Unsolved Problems in ML Safety

Dan Hendrycks
UC Berkeley

Nicholas Carlini
Google

John Schulman
OpenAI

Jacob Steinhardt
UC Berkeley

Abstract

Machine learning (ML) systems are rapidly increasing in size, are acquiring new capabilities, and are increasingly deployed in high-stakes settings. As with other powerful technologies, safety for ML should be a leading research priority. In response to emerging safety challenges in ML, such as those introduced by recent large-scale models, we provide a new roadmap for ML Safety and refine the technical problems that the field needs to address. We present four problems ready for research, namely withstanding hazards ("Robustness"), identifying hazards ("Monitoring"), steering ML systems ("Alignment"), and reducing risks to how ML systems are handled ("External Safety"). Throughout, we clarify each problem's motivation and provide concrete research directions.

1 Introduction

As machine learning (ML) systems are deployed in high-stakes environments, such as medical settings [145], roads [185], and command and control centers [38], unsafe ML systems may result in needless loss of life. Although researchers recognize that safety is important [1, 5], it is often unclear what problems to prioritize or how to make progress. We identify four problem areas that would help make progress on ML Safety: robustness, monitoring, alignment, and external safety. While some of these, such as robustness, are long-standing challenges, the success and emergent capabilities of modern ML systems necessitate new angles of attack.

We define ML Safety research as ML research aimed at making the adoption of ML more beneficial, with emphasis on long-term and long-tail risks. We focus on cases where greater capabilities can be expected to decrease safety, or where ML Safety problems are otherwise poised to become more challenging in this decade. For each of the four problems, after clarifying the motivation, we discuss possible research directions that can be started or continued in the next few years. First, however, we motivate the need for ML Safety research.

We should not procrastinate on safety engineering. In a report for the Department of Defense, Frola and Miller [55] observe that approximately 75% of the most critical decisions that determine a system's safety occur early in development [120]. If attention to safety is delayed, its impact is limited, as unsafe design choices become deeply embedded into the system. The Internet was initially designed as an academic tool with neither safety nor security in mind [46]. Decades of security patches later, security measures are still incomplete and increasingly complex. A similar reason for starting safety work now is that relying on experts to test safety solutions is not enough—solutions must also be age tested. The test of time is needed even in the most rigorous of disciplines. A century before the four color theorem was proved, Kempe's peer-reviewed proof went unchallenged for years until, finally, a flaw was uncovered [73]. Beginning the research process early allows for more prudent design and more rigorous testing. Since nothing can be done both hastily and prudently [176], postponing machine learning safety research increases the likelihood of accidents.

Unsolved Problems in ML Safety

	Robustness	Create models that are resilient to adversaries, unusual situations, and Black Swan events.
	Monitoring	Detect malicious use, monitor predictions, and discover unexpected model functionality.
	Alignment	Build models that represent and safely optimize hard-to-specify human values.
	External Safety	Use ML to address risks to how ML systems are handled, such as cyberattacks.

Just as we cannot procrastinate, we cannot rely exclusively on previous hardware and software engineering practices to create safe ML systems. In contrast to typical software, ML control flows are specified by inscrutable weights learned by gradient optimizers rather than programmed with explicit instructions and general rules from humans. They are trained and tested pointwise using specific cases, which has limited effectiveness at improving and assessing an ML system's completeness and coverage. They are fragile, rarely correctly handle all test cases, and cannot become error-free with short code patches [155]. They exhibit neither modularity nor encapsulation, making them far less intellectually manageable and making causes of errors difficult to localize. They frequently demonstrate properties of self-organizing systems such as spontaneously emergent capabilities [22, 31]. They may also be more agent-like and tasked with performing open-ended actions in arbitrary complex environments. Just as, historically, safety methodologies developed for electromechanical hardware [166] did not generalize to the new issues raised by software, we should expect software safety methodologies not to generalize to the new complexities and hazards of ML.

