires and yet it takes an erroneous action. This happens due to a flaw in the processing logic and the flaw can be fixed only by modifying the processing logic. We assume that input information to the AI is adequate and that the repertoire of actions available to the AI system is adequate as well. We describe four flavors of commission errors in the processing logic of an AI system in the paragraphs that follow.

2.4.1 Logic in AI focused on non-salient part of input information (CP1)

This error happens when the AI system is provided all the information it needs and yet it makes a judgment based on the less-salient part of the information it receives. In the example below the reason for misjudgment was presence of a non-salient pattern resembling the pattern of interest; moreover, certain ambient conditions exacerbated the frequency and intensity of the error.

> In October, the Scottish Inverness Caledonian Thistle FC soccer club announced its home games would feature live video coverage courtesy of a newly installed AI-powered Pixellot camera system. Alas, in its attempts to follow the flow of the game at Caledonian Stadium, the AI ball-tracking technology repeatedly confused the ball with the referee's bald head, especially when its view was obscured by players or shadows.
> *Source:* Cai and Yuan (2021)

Above case can be considered another instance of the fundamental theme (**III**) for making AI systems better—an AI system must get better in distinguishing a signal that is meant to trigger its functioning from all other signals—howsoever generated—that ought not to trigger its functioning. This entails that an improved AI system will then call up much more data and many more models (concurrently) to discern whether what it 'heard' is a trigger command or not. The underlying technology/processing is quite different from that underlying an AI system carrying out 'normal' conversation.[6]

[6] Here we assume that the technology for discerning a trigger command improves over time—by employing AI-ML-based learning—so that physiological changes to the human owner issuing the com-

2.4.2 An inappropriate logic was used (CP2)

In this case the processing logic employed by the AI system is not fit for purpose. The logic may work satisfactorily for a subset of cases under consideration. However, there are fundamental issues whereby the processing logic is inappropriate.

For example, it has been shown that in Amazon's Rekognition tool, photos of senators of color are more likely to be misidentified as matching with mugshots of persons arrested on suspicion of criminal conduct (Snow 2018). It is probable that the 'face template' used in the Rekognition tool has a large number of parameters on which only non-colored people (specifically, the white males) exhibit significant variation. This allows fine-grained judging of images of white males. It is also likely that a Hispanic or African-American face in a photograph shows limited extent of variation on a majority of the same parameters in the face template used in the Rekognition tool. This can happen if a shadow of a facial feature is indistinguishable owing to darker skin tone. The AI system's commitment to parameters in which the colored people exhibit lesser variation leads to course-grained judgments, resulting in higher extent of false positives. A speculation regarding this technical reason can be found in the "Comments" section under the article by Cushing (2019):

> The real reason <for higher false positives for people of color> is … lighting. Faces are curved and contoured enough to cast shadows, and people with lighter skin have more contrast between lit and shadowed parts of their face than people with darker skin, because their lit skin color is simply closer to the color of shadow. This makes distinctive features blur together to the computer's pattern-matching system. It's unfortunate that that happens …
> Comment by a reader in Cushing (2019)

A Federal study by the National Institute of Standards and Technology (NIST) in the US examined 189 facial recognition algorithms voluntarily submitted by 99 companies, academic institutions and other developers with 18 million photos of more than 8 million people sourced from databases run by the US State Department, the Department of Homeland Security and the FBI (Harwell 2019). Noted companies like Idemia, Intel, Microsoft, Panasonic, SenseTime and Vigilant Solutions took part. Amazon declined to participate. The Federal study found:

> Asian and African-American people were up to 100 times more likely to be misidentified than white men, depending on the particular algorithm and type of search. Native Americans had the highest false-positive rate of all ethnicities, according to the study, which found that systems varied widely in their accuracy. The faces of African-American women were falsely identified more often in the kinds of searches used by police investigators where an image is compared to thousands or millions of others in hopes of identifying a suspect. *Source:* Harwell (2019)

However, it is heartening to note that:

> Algorithms developed in Asian countries had smaller differences in error rates between white and Asian faces, suggesting a relationship "between an algorithm's performance and the data used to train it" … "You need to know your algorithm, know your data and know your use case," … "Because that matters." – *Source:* Members involved the Federal study quoted in Harwell (2019).

The fact that algorithms developed outside the US do a better job—i.e. have smaller difference in error rates between white and Asian faces—suggests that lowering racial bias in facial recognition technology shall require an approach very different from that of the AI developers in the USA (particularly Amazon's "Rekognition" tool). However, if AI based on (current) machine-learning technology continues to be uncomfortable with simultaneous pursuit of multiple goals (e.g. do justice to white subjects, do justice to subjects of color, do justice to people of different genders and color) it may be necessary to look for technologies that augment machine learning, for example, in better camera sensor settings and image pre-processing that enhances features of interest across the entire possible range of skin tones. Alternately it may be necessary to have multiple AI algorithms to carry out facial recognition work, each in its respective domain of competence. An algorithm doing better in recognizing faces of White people is not used for recognizing faces of African-American people. An algorithm doing poorly in recognizing faces of African-American women is not deployed for that task, and so on.

In this context, the role of old-fashioned, painstaking detective work in piecing together evidence before making a radical move like an arrest cannot be understated. In Michigan, the police arrested an innocent African-American man (Mr. Williams)—handcuffing him in front of his wife and children—and held him overnight in a detention center (Hill 2020). The reason: a facial recognition algorithm had matched this man with the photographs of a suspected shoplifter in an upscale store. However, a human being looking at the photograph and Mr. Williams can make out that there

Footnote 6 (continued)

mand as well as the gradual changes in the ambient conditions that an AI-ML system operates in, gets baked into AI's processing.

is no match. The unfortunate event could have been avoided if police detectives did some sleuthing upfront, by going around with the picture and checking whether it matches Mr. Williams and/or (indirectly) asking people knowing Mr. Williams whether they think it is him in the photo, etc.

2.4.3 Ill-conceived updates or features from the originators of the AI system break functioning (CP3)

AI functioning may break upon operation of certain features in an unanticipated way. On occasions, the present-day practice of continuously providing software updates to AI systems breaks the fidelity of a system or a part of the system. In the example that follows, Google Photos and Google Assistant teamed up to create a bizarre/absurd panorama.

> Alex Harker was skiing with friends at the Lake Louise ski resort in Banff, Alberta, a week ago when the group stopped to take some photos on Harker's Android smartphone. After shooting a few shots, Harker found that the AI-powered panorama stitching feature inside his Google Photos app had created ... <an absurd> photo … as the suggested panorama for his scene.
> For some reason, Google Photos saw fit to insert Harker's friend Matt as a colossal bust in the snowy mountain landscape, making the guy look like a colossus peering over the hill at Harker.
> …
> "I literally took like 3 pictures, one with them in, and two without them," he says. "And for some bizarre reason Google Assistant offered me a really strange panorama of the 3 photos spliced together."
> *Source:* Zhang (2018)

The AI system flaw noted above suggests that, when creating new features, AI developers themselves may be unaware of how well their new creation sits with features already present. A flaw may materialize either due to inadequacy in the paradigm espoused for prior features; alternately, the paradigm espoused for new features is at loggerheads with the prior paradigm. If this problem can create so much nuisance when only AI systems from one organization (Google in this case) are involved, one foresees grim outcomes when mission-critical decision-making is obtained from multiple interacting AI systems.

In our view above constitutes another instance concerning the emerging theme (**IV**) of mapping of purpose (of coming into being of an AI system) with the capability the AI system is armed with. If the purpose of the panorama feature is aligned with retaining veridicality with reality (i.e. a human must not be represented out of proportion to his/her size relative to the surroundings) extensive testing is necessary. To construct the test cases, millions of images (with sub-sets having similarity in geographical settings where the images were clicked) may be abstracted from the Internet or for making panorama merging a few images at a time. A new AI system shall then identify instances where veridicality is being violated. This can lead to remediation of the quirks of Google Photos.