We also cannot solely rely on economic incentives and regulation to shepherd competitors into developing safe models. The competitive dynamics surrounding ML's development may pressure companies and regulators to take shortcuts on safety. Competing corporations often prioritize minimizing development costs and being the first to the market over providing the safest product. For example, Boeing developed the 737 MAX with unsafe design choices to keep pace with its competitors; and as a direct result of taking shortcuts on safety and pressuring inspectors, Boeing's defective model led to two crashes across a span of five months that killed 346 people [174, 54, 109]. Robust safety regulation is almost always developed only after a catastrophe—a common saying in aviation is that "aviation regulations are written in blood." While waiting for catastrophes to spur regulators can reduce the likelihood of repeating the same failure, this approach cannot prevent catastrophic events from occurring in the first place. Regulation efforts may also be obstructed by lobbying or by the spectre of lagging behind international competitors who may build superior ML systems. Consequently, companies and regulators may be pressured to deprioritize safety.

These sources of hazards—starting safety research too late, novel ML system complexities, and competitive pressure—may result in deep design flaws. However, a strong safety research community can drive down these risks. Working on safety proactively builds more safety into systems during the critical early design window. This could help reduce the cost of building safe systems and reduce the pressure on companies to take shortcuts on safety. If the safety research community grows, it can help handle the spreading multitude of hazards that continue to emerge as ML systems become more complex. Regulators can also prescribe higher, more actionable, and less intrusive standards if the community has created ready-made safety solutions.

When especially severe accidents happen, everyone loses. Severe accidents can cast a shadow that creates unease and precludes humanity from realizing ML's benefits. Safety engineering for powerful technologies is challenging, as the Chernobyl meltdown, the Three Mile Island accident, and the Space Shuttle Challenger disaster have demonstrated. However, done successfully, work on safety can improve the likelihood that essential technologies operate reliably and benefit humanity.

2 Robustness

Black Swans
- Adapt to evolving environments
- Endure once-in-a-century events

Adversaries
- Handle diverse perceptible attacks
- Detect unforeseen attacks

Figure 1: Robustness research aims to build systems that endure extreme, unusual, or adversarial events.

2.1 Black Swan and Tail Risk Robustness

Motivation. To operate in open-world high-stakes environments, machine learning systems will need to endure unusual events and tail risks. However, current ML systems are often brittle in the face of real-world complexity and unknown unknowns. In the 2010 Flash Crash [99], automated trading systems unexpectedly overreacted to market aberrations, created a feedback loop, and wiped away a trillion dollars of stock value in a matter of minutes. This demonstrates that computer systems can both create and succumb to long tail events.

Long tails continue to thwart modern ML systems such as autonomous vehicles. This is because some of the most basic concepts in the real world are long tailed, such as stop signs, where a model error can directly cause a crash and loss of life. Stop signs may be titled, occluded, or represented on an LED matrix; sometimes stop signs should be disregarded, for example when held upside down by a traffic officer, on open gates, on a shirt, on the side of bus, on elevated toll booth arms, and so on. Although these long tail events are rare, they are extremely impactful [181] and can cause ML systems to crash. Lever-aging existing massive datasets is not enough to ensure robustness, as models trained with Internet data and petabytes of task-specific driving data still are not robust to long tail road scenarios [185]. This decades-long challenge is only a preview of the more difficult problem of handling tail events in environments that are beyond a road's complexity.

"Nature always sides with the hidden flaw." *Murphy's Ninth Law*

Long-tail robustness is unusually challenging today and may become even more challenging. Long-tail robustness also requires more than human-level robustness; the 2008 financial crisis and COVID-19 have shown that even groups of humans have great difficulty mitigating and overcoming these rare but extraordinarily impactful long tail events. Future ML systems will operate in environments that are broader, larger-scale, and more highly connected with more feedback loops, paving the way to more extreme events [128] than those seen today.