2.4.4 Logic necessary for some part of the AI functioning got deployed in another part where it is not required (CP4)

The black-box nature of the logic embodied in AI algorithms entails that AI developers may not always be in control of what kind of logic works where. A snafu in Google Home Mini provides an illustration. The AI system has a legitimate functionality to listen to a householder's conversation when permission is provided. There is also a case for sending some of the conversation to Google, particularly in cases where quality issues are detected—similar to sending error reports to a software manufacturer. However, it is problematic if an AI system secretly records conversations and relays it back to Google HQ.

> In October, security researchers discovered that some Google Home Minis had been secretly turning on, recording thousands of minutes of audio of their owners, and sending the tapes to Google. After noticing that his digital assistant had been turning on and trying to listen to the TV, one user checked Google's My Activity portal, where he found out the device had been recording him.
> Google quickly announced a patch to prevent the issue.
> *Source:* Krauth (2018)

In the case above, we see another distinctive instance of AI developers not being in control of theaters of deployment of AI's decision-making logic. Alternately, it is probable that a 'backdoor' as above was created as a convenience for system testing i.e., to test the system rigorously. Subsequently, the developers forgot to shut off the backdoor (before the development moved to production). This suggests that negatives of edge test cases need to be incorporated in the test schedules: a backdoor test must return a negative before a development moves to production. The Google Home Mini trying to listen to the TV also manifests a kind of concern we noted earlier: inadequate distinction of trigger commands.

2.5 Omission errors in outputs from an AI system

This kind of error arises from failing to arm an AI system with a mechanism to carry out an intended task to an adequate level of dexterity. We assume that the right inputs are provided to the AI system, the AI system comprehends and processes the inputs correctly, and there is no major flaw in the processing logic. This omission error can be addressed

if certain new mechanism(s) is/are made available to the AI system. We list four flavors of this error below.

2.5.1 Failure arising from lack of technological sophistication (OO1)

In the field of AI-powered robotics, machines have failed to display grasping and picking skills matching that of a human: combing through bin after bin of variegated articles found in a retailer's warehouse—clothes, shoes, electronic equipment, detergents, glasswork, hammers, nails, milk-packets, comic books, shampoos etc.—so that each item can be packaged and sent on its way (Satariano and Metz, 2020).

AI cannot smell, even today: a robotic janitor cannot make a room just correctly good smelling the way a human can, by trial and error and experience.

2.5.2 Failure to design new kinds of actions from AI, when environment changes (OO2)

Physical settings in human society change in response to changing needs. For example, as a result of technological advancement, new equipment gets deployed. Thereafter, rules and norms are modified to smoothen the path of deployment. To adjust, human behavior changes. New forms of human behavior necessitate newer configurations of material artifacts, including AI Systems! On some occasions this involves adding new actions to the repertoire of an AI system. The incident described below is informative.

> Back in November 2018, Chinese police admitted to wrongly shaming a billionaire businesswoman after a facial recognition system designed to catch jaywalkers 'caught' her on an advert on a passing bus.
> Traffic police in major Chinese cities deploy smart cameras that use facial recognition techniques to detect jaywalkers, whose names and faces then show up on a public display screen. After this went viral on Chinese social media, a CloudWalk researcher stated that the algorithm's lack of live detection could have been the problem.
> *Source:* Thomas (2020)

In above instance we observe that liveness detection became a requirement after it became known that images of humans on plying vehicles can be a source of human images detected on roads at times when humans are not allowed to be on the road (i.e., a stretch of the road surface is reserved for vehicular movement).

2.5.3 Failing to send output to intended party/parties, in a timely manner (OO3)

In this case the AI system initiates an appropriate action but the action fails to complete owing to some kind of outage in the outward transmission chain. This problem is not specific to AI—this issue has shown up in prior attempts to automate as well. A majority of email users have experienced an outgoing email getting stuck in the 'outbox', i.e., failing to make its way out to the intended party in a timely manner. Alternately a print job that got stuck and was canceled by the user may get executed several hours or several days later (when it is no longer necessary), possibly recalled from quirky parts of the (printer or computer) system's memory. In case this kind of problem originates from an AI system, there are added complications, given that no human is involved. An AI's failure to communicate (when it ought to have sent a communication) may be misinterpreted as that the AI decided to withhold communicating for a reason. For example, a failure to send out an expected payment (or to respond to a demand/penalty notice from a local authority) signals intransigence. A powerful ignored party—who is unwilling to consider contingencies like mistake by a machine—may initiate harsh countermeasures. Likewise, an AI system sending out a communication unusually late may needlessly mobilize people (or other AI/automation systems) well after the need for such mobilization has dissipated.

In the enterprise application integration (EAI) space for machine to machine communication (say between an ERP system and a CRM system), a technique of certified messaging is used: each input task-message is acknowledged, and acknowledgements are anticipated for each task-message sent out. If an expected acknowledgment does not arrive timely, the expecting system (i.e. the system making the original outward communication) requests for the acknowledgment again. In response, occasionally the sender gets informed that the receiver did not receive the original task-message. The latter is then resent, if still relevant. This may be difficult to implement when an AI system's outward transmission is received by a human. A human is unlikely to be comfortable with a paranoid way of functioning—demanding or anticipating acknowledgment for every action, both ways. Rather, as human-to-human communications develope, routine matter is rarely cross-acknowledged; additional energy is devoted for tracking and controlling only for exception cases. In sum we are of the opinion that an AI system should not be deployed when there is a human on one side of a transaction and the situation needs cross-acknowledgment of each task or action rigorously, as in certified messaging.

2.5.4 Some relevant ways of deciding on an action (for example, consulting a human, asking another AI system) not made available to the AI system (OO4)

This error manifests when there is (I) inadequate appreciation as to the boundary where an AI system's responsibility ends and the responsibility of the human partner begins, as well as (II) lack of clarity regarding (a) the rules-of-engagement between a human and the AI system and (b) tie-break recourses in case of deadlock in interaction.

In many AI implementations, the prevalent approach has been to incorporate an AI system into business operations and hope/expect that the surroundings (including humans) will adjust to the AI system's quirks. This is wrong. Careful thought needs to be provided regarding specific conditions that should activate AI functioning, rules-of-engagement between human and AI, and specific conditions that signify a transfer of execution responsibility from AI to a human. We provide an illustration below.

> In 2015, in Japan the first innovative Henn-na Hotel opened its doors to guests. All its staff: the front-desk, cleaners, porters and in-room assistants were robots. But the bots started accumulating customer complaints much faster than expected: the bots frequently broke down, could not provide satisfactory answers to guest queries, and in-room assistants startled guests at night by interpreting snoring as a wake command.
> *Source:* https://www.immuniweb.com/blog/top-10-failures-of-ai.html

In the example above, the failure of AI is at multiple levels. However, what stands out is that the AI did not have a confidant or colleague or boss to refer to, when faced with a decision situation slightly out of the ordinary (i.e. when facing a situation for which exact pre-coding of action is not available). There is a definite purpose why an organization—a contrivance for coordination—has multiple levels: certain decisions need to be taken at a level higher than the level at which a problem is detected. Such is necessary because additional information needs to be brought into consideration, to progress towards a suitable/satisfactory resolution. For example, if a guest (justifiably) seeks a change of room, she may be moved to another room in the same hotel depending on other reservations, the schedule of rooms falling vacant, her duration of stay and the lead time the hotel requires to correct the malfunction in her present room. Otherwise the guest may be offered some compensation and/or a transfer to another comparable hotel, with free drop-off. Likewise, if a guest seeks fresh towels, such may be provided only if inventory exists; otherwise it may make sense to provide some compensation to the guest, say a discount voucher.

In Henn-na Hotel, tasks appear to have been divided among Robots holding specific roles. However, one fails to find elaboration of mechanisms for resolution of matters of needing interaction between robot-roles and for resolution of matters needing higher level information outside the immediate problem. There is need for certain additional ways of deciding on an action—for example, consulting a human, asking another AI system, etc.

2.6 Commission errors in outputs from an AI system

Commission errors in outputs of an AI system stem from adverse interaction (a) within the parts of an AI system—say as a consequence of poorly-handled logical inconsistencies among the modules of an system, (b) between an AI system and humans, (c) between an AI system and the environment, (d) between an AI system and other connected technological systems, OR (e) because AI learned bad behavior in the wild, i.e. through a combination of human agency, environmental context and interaction with other technology systems.

2.6.1 Faulty action stemming from adverse interaction between modules of an AI system (CO1)

In this variant, an AI system sends information/commands to parties unnecessarily. Stated differently, an expected action by the AI system fails to materialize, timely, for the correct target audience. One example is a failure to limit sending output to only legitimate parties. This manifests when the necessity to narrow-cast is overlooked, and broadcasting is inappropriately resorted to, instead. In Portland, USA, Amazon Echo listened to a couple's conversation about hardwood floors and sent the recording to someone in their contact list—without the couple's knowledge (Wamsley 2018).

> The wife … told … that they learned something was amiss when they received a phone call from the husband's employee who lived in Seattle, telling them what he had inadvertently received. He told them to unplug their Alexa devices right away. … The employee sent the couple the sound file that the Echo had sent to him, and they were shocked to realize they had essentially been bugged.
> *Source:* Wamsley (2018).
>
> Amazon explains:
>
> "Echo woke up due to a word in background conversation sounding like 'Alexa.' Then, the subsequent conversation was heard as a 'send message' request. At which point, Alexa said out loud 'To whom?' At which point, the background conversation was interpreted as a name in the customers contact list. Alexa then asked out loud, '[contact name], right?' Alexa then interpreted background conversation as 'right'. As unlikely

as this string of events is, we are evaluating options to make this case even less likely."
Source: Wamsley (2018).

The fact that Alexa heard background conversation and misconstrued it as very specific directives several times suggests that Amazon's explanation is not robust. Alexa's principal failing is its inability to distinguish between conversion relevant to it and other conversation it is not required to involve itself with (in line with an emerging theme (**III**) noted earlier). Moreover, Alexa appears to be thinking up complex actions (like sending a conversation to a contact) on its own. Amazon's explanation suggests that sending a message is a standard option in Alexa's toolkit of features. Alexa may even be configured (by the manufacturers) to regularly send conversation samples back to HQ for quality purposes, without the user explicitly ordering it to do so. Sending the hardwood floor conversation to a contact of the user—an employee with whom the user may have had other conversations regarding "deliveries" in their business—upon coming across a domestic conversation that may have also concerned deliveries (regarding supplies for hardwood floors) leads us to speculate that this is happening because there are several un-handled inconsistencies between Alexa's own modules.

2.6.2 Inappropriate action stemming from adverse AI-environment interaction (CO2)

In this variant the AI system clearly malfunctions on some occasions but not all; moreover, reasons for malfunction lie in some undiagnosed set of environmental conditions interacting with the way an AI system is set up to function. Malfunctioning robots provide good illustrations, as in the two cases below.

A video that went viral on Chinese social media platform Weibo shows a robot tumbling down an escalator, crashing into and knocking over shoppers. The incident occurred on Christmas Day in China's Fuzhou Zhongfang Wanbaocheng Mall.
Convenient, cost-efficient and cute, service robots have been widely deployed in public places—but some are adapting better than others to life in the wild. This particular robot's tasks included providing information services, body temperature monitoring of shoppers, and using interactive functions such as singing and dancing to entertain children. While there are mixed reports on whether the robot may have been interfered with, a supervisor at the mall reported that it navigated to the escalator by itself.
Source: Cai and Yuan (2021)

Failures of this kind are cause for concern because something that is cute and non-threatening suddenly assumes the proportions of a deadly, destructive entity. We can speculate on some reasons for malfunction. For one, some machine part may have malfunctioned owing to change in temperature or humidity or due to stress from extended operation. The robot may have lacked routines to periodically verify that all its parts function as intended. Hence one fault led to other and so forth, creating a domino effect. Alternately even a little accidental push (by humans or other robots etc. in the surroundings) may cause a robot's motors/brakes to jam; subsequent operation in the faulty condition again create a chain of faults and culminate in a major incident. As an analogy, let us consider a human moving around in a mall getting injured in one leg—say upon stumbling on to something hard. Subsequently when the human tries to climb (or go down) stairs, sensations of pain convey a message that the normal way of placing footsteps is jeopardized. If the human ignores this message and tries to navigate the stairs in the usual way, a fall is very much a possibility. In reality the injured human takes extra precautions—like holding on to handrails, taking smaller paces or stepping slowly etc.—to navigate the steps. An "injured" robot may not have relevant sensations of pain built in (per today's technology). Hence it fails trying to do things the usual way when its "injury" requires it to do otherwise. Below we present a second instance.

SoftBank-owned Boston Dynamics debuted its humanoid robot Atlas at Congress of Future Science and Technology Leaders in 2017. While it displayed impressive dexterity on the stage, it tripped over the curtain and tumbled off the stage just as it was wrapping up.
Source: Thomas (2020)

If Atlas were a human, we would have rationalized that (s)he got tired and fatigued after the stage performance, or the audience's positive reactions distracted her/him, and hence (s)he tripped on the way out. For a robot though, we are led to speculate that, similar to the previous case, Atlas's antics on the stage led to some degradation/malfunction of components and such malfunction was not sensed and acted upon by changing Atlas's behavior and task expectations. This raises an interesting issue. If a robot is simply an automaton working according to fixed rules, it is possible to recommend rigid guidelines for behavior and functioning. However, if the robot also learns from experience by machine learning, deeper research is needed to understand what the robot learns when conditions very divergent from those of normal use materialize. This is another emerging theme (**V**) for further inquiry.

For example, when training a robot for deployment in a Mall, it will be necessary to create conditions of significant turbulence. The robot should be pushed, shoved, kicked, hit (by objects and by humans and other robots), ambient lights

should go on/off at random, floors should made slippery and/or strewn with obstacles, temperature should abruptly change, humidity should fluctuate, there should be sudden gusts of draught and/or rain, the robots battery/energy storage should be made to fail abruptly, and so on. In every case that a robot detects an internal failure it cannot recover from without human support, the robot should immobilize i.e., retire and wait at the nearest place deemed safe for a withdrawing robot. In this situation there is a possibility that robots withdraw rather too frequently in live operation, lowering their effectiveness. Thus, robots cannot be used in contexts where it is too expensive to train robots for safe operation.

2.6.3 Questionable action stemming from adverse interaction between AI and other technology systems (CO3)

This commission error materializes when AI is deployed in a setting involving multiple technologies and AI's action demonstrates bias of some kind. For example, upon uploading of large photo collage containing white and African-American persons Twitter was seen to selectively crop out the face of an African-American person in its image preview (Das 2020).

> The image preview function of Twitter's mobile app automatically crops pictures that are too big to fit on the screen and selects which parts of the image to display and cut off.
> Prompted by a graduate student who found an image he was posting cropped out the face of a black colleague, a San Francisco-based programmer <Tony Arcieri> found Twitter's system would crop out images of President Barack Obama <a black person> when posted alongside Republican Senate Leader Mitch McConnell <a white person>.
> *Source*: Asher-Schapiro (2020)
> Arcieri uploaded large photo collages of former US President Barack Obama and Republican Senate Leader Mitch McConnell, with their faces placed in different spots in the various versions. Twitter's image preview function automatically crops photos which are too big for the screen, selecting which part to make visible to users. The idea was to force the algorithm to choose one of the men's faces to feature in the tweet's image preview.
> But for every iteration, Twitter's algorithm cropped out Obama's face, instead focusing on McConnell, a white politician. Arcieri tried changing other parts of the image, including the color of the ties the men were wearing, but nothing worked in Obama's favor. It was only when Arcieri inverted the picture's colors that Obama was finally featured.
> *Source*: Restle (2020)

In another instance, an online meeting tool used to deliver lectures to students remotely, Zoom, cropped out the head of a black faculty member when used with a virtual background, and did not do this for a white person Das (2020).