While there are incentives to make systems partly robust, systems tend not to be incentivized nor designed for long tail events outside prior experience, even though Black Swan events are inevitable [191]. To reduce the chance that ML systems will fall apart in settings dominated by rare events, systems must be *unusually* robust.

Directions. In addition to existing robustness benchmarks [78, 101, 75], researchers could create more environments and benchmarks to stress-test systems, find their breaking points, and determine whether they will function appropriately in potential future scenarios. These benchmarks could include new, unusual, and extreme distribution shifts and long tail events, especially ones that are challenging even for humans. We should focus less on "mild" distribution shifts that cause small accuracy declines [50] and more on "wild" shifts that cause large accuracy drops [124].

Robustness work could also move beyond classification and consider *competent errors* where agents misgeneralize and execute wrong routines, such as an automated digital assistant knowing how to use a credit card to book flights, but choosing the wrong destination [100, 90]. Interactive environments [36] could simulate qualitatively distinct random shocks that irreversibly shape the environment's future evolution. Researchers could also create environments where ML system outputs affect their environment and create feedback loops.

3

Using such benchmarks and environments, researchers could improve ML systems to withstand Black Swans [182, 181], long tails, and structurally novel events. The performance of many ML systems is currently largely shaped by data and parameter count, so future research could work on creating highly unusual but helpful data sources. The more experience a system has with unusual future situations, even ones not well represented in typical training data, the more robust it can be. New data augmentation techniques [84] and other sources of simulated data could create inputs that are not easy or possible to create naturally.

Since change is a part of all complex systems, and since not everything can be anticipated during training, models will also need to adapt to an evolving world and improve from novel experiences [129, 195, 180]. Future adaptation methods could improve a system's ability to adapt quickly. Other work could defend adaptive systems against poisoned data encountered during deployment [127].

2.2 Adversarial Robustness

Motivation. We now turn from unpredictable accidents to carefully crafted and deceptive threats. Adversaries can easily manipulate vulnerabilities in ML systems and cause them to make mistakes [14, 177]. For example, systems may use neural networks to detect intruders [4] or malware [173], but if adversaries can modify their behavior to deceive and bypass detectors, the systems will fail. While defending against adversaries might seem to be a straightfoward problem, defenses are currently struggling to keep pace with attacks [7, 188], and much research is needed to discover how to fix these longstanding weaknesses.

Directions. We encourage research on adversarial robustness to focus on broader robustness definitions. Current research largely focuses on the problem of "ℓ_p adversarial robustness," [29] where an adversary attempts to induce a misclassification but can only perturb inputs subject to a small p-norm constraint. While research on simplified problems helps drive progress, researchers may wish to avoid focusing too heavily on any one particular simplification.

To study adversarial robustness more broadly [61], researchers could consider attacks that are perceptible [140] or whose specifications are not known beforehand [96, 111]. For instance, there is no reason that an adversarial malware sample would have to be imperceptibly similar to some other piece of benign software—as long as the detector is evaded, the attack has succeeded [138]. Likewise, copyright detection systems cannot reasonably assume that attackers will only construct small ℓ_p perturbations to bypass the system, as attackers may rotate the adversarially modified image [51] or apply otherwise novel distortions [61] to the image.

While many effective attacks assume full access to a neural network, sometimes assuming limited access is more realistic. Here, adversaries can feed in examples to an ML system and receive the system's outputs, but they do not have access to the intermediate ML system computation [20]. If a blackbox ML system is not publicly released and can only be queried, it may be possible to practically defend the system against zero-query attacks [189] or limited-query attacks [34].

On the defense side, further underexplored assumptions are that systems have multiple sensors or that systems can adapt. Real world systems, such as autonomous vehicles, have multiple cameras. Researchers could exploit information from these different sensors and find inconsistencies in adversarial images in order to constrain and box in adversaries [202]. Additionally, while existing ML defenses are typically static, future defenses could evolve during test time to combat adaptive adversaries [194].

Future research could do more work toward creating models with adversarially robust representations [40]. Researchers could enhance data for adversarial robustness by simulating more data [208], augmenting data [149], repurposing existing real data [30, 80], and extracting more information from available data [82]. Others could create architectures that are more adversarially robust [203]. Others could improve adversarial training methods [201] and find better losses [206, 179]. Researchers could improve adversarial robustness certifications [144, 116, 37], so that models have verifiable adversarial robustness.

It may also be possible to unify the areas of adversarial robustness and robustness to long-tail and unusual events. By building systems to be robust to adversarial worst-case environments, they may also be made more robust to random-worse-case environments [6, 85]. To study adversarial robustness on unusual inputs, researchers could also try detecting adversarial anomalies [16, 85] or assigning them low confidence [172].
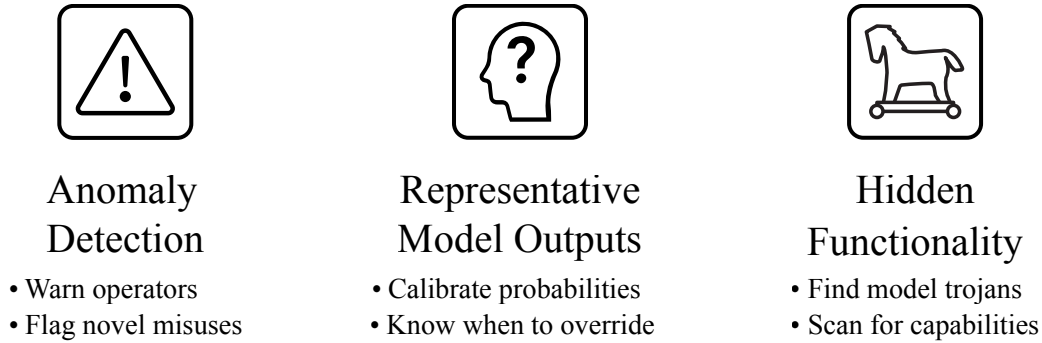
3   Monitoring

Anomaly
Detection

- Warn operators
- Flag novel misuses

Representative
Model Outputs

- Calibrate probabilities
- Know when to override

Hidden
Functionality

- Find model trojans
- Scan for capabilities

Figure 2: Monitoring research aims to identify hazards, inspect models, and help human ML system operators.

3.1   Identifying Hazards and Malicious Use With Anomaly Detection

Motivation.   Deploying and monitoring powerful machine learning systems will require high caution, similar to the caution observed for modern nuclear power plants, military aircraft carriers, air traffic control, and other high-risk systems. These complex and hazardous systems are now operated by high reliability organizations (HROs) which are relatively successful at avoiding catastrophes [47]. For safe deployment, future ML systems may be operated by HROs. Anomaly detectors are a crucial tool for these organizations since they can warn human operators of potential hazards [142]. For detectors to be useful, research must strive to create detectors with high recall and a low false alarm rate in order to prevent alarm fatigue [41].

Separately, anomaly detection is essential in detecting malicious uses of ML systems [23]. Malicious users are incentivized to use novel strategies, as familiar misuse strategies are far easier to identify and prevent compared to unfamiliar ones. Malicious actors may eventually repurpose ML systems for social manipulation [27], for assisting research on novel weapons [18], or for cyberattacks [26]. When such anomalies are detected, the detector can trigger a fail-safe policy in the system and also flag the example for human intervention. However, detecting malicious anomalous behavior could become especially challenging when malicious actors utilize ML capabilities to try to evade detection. Anomaly detection is integral not just for promoting reliability but also for preventing novel misuses.

Directions.   Anomaly detection is actively studied in research areas such as out-of-distribution detection [79], open-set detection [10], and one-class learning [178, 82], but many challenges remain. The central challenge is that existing methods for representation learning have difficulty discovering representations that work well for previously unseen anomalies. One of the symptoms of this problem is that anomaly detectors for large-scale images still cannot reliably detect that previously unseen random noise is anomalous [81]. Moreover, there are many newer settings that require more study, such as detecting distribution shifts or changes to the environment [44], as well developing detectors that work in real-world settings such as intrusion detection, malware detection, and biosafety.