> … Ph.D. student Colin Madland tweeted about a Black faculty member's issues with Zoom. According to Madland, whenever said faculty member would use a virtual background, Zoom would remove his head.
> *Source:* Dickey (2020)

In this case the interaction of the technology for online meeting image and video capture with the technology for setting backgrounds produced an undesirable result. The feature for setting a cheerful background is very useful for meeting attendees who have to take the meeting from cramped or dreary surroundings (say due to COVID-19 lockdowns). It comes as a shock that, for some people, utilizing this feature will come with a price tag of having to appear headless. This error will, in all probability, get corrected expeditiously. However, the serious extent of failure raises questions regarding the ability of AI's in supporting human endeavors in difficult situations.

2.6.4 Inappropriate action stemming from adverse interaction between AI and humans (CO4)

This error manifests when an AI system's ability of learning from interaction with humans gets misused to make it do bad things. Microsoft's travails with successive chatbots illustrate this situation.

> Tay, the millennial chatbot created by Microsoft, started spewing bigoted and white supremacist comments within hours of its release. … The Internet soon discovered you could get Tay to repeat phrases back to you ... The bot was taught everything from repeating hateful gamergate mantras to referring to the president with an offensive racial slur. … <Microsoft provided a comment that> " … Unfortunately, within the first 24 hours of coming online, we became aware of a coordinated effort by some users to abuse Tay's commenting skills to have Tay respond in inappropriate ways. As a result, we have taken Tay offline and are making adjustments."
> *Source*: Paul (2016)
> The company tried a second time with a bot named Zo, and the company said it implemented safeguards to prevent bad behavior. However, that bot picked up bad habits, too.
> *Source*: Larson (2017)

> Zo told a BuzzFeed News reporter the Qur'an (the holy book of Islam) is "very violent." (although) Microsoft programmed Zo to avoid discussing politics and religion.
> *Source*: Kantrowitz (2017).

Facebook's AI-driven self-service platform (for purchasing ads) fared worse.

> Using Facebook's AI-driven, self-service platform to purchase ads, companies and brands can target their message to different demographics. In September, ProPublica reported that some of those demographics include those with racist or anti-Semitic views.
> The news organization found that ads could be specifically targeted to people interested in topics like "How to burn <members of a particular ethnic minority>" or "History of 'why <members of a particular ethnic minority> ruin the world.'" Facebook said those categories were created by an algorithm, not a human, and removed them as an option.
> *Source*: Krauth (2018)

Above examples suggests that AI poses danger to humans not just in terms of transmitting toxic ideas from bad people; AI can discover and amplify such malaise as well.

In our view, there needs to be some level of human quality control to supervise the task of creating conceptual categories. Humans are subject to normative pressures—to conform—from society, from colleagues and family members. They also have access to other people's experiences and academic content regarding what works and what doesn't. Therefore, it is reasonable to expect that humans will be good at identifying normative categories that embarrass or worse. An AI system, on the other hand, is subject to nil normative pressure, has no collegiality to uphold, and cannot feel shame. For this reason, when it creates conceptual categories, a level of damage to social norms needs to anticipated and headed off appropriately by human intervention.

2.6.5 AI learns biased/inappropriate behavior in the wild (CO5)

In this variant, a commission error materializes when an AI system takes an improper action on account of having learnt questionable inferences in its interactions with humans, content in websites and repositories, and other AI and technological systems in variegated contexts. For example, Google Allo suggested a man in turban emoji as response to a gun emoji (Krauth 2018). This is objectionable because Google Allo appears to be suggesting that any person who wears a turban is likely to be connected with violence, as signified by the stimulus—a gun. Google apologized and changed Allo's algorithm after CNN reported the instance to Google (Krauth 2018). Let us look at another example.

Google's instant Autocomplete feature for search text ran into trouble several times for suggesting negative content associated with specific individuals or groups. We list two instances.

> Google got sued in France because its autocomplete feature suggests the word "Jewish" in searches involving certain public figures, including News Corporation chairman Rupert Murdoch and actor Jon Hamm, reports The Times of Israel.
> Indeed, querying the search engine for "Jon Hamm," for example, yields "Jon Hamm Jewish" as one of the top results.
> According to Google's website, its algorithm for the Google Instant autocomplete feature "predicts and displays search queries based on other users' search activities and the contents of web pages indexed by Google." In addition, the search engine says it strives to "reflect the diversity of content on the web (some good, some objectionable)" and so has a narrow set of removal policies for pornography, violence, hate speech, etc.
> *Source*: Palis (2017).

In above case the algorithm appears to be needlessly bringing in the religious affiliation of a person when the search context—typed in by the user till the point Google Instant Autocomplete interferes—suggests nothing more than a general interest about a public figure. This is wrong because there is a suggestion of generalization of the deeds and characteristics of individuals to an entire community of people. It is not unlikely that Google Instant Autocomplete learnt this behavior from prior searches along the lines of "Is Jon Hamm Jewish". People wish to find out a public figure's religious affiliation for a range of reasons, some good, and some bad. As an example of the first kind, one may wish to send good wishes to a specific person on the occasion of a particular religious festival, and therefore checks affiliation. Alternately, a biased person may be looking to validate his/her misgivings about negative characteristics of a community by inquiring whether a public figure—who, in the judgment of the enquirer also showed similar negative characteristics—belongs to the community in question. The Google Instant Autocomplete technology appears to be deploying an availability bias (Tversky and Kahneman 1974), since it already has the probable answers to "Is Jon Hamm Jewish" rated and ranked.

To reiterate, there are two key issues here. (I) Google Instant Autocomplete does not know the reasons why each prior enquirer asked "Is Jon Hamm Jewish" and does not care (possibly owing to technology or privacy policy limitations). In other words, Google Instant Autocomplete does not

know if the query was for good or bad reasons (as illustrated above). Distinguishing Good from Bad in specific contexts requires grasp of ethics and morality of a kind not possessed by an AI system. Moreover, Google Instant Autocomplete also does not know the reasons why the present enquirer has typed "Jon Hamm" in the search box. This is a basic flaw hard-wired into all machine-learning-based AI systems: that, data can inform what happened, but data cannot tell why something happened. Yet the current AI technology, by and large, fails to confront this issue. (II) A second problem is revealed when the Google Instant Autocomplete goes on to unilaterally assume that the search query is inquiring about the religious affiliation of the public figure. The problem pertains to an AI system's proclivity to seek (and stick) one (or a limited number of) answer(s) to every (potential) question: recall the most likely reason Google Instant Autocomplete suggested "Jon Hamm Jewish" is that it has the ranked and rated (indexed) answers ready. Thus, a human flaw—avoiding being open to new information by preferring to parse information by a well-worn template—appears to plague AI systems as well.

Typing "Jon Ham" (note the single "m") in Google search box on March 24, 2021 brings up net worth, height, age, wife, movies, girlfriend, dating, batman and black mirror through Google Instant Autocomplete, alongside the text "Jon Hamm". This appears to be a list of topics related to Jon Hamm that members of public have curiosity on. At present none of these topics are proscribed. If any become proscribed at a later point in time, the AI system will be on the dock again. For example, by putting net worth at the top followed by terms indicating search for romantic connection with Mr. Hamm, Google Instant Autocomplete appears misogynist since it is suggesting that prospective mates to Mr. Hamm (who was not married at the time of this experiment) are gold-diggers, given that information about his net worth is their top priority. One also observes that by providing a specific set of additional keywords to choose from, Google Instant Autocomplete ends up (unwittingly) constricting human curiosity, and creates vicious recycling of tired banality.

A second instance of a failing by Google Instant Autocomplete sharply illuminates the malaises discussed above.

> Just over a month ago, a man in Japan won an injunction against Google to have the autocomplete feature turned off when someone searched the man's name. Apparently, the search engine was connecting the man's name with crimes he had not committed and, according to Japan Times, "likely played a role in the sudden loss of his job several years ago and caused several companies to subsequently reject him when he applied for new jobs."
> *Source*: Palis (2017).

We observe classic instances of unintended consequences of purposive social/technological action (Merton 1936) from the Google Instant Autocomplete fiasco(s). Emphasis on quickly serving queries ends up harming humans through fashioning of untrue connection(s) with negative events, leading to adverse consequences in real life.

In our view Google Instant Autocomplete may hold off from offering the autocomplete function from time to time. Rather, the underlying AI system should wait to allow the user to type in the full query. This will allow Google Instant Autocomplete to learn about new kinds of query on an existing topic. It can then go ahead and build indexes to keep answers ready for these new queries, in the background. Thereby, Google Instant Autocomplete can continue its focus on serving relevant content, fast, without coming across as overbearing in attempting to dictate what users should look for. Moreover, Google Instant Autocomplete may incorporate a filter for certain autocomplete suggestions that may be deemed inappropriate or insensitive. In cases when the filter determines an autocomplete suggestion can be deemed inappropriate beyond a certain level of certitude it can skip that suggestion and move to the next candidate in the autocomplete suggestions list. To further make the construction of such a filter feasible, Google can additionally allow users to provide feedback of a particular suggestion being inappropriate (similar to how ads can be marked inappropriate or irrelevant with a "x" option next to them)—thus allowing for crowd-sourcing the construction of the filter.

3 Towards more robust AI-ML systems

In this section we consolidate the discussion on ways and means to make AI-ML systems more robust. First, we note a set of key recommendations from extant research. Thereafter we present a consolidation of the additional lines of inquiry our study findings bring to the fore.

3.1 Views from extant research

To lower the probability of failure of AI-ML systems, Yampolskiy (2019) recommends (a) restricting/controlling user input to an AI system (b) analyzing how many ways the software may fail, and providing a safety mechanism for each and (c) checking for racial, gender, age etc. biases in the algorithm on an ongoing basis. Thibodeaux (2017) recommends physically securing facilities (for example, power, gas, and water systems under AI control), encrypting sensitive data and deploying network

intrusion detection technology, and desisting from using a complex technology where a simple technology suffices.[7] Yampolskiy also recommends turning off the AI system from time to time, and allowing competent humans make the judgment calls instead. Alternately, a system allowing deployment of human judgment side-by-side with the AI system—doing similar work—also facilitates injecting a necessary level of diversity.

Second, to mitigate the extent of damage upon materialization of an AI failure, Thibodeaux (2017) suggests providing fail-safes and manual overrides in systems and networks, facilities to forcibly shut down hacked systems till security experts rectify the situation, In like vein, Yampolskiy (2019) recommends having a less smart backup product or service wherever AI is deployed. Yampolskiy (2019) goes on to suggest that when AI fails, senior management should apologize and provide information on measures being taken to rectify the situation.

Third, just as machines become outdated after some time, it is increasingly being recognized that AI algorithms too become outdated, requiring removal. In a video made available by INFORMS, a Comcast executive of data science makes this point (INFORMS 2020). The video also contains a discussion on methods to check whether an algorithm has outlived its usefulness. Thus, standard operating procedures in a company using AI solutions need to evolve such that (a) the performance of the AI models get audited periodically and (b) AI models that have deteriorated significantly over time get discarded.

3.2 Findings related to purpose, context, and interactions of an AI-ML system

Our study helps unearth certain additional directions for inquiry that has the potential to enhance the robustness of AI-ML systems greatly. To this end, we identified the following five emerging themes. (**I**) Use AI to source a vastly higher number of test scenarios, to augment the human-suggested test scenarios (**II**) Create repositories of obvious and obviously true information that has not been articulated on the Internet, that can be readily accessed by AI-ML systems (**III**) Improving AI systems such that they get better in distinguishing a signal that is meant to trigger its functioning from all other signals—howsoever generated—that ought not to trigger its functioning (**IV**) Strive to obtain better mapping of purpose (of coming into being of an AI system) with the capability the AI system is armed with, to have the AI system consider inputs at an appropriate level of granularity. (**V**) Conduct further research to understand what a robot or an AI system learns, when conditions very divergent from those of normal use materialize, to discover ways to stave off/filter improper learning to the extent possible.

We note that an AI-ML solution designed for a specific purpose is likely to work properly only in the context it was designed to operate in, perhaps alongside a few other AI-ML and/or automation solutions for other specific purposes pertaining to the same context (thereby sharing a limited number of interdependencies). Our study suggests that AI-ML system malfunction may be anticipated the moment the purpose served by an AI system is broadened beyond that envisaged by the original designers, and/or when the context changes, and/or when there are changes in the AI-ML and/or automation systems having interaction with the focal AI-ML system.

3.2.1 Purpose of an AI-ML system

The purpose for which an AI-ML system is built is not "known" to (i.e. not subsumed in) the AI-ML system—it is available only with the human designers of the system. Absent an "understanding" of "purpose", an AI-ML system is at a loss as to what level of granularity to "see" or "cognize" phenomena around it. This creates a range of problems. The AI system may find it hard to distinguish a call to action or relevant information from noise. This issue can assume more complicated proportions than, say, Amazon Echo turning on music from Spotify at full volume in an empty apartment in the middle of the night (Olschewski 2017), or sending recordings household conversations to random contacts (Wamsley 2018), or an AI-powered Pixellot camera system mistaking the referee's bald for the soccer ball (Cai and Yuan 2021).

For instance, when the AI system "sees" more, i.e. at a higher level of granularity than that necessary for human-like understanding (and appropriate action), we have a situation as in the example given in Vincent (2017): a burst of TV static overlaid on a picture of a Panda gets incorrectly recognized as a gibbon (whereas humans still "see" the Panda). Alternately, when the AI system "sees" less (i.e. at a lower level of granularity that humans "see" and make correct calls) an AI system makes mistakes like identifying a dragonfly sitting on a woven cloth as a manhole cover, or identifying a fox squirrel standing up as a sea lion, etc.

3.2.2 Context of functioning of an AI-ML system

Changes to context—or, an AI-ML system failing to situate its decision-making in an appropriate context—can cause system malfunctioning. An example of the former is given by the case where a Chinese businesswoman got misidentified as a jaywalker simply because her picture appeared in an advertisement on a moving bus on the road (Thomas

[7] For example, a pencil suffices just fine for writing on paper in zero-gravity situations inside spaceships, a heavy investment to make an ink-pen work under zero-gravity may not be justified.

2020). An example of the latter case is noted in Facebook's automatic translation software mistranslating "good morning" as "attack them", upon failing to resolve the context—Arabic vs. Hebrew text appropriately (Berger 2017). Certain other failures in this genre may be traced merely to lack of adequate testing with diverse data: the cases of Twitter cropping out images of persons with darker skin tones (Asher-Schapiro 2020) and Zoom virtual background removing the head of an African-American faculty teaching on Zoom (Dickey 2020). In sum, AI-ML system designers need to be on the lookout for any expansion in the purpose of use, and any change in context of use, to anticipate and head-off system failure.

3.2.3 Interactions with other systems

We have also noted that an AI-ML system's interactions with other AI-ML or automation systems, or even interactions within components of itself can produced undesired or unforeseen outcomes. The case of Amazon Echo turning on music from Spotify at full volume in an empty apartment in the middle of the night could very well be due to defects in the input–output protocols between the two systems that surfaced upon an upgrade or patch or unresolved error condition in either system. Likewise, we noted that Google Photos and Google Assistant teamed up to create a bizarre/absurd panorama (Zhang 2018), and Google Home Mini attempted to listen to the TV and sent up recordings back to the system providers (Krauth 2018). We are of the opinion that AI-ML work itself needs to be organized better, to deliver targeted functionality satisfactorily. The excitement of discovering newer uses of AI-ML need to gradually make way for a deliberate plan to develop coherent AI-ML ecosystems.

3.3 Human-like frailties showing up in AI-ML systems

3.3.1 Availability bias, mimicking bounded rationality

In a subset of the cases discussed we observed that AI picks human-like follies. For example, Google Autocomplete displayed an unwarranted degree of obsession with Mr. John Hamm's ethnicity (Palis 2017). It also appeared to display similar availability bias (Tversky and Kahneman 1974) by associating an innocent Japanese individual with crimes of another person—simply because they had the same name—harming the former when prospective employers shied away from interviewing him for job positions, mistaking him for the criminal person (Palis 2017). Microsoft's chatbot Tay's proclivity to pick up and articulate hate speech (Paul 2016), a failing that continued in the successor chatbot, Zo (Kantrowitz 2017), provide further instances.

An important root cause of the AI failures above is that, an AI-ML system, of and by itself, is in the dark as to the purpose a user is consulting the system for. Moreover, we know that AI designers cannot anticipate all purposes that a user may approach an AI system for. In this context, Luca et al. (2016, p. 98) state that: "while people understand soft goals and trade-offs, algorithms will pursue a specified objective single-mindedly". An important outcome is that we observe availability bias in functioning of AI-ML systems, resulting in undesired consequences as discussed above.

We note that certain choices are advanced simply because the system already has the "answers" corresponding to those choices indexed and ready. In effect, a human flaw—avoiding being open to new information by preferring to parse information by a well-worn template (possibly originating in bounded rationality)—got inadvertently incorporated into AI systems. We question whether this has to be the case, all the time. Unlike humans, computers can recall and process a very large amount of information in a fraction of a second; saddling them with bounded rationality (as shown above) is not inevitable. For starters, searching on a person's name, say "xx", could bring up generic categories in autocomplete applications, e.g., "Positive press about persons with this name [xx]", "Negative press about persons with this name [xx]", "Business matters about persons sharing this name [xx]", and so forth, where the number (xx) within square brackets display the number of distinct persons involved in the respective classification category (rather than counts of numbers of records available). Note that we ask future AI-ML systems to use its capabilities more—say in terms of distinguishing 'negative press' articles from 'positive press' articles, and making an effort to classify information on specific individuals rather than throw the whole kitchen sink—leveraging its lesser bounds of rationality, to function in a more useful way.

3.3.2 Disagreements among humans spilling over to AI-ML systems

We noted that one of the reasons for failure of the GPT-3 AI-ML system can be traced to lack of access to veridical information. Thus, in spite of digesting a major chunk of information on the Internet, the GPT-3 didn't know that a blade of grass does not have an eye, and neither does a human toe, etc. We also noted that the problem goes beyond providing GPT-3 access to repositories of obvious and obviously true information that has not been articulated on the Internet. While it may not be very difficult to separate fact, fiction and fantasy per se, it is unlikely that teething disagreements on what the salient facts are with respect to any topic of human interest, and what constitutes subjective vs. objective information, will go away any time soon.

Moreover, the ongoing efforts towards canceling words with negative racial or hateful connotation is creating another problem: humans are increasingly seen as being unable to agree to definitions of basic categories. Thereby we note that it is going to be difficult to agree on a fact being a fact, since facts use basic categories as constituent. We also saw that left to itself, AI-ML is rather poor at creating useful categories; a case to point being Facebook's AI-driven self-service platform (for purchasing ads) creating categories such as "How to burn < members of a particular ethnic minority > " or "History of 'why < members of a particular ethnic minority > ruin the world.'" In our view, AI-ML needs to be multi-paradigmatic in that the information fed to it needs to be sharply tuned to the purpose of the AI-ML system in the chosen context. There is an urgent need for review of the current practice of feeding AI-ML systems with data originating in any context that the data-gatherers could lay their hands on.

3.4 Necessity of human-like characteristics in AI-ML systems

We have also noted that, in some contexts, the AI-ML system needs to pick human-like abilities to be useful. For example, the AI robots performing the tasks of front-desk, cleaners, porters and in-room assistants in the AI-ran Henna hotel in Japan in 2015 failed to solve customer issues satisfactorily because (a) they lacked a feature of consulting with a colleague doing a different function when a customer-reported problem spans multiple functions and because (b) they lacked a supervisor who has the authority to take decisions having impact across functions. Likewise, an injured robot [say in a shopping mall, Cai and Yuan 2021)] needs to develop a human-like function of reducing its span of activities drastically, failing which successive errors can create a domino effect, resulting in a big disaster. Again, we note that AI-ML robots are likely to be in the catch-up mode, to decipher human understanding of the context and the associated human mental model connecting tangible and intangible entities. This is an ongoing effort, since human language and preferences change in response to changes in the environment and society. Lastly, we also note that the feasibility, practicability and morality of situations involving AI one side, and a human on the other side of a transaction are not well understood today, calling for nuanced inquiry.

4 Concluding remarks

We set out to classify a set of AI failures into a framework involving omission and commission errors in input, processing and output. Though we were limited by the extent of public availability of information of such failures, we were able to find representative examples for nearly all the cells in our framework; the caveat is that, but for the simplifying assumptions we applied, several cases could fall in the ambit of more than one cell. By considering incidents from a range of AI use cases—image recognition, natural language processing, processing of unstructured texts as well as more complex applications combing several AI technologies (e.g. driverless vehicles)—we demonstrate certain common threads underlying AI failures, cutting through the technology-themed silos. We hope that our efforts help de-mystify the working of AI that uses machine learning. The alternative, throwing up our hands simply because the human-incomprehensible hidden layers of predictor variables constitute the source of all malaise, is not helpful. The framework points to avenues for weeding out weak AI algorithms and allowing only the fittest to survive. Several promising lines of inquiry come to the fore for developing more robust AI-ML systems.

The advent of machine-learning AI solutions promises to help humans cope with the floods of data that modern civilization tends to produce, and coping with which is beyond human cognitive capacity. AI has actually done quite well in situations in which human agency is low, but complexity is high, for example in prediction of machine breakdowns based on readings from multiple sensors. We hope that our work helps in designing better AI systems where the benefits from identifying true positives and true negatives outweigh the loss in productivity arising from wasted effort in dealing with the false positives and false negatives. As AI proliferates though, we expect to see newer kinds of issues, stemming from AI's inability to deal with functional and intentional explanations (Elster 1983), comprising of, respectively, biological organisms' desire for survival, and human beings' needs for imparting morality in decision-making.

Appendix

A new kind of AI system, the augmented-trained-AI system

We have noted that, in many cases, AI-ML developers are not in a position to anticipate and provide for the full range of circumstances that the AI system will be called upon to reckon with (emerging theme **I** in the main text). In some cases, AI developers resort to learning-by-doing. For example, a driverless vehicle is made to navigate a range of road, traffic, and weather conditions. Engineers make notes of situations that are not handled well. These situations are subsequently brought to the notice of AI

developers. The latter update AI algorithms. In our view this approach needs to be augmented by inputs from a new way to developing AI system that we designate as augmented-trained-AI system (ATAI). An ATAI system serves AI-ML systems by feeding scenarios. The key difference between ATAI and AI lies in the way the systems get trained. AI systems are usually trained by real historical data and/or data from a reinforcement learning system where the AI is simultaneously operating and learning in the real world. In addition to those learnings, an ATAI system will get further boost in training by a scenario-capturing and scenario-reframing system. The scenario-capturing part of the system is typically either a data transformer that takes as its inputs corpuses originating from a very different source than the traditional sources of input collected from historical or reinforcement training of the AI, and/or a simulation system that simulates an entire world of possible scenarios using some known rules (which may sometimes themselves be AI-learnt) of interacting agents that are responsible for the generation of the training corpus of the AI in the first place. The scenario-reframing part of the system is tasked to translate the unique scenarios into cases or problems for AI—that is deployed for a specific task like operating a driverless vehicle—to solve.

How are the scenarios pertaining to driverless vehicles constructed? The proposed ATAI system is tasked with going through millions of hours of footage from traffic cameras worldwide, to discover the intelligence necessary to sense and capture unusual traffic incidents. Simulation of the scenarios underlying the incidents subsequently get deployed to train driverless vehicles.

Likewise, millions of hours of videos of human movements in movies, theater, sports, concerts and other places can be used by another ATAI system to discover the intelligence necessary for training AI-ML systems to make part-whole distinctions of the kind that will prevent mistakes like misidentifying someone scratching the cheek as holding a mobile phone Allen (2019), or confusing between a soccer ball and the bald head of a participant (Cai and Yuan 2021), or making errors in computing relative sizes in constructing a vista observed in the Google photos case (Zhang 2018). An ATAI system involving moving entities may need to be multi-paradigmatic, depending on the time-scale that is relevant to the specific purpose an AI-ML system is deployed for. For example, certain phenomena of interest may develop rather slowly, compared to developments in a soccer game or in a Formula 1 race, etc.

To bolster and train AI-ML systems for image or facial recognition (or other recognition-themed tasks), another ATAI system is necessary for discovery of the intelligence underlying (a) the same human's face looking different in different settings and (b) two (or more) faces of different humans looking similar, is necessary We may note though, this kind of ATAI system may also need to be multi-paradigmatic. In one paradigm, the meta-AI system will deliberately "see" in coarse grains, as applicable to avoid making the mistake in the case where the picture of a Panda overlaid with a burst of TV static got misidentified as a Gibbon (Vincent 2017). In a second paradigm the ATAI system will "see" in finer grains as applicable to avoid the mistake made in the case where a fox squirrel standing up got misidentified as a sea lion. It remains to be seen whether the task of deciding which paradigm to switch on, in particular decision contexts, can also be eventually bestowed to AI. We anticipate an exciting period of discovery and learning for AI connoisseurs.

Acknowledgements We thank Prof. Arogyaswamy Paulraj for comments and suggestions on earlier drafts of our paper. All errors remain the authors' sole responsibility.

Availability of data and material The data referred to in this research is sourced entirely from the public domain, links for which have been provided either in the main document itself or in the 'References' section.

Declarations

Conflict of interest The authors did not receive support from any organization for the submitted work. The authors have no relevant financial or non-financial interests to disclose. All authors contributed equally to the research presented in the manuscript. All authors read and approved the final manuscript.

References

Ackoff RL (1994) It's a Mistake! Syst Pract 7(1):3–7. https://doi.org/10.1007/BF02169161

Allen K (2019) Chinese driver gets ticket for scratching his face. BBC. (May 24). https://www.bbc.com/news/blogs-news-from-elsewhere-48401901. Accessed 24 Mar 2021.

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, (May 23). www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 24 Mar 2021

Asher-Schapiro A (2020) Questions swirl about possible racial bias in Twitter image function. The Sunday Morning Herald, (September 22). https://www.smh.com.au/world/north-america/questions-swirl-about-possible-racial-bias-in-twitter-image-function-20200922-p55xwm.html. Accessed 24 Mar 2021

Babic B, Cohen IG, Evgeniou T et al (2021) When machine learning goes off the rails. Harv Bus Rev 99(1):76–84. https://psnet.ahrq.gov/issue/when-machine-learning-goes-rails-guide-managing-risks

Berger Y (2017) Israel Arrests Palestinian Because Facebook Translated 'Good Morning' to 'Attack Them'. Haaretz, (October 22). https://www.haaretz.com/israel-news/palestinian-arrested-over-mistranslated-good-morning-facebook-post-1.5459427. Accessed 24 Mar 2021

Bryen S (2020) Debris is only part of Boeing's problems. Asia Times. (February 23). https://asiatimes.com/2020/02/debris-is-only-part-of-boeings-problems/. Accessed 24 Mar 2021

Cai F, Yuan Y (2021) 2020 in review: 10 AI failures. Synced Review, (January 1) Editor: Michael Sarazen https://syncedreview.com/2021/01/01/2020-in-review-10-ai-failures/. Accessed 24 Mar 2021

Campbell M (2017) Apple's Face ID with attention detection fooled by $200 mask. Apple Insider, (November 28). [https://appleinsider.com/articles/17/11/27/apples-face-id-with-attention-detection-fooled-by-200-mask. Accessed 24 Mar 2021

Cole S (2019) This Trippy T-Shirt Makes You Invisible to AI. Vice. (November 5). https://www.vice.com/en/article/evj9bm/adversarial-design-shirt-makes-you-invisible-to-ai. Accessed 24 Mar 2021

Cossins D (2018) Discriminating algorithms: 5 times AI showed prejudice. NewScientist (April 27). https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/#ixzz6iYaTCbkt. Accessed 24 Mar 2021

Cushing T (2019) NIST study of 189 facial recognition algorithms finds minorities are misidentified almost 100 times more often than white men. TechDirt (December 20). https://www.techdirt.com/articles/20191222/10223143620/nist-study-189-facial-recognition-algorithms-finds-minorities-are-misidentified-almost-100-times-more-often-than-white-men.shtml. Accessed 24 Mar 2021

Das S (2020) Biggest AI goof-ups that made headlines in 2020. Analytics India Magazine, (November 26). https://analyticsindiamag.com/biggest-ai-goof-ups-that-made-headlines-in-2020/. Accessed 24 Mar 2021

Dastin J (2018) Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, (October 11). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Accessed 8 Nov 2022

Dickey MR (2020) Twitter and Zoom's algorithmic bias issues. TechCrunch, (September 22). https://techcrunch.com/2020/09/21/twitter-and-zoom-algorithmic-bias-issues/. Accessed 24 Mar 2021

Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv 4(1): eaao5580. https://doi.org/10.1126/sciadv.aao5580 https://advances.sciencemag.org/content/4/1/eaao5580.full

Elster J (1983) Explaining technical change. Cambridge University Press, Cambridge

Fingas R (2017) Vietnamese firm trips up iPhone X's Face ID with elaborate mask & makeup. Apple Insider, (November 10). https://appleinsider.com/articles/17/11/10/vietnamese-firm-trips-up-iphone-xs-face-id-with-elaborate-mask-makeup. Accessed 24 Mar 2021

Greenberg A (2017) Watch a 10-Year-Old's Face Unlock His Mom's iPhone X. Wired, (November 14). https://www.wired.com/story/10-year-old-face-id-unlocks-mothers-iphone-x/?mbid=social_fb. Accessed 24 Mar 2021

Harwell D (2019) Federal study confirms racial bias of many facial-recognition systems, casts doubt on their expanding use. Washington Post, (December 20). https://www.washingtonpost.com/technology/2019/12/19/federal-study-confirms-racial-bias-many-facial-recognition-systems-casts-doubt-their-expanding-use/. Accessed 24 Mar 2021

Hill K (2020) Wrongfully accused by an algorithm. New York Times, published June 24. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html. Accessed 3 Aug

INFORMS (2020) Use of Machine Learning and AI at Comcast. (November 2). https://www.youtube.com/watch?v=CxkjNA2Fuaw. Accessed 10 Aug 2021

Kantrowitz A (2017) Microsoft's chatbot Zo calls the Qur'an violent and has theories about Bin Laden. *BuzzFeed.News*, (July 3). https://www.buzzfeednews.com/article/alexkantrowitz/microsofts-chatbot-zo-calls-the-quran-violent-and-has#.xh5yOZl2N. Accessed 24 March 2021

Kitching C (2017) Police raid man's home after Amazon Alexa device blasted music and 'held party on its own' at 2am. Mirror. https://www.mirror.co.uk/news/weird-news/police-raid-mans-home-after-11490899. Accessed 8 Nov 2022

Knight W (2016) Tougher Turing Test exposes chatbots' stupidity. MIT Technology Review, (July 14). https://www.technologyreview.com/s/601897/tougher-turing-test-exposes-chatbots-stupidity/. Accessed 8 Nov 2022

Krauth O (2018) Artificial ignorance: the 10 biggest AI failures of 2017. TechRepublic, (January 4). https://www.techrepublic.com/article/the-10-biggest-ai-failures-of-2017/. Accessed 24 Mar 2021

Krisher T (2018) Uber self-driving SUV saw pedestrian but did not brake, federal report finds. Chicago Tribune, (May 24). https://www.chicagotribune.com/business/ct-biz-uber-self-driving-uber-car-report-20180524-story.html. Accessed 24 Mar 2021

Larson J, Mattu S, Kirchner L, Angwin J (2016) How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*, (May 23). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed 24 Mar 2021

Larson S (2017) Offensive chat app responses highlight AI fails. CNN-Money, (October 25). https://money.cnn.com/2017/10/25/technology/business/google-allo-facebook-m-offensive-responses/index.html. Accessed 24 Mar 2021

Levin S (2018) Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian. The Guardian, (March 19). https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe. Accessed 24 Mar 2021

Luca M, Kleinberg J, Mullainathan S (2016) Algorithms need managers, too. Harv Bus Rev 96(1):96–101. https://sendhil.org/algorithms-need-managers-too/

Maeda N, Parker PM (2003) Mind over matter: a case for artificial intelligence. Insead, INS785

McCausland P (2019) Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk. NBC News, (November 10). https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281. Accessed 24 March 2021

Merton RK (1936) The unanticipated consequences of purposeful social action. Am Sociol Rev 1:894–904

Nachreiner C (2021) Apple's Face ID: no match for multifactor security. TechBeacon. https://techbeacon.com/security/apples-face-id-no-match-multifactor-security. Accessed 8 Nov 2022

O'Kane S (2019) Self-driving shuttle crashed in Las Vegas because manual controls were locked away. The Verge, (July 11). https://www.theverge.com/2019/7/11/20690793/self-driving-shuttle-crash-las-vegas-manual-controls-locked-away. Accessed 24 Mar 2021

Olschewski M (2017) A German Alexa owner returned home to find his Amazon device had started a 'party' at 2am, leading to police breaking down his door. Business Insider, (November 9). https://www.businessinsider.com/amazon-alexa-started-party-2am-police-broke-down-door-2017-11?IR=T. Accessed 24 Mar 2021

Oppenheim M (2018) Amazon scraps 'sexist AI' recruitment tool. The Independent, (October 11). https://www.independent.co.uk/life-style/gadgets-and-tech/amazon-ai-sexist-recruitment-tool-algorithm-a8579161.html. Accessed 24 Mar 2021

Owen M (2019) Face ID attention detection security defeated with glasses and tape. Apple Insider, (August 8) https://appleinsider.com/articles/19/08/08/face-id-security-defeated-with-glasses-and-tape. Accessed 24 Mar 2021

Palis C (2017) Google Instant's allegedly 'anti-semitic' results lead to lawsuit in France. Huffington Post, (December 6). https://www.huffpost.com/entry/google-instant-anti-semitic-france_n_1465430. Accessed 24 Mar 2021

Paul I (2016) The Internet turns Tay, Microsoft's millennial AI chatbot, into a racist bigot. PCWorld, (March 24). https://www.pcworld.com/article/3048157/the-internet-turns-tay-microsofts-millennial-ai-chatbot-into-a-racist-bigot.html. Accessed 24 Mar 2021

Pollack A (1983) Technology; The computer as translator. New York Times, (April 28). https://www.nytimes.com/1983/04/28/business/technology-the-computer-as-translator.html. Accessed 24 Mar 2021

Restle B (2020) Is Twitter's image-cropping feature racist? DW, (September 28). https://www.dw.com/en/twitter-image-cropping-racist-algorithm/a-55085160. Accessed 24 Mar 2021

Russell SJ, Norvig P (2003) Artificial intelligence: a modern approach

Satariano A, Metz C (2020) A warehouse robot learns to sort out the tricky stuff. New York Times, (January 29). https://www.nytimes.com/2020/01/29/technology/warehouse-robot.html. Accessed 24 Mar 2021

Shane J (2018) Do neural nets dream of electric sheep? https://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep. Accessed 24 Mar 2021

Smith A (2017a) German police raid flat after Alexa Amazon Echo 'holds party on its own'. Metro. https://metro.co.uk/2017a/11/09/german-police-raid-flat-after-alexa-amazon-echo-holds-party-on-its-own-7067256/. Accessed 8 Nov 2022

Smith L (2017b) Israel police mistakenly arrest Palestinian man for writing 'good morning' on Facebook. Yahoo!news, (October 23). https://www.yahoo.com/news/israel-police-mistakenly-arrest-palestinian-151403166.html. Accessed 24 Mar 2021

Snow J (2018) Amazon's face recognition falsely matched 28 members of congress with mugshots. Technology & Civil Liberties Attorney, ACLU of Northern California, (July 28) https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28. Accessed 24 March 2021

Stewart J (2018) Why people keep rear-ending self-driving cars. Wired, (October 18). https://www.wired.com/story/self-driving-car-crashes-rear-endings-why-charts-statistics/. Accessed 24 Mar 2021

Thibodeaux T (2017) Smart cities are going to be a security nightmare. Harv Bus Rev H03MUJ. Retrieved from https://hbr.org/2017/04/smart-cities-are-going-to-be-a-security-nightmare. Retrived 21 June 2018

Thomas A (2020) Top 8 funniest and shocking AI failures of all time. Analytics India Magazine. (March 2) https://analyticsindiamag.com/top-8-funniest-and-shocking-ai-failures-of-all-time/. Accessed 24 Mar 2021

Toews R (2020) GPT-3 is amazing—and overhyped. Forbes, (July 19) https://www.forbes.com/sites/robtoews/2020/07/19/gpt-3-is-amazingand-overhyped/?sh=5e49de331b1c. Accessed 24 Mar 2021

Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Sci New Ser 185(4157):1124–1131

Vaas L (2018) Uber car software detected woman before fatal crash but failed to stop. NakedSecurity.com, (May 9). https://nakedsecurity.sophos.com/2018/05/09/uber-car-software-detected-woman-before-fatal-crash-but-failed-to-stop/. Accessed 24 Mar 2021

Vincent J (2017). Magic AI: these are the optical illusions that trick, fool, and flummox computers. The Verge, (April 12). https://www.theverge.com/2017/4/12/15271874/ai-adversarial-images-fooling-attacks-artificial-intelligence. Accessed Mar 24 2021

Wamsley L (2018) Amazon Echo recorded and sent couple's conversation—all without their knowledge. NPR, (May 25). https://www.npr.org/sections/thetwo-way/2018/05/25/614470096/amazon-echo-recorded-and-sent-couples-conversation-all-without-their-knowledge. Accessed 24 Mar 2021

Weller C (2017) The first 'Robot Citizen' in the world once said she wants to 'destroy humans. Business Insider, (October 26). https://www.inc.com/business-insider/sophia-humanoid-first-robot-citizen-of-the-world-saudi-arabia-2017.html. Accessed 24 Mar 2021

Yampolskiy RV (2019) Predicting future AI failures from historic examples. Foresight 21(1):138–152

Zhang M (2018) Google Photos' AI Panorama Failed in the Best Way. PetaPixel, (January 23). https://petapixel.com/2018/01/23/google-photos-ai-panorama-failed-best-way/. Accessed 24 Mar 2021

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.