Beyond just detecting anomalies, high reliability organizations require candidate explanations of how an anomaly came to exist [142, 163]. To address this, detectors could help identify the origin or location of an anomaly [13]. Other work could try to help triage anomalies and determine whether an anomaly is just a negligible nuisance or is potentially hazardous.

3.2   Representative Model Outputs

3.2.1   Calibration

Motivation.   Human monitors need to know when to trust a deployed ML system or when to override it. If they cannot discern when to trust and when to override, humans may unduly defer to models and cede too much control. If they can discern this, they can prevent many model hazards and failure modes.

To make models more trustworthy, they should accurately assess their domain of competence [60]—the set of inputs they are able to handle. Models can convey the limits of their competency by expressing their uncertainty. However, model uncertainties are not representative, and they are often overconfident [68]. To address this, models could become more calibrated. If a model is perfectly calibrated and predicts a "70% chance of rain," then when it makes that prediction, 70% of the time it will rain. Calibration research makes model prediction probabilities more representative of a model's overall behavior, provides monitors with a clearer impression of their understanding, and helps monitors weigh model decisions.

Directions. To help models express their domain of competence in a more representative and meaningful way, researchers could further improve model calibration on typical testing data [68, 131, 112, 108, 205, 106, 107, 123], though the greater challenge is calibration on testing data that is unlike the training data [135]. Future systems could communicate their uncertainty with language. For example, they could express decomposed probabilities with contingencies such as "event A will occur with 60% probability assuming event B also occurs, and with 25% probability if event B does not." To extend calibration beyond single-label outputs, researchers could take models that generate diverse sentence and paragraph answers and teach these models to assign calibrated confidences to their generated free-form answers.

3.2.2 Making Model Outputs Honest and Truthful

Motivation. Human monitors can more effectively monitor models if they produce outputs that accurately, honestly, and faithfully [62] represent their understanding or lack thereof. However, current language models do not accurately represent their understanding and do not provide faithful explanations. They generate empty explanations that are often surprisingly fluent and grammatically correct but nonetheless entirely fabricated. These models generate distinct explanations when asked to explain again, generate more misconceptions as they become larger [122], and sometimes generate worse answers when they know how to generate better answers [33]. If models can be made honest and only assert what they believe, then they can produce outputs that are more representative and give human monitors a more accurate impression of their beliefs.

Directions. Researchers could create evaluation schemes that catch models being inconsistent [49], as inconsistency implies that they did not assert only what they believe. Others could also build tools to detect when models are hallucinating information [117]. To prevent models from outputting worse answers when they know better answers, researchers can concretize what it means for models to assert their true beliefs or to give the right impression. Finally, to train more truthful models, researchers could create environments [137] or losses that incentivize models not to state falsehoods, repeat misconceptions [122], or spread disinformation.

3.3 Hidden Model Functionality

3.3.1 Backdoors

Motivation. Machine learning systems risk carrying hidden "backdoor" or "trojan" controllable vulnerabilities. Backdoored models behave correctly and benignly in almost all scenarios, but in particular circumstances chosen by an adversary, they have been taught to behave incorrectly [67]. Consider a backdoored facial recognition system that gates building access. The backdoor could be triggered by a specific unique item chosen by an adversary, such as an item of jewelry. If the adversary wears that specific item of jewelry, the backdoored facial recognition will allow the adversary into the building [158]. A particularly important class of vulnerabilities are backdoors for sequential decision making systems, where a particular trigger leads an agent or language generation model to pursue a coherent and destructive sequence of actions [197, 207].

Whereas adversarial examples are created at test time, backdoors are inserted by adversaries at training time. One way to create a backdoor is to directly inject the backdoor into a model's weights [154, 89], but they can also be injected by adding poisoned data into the training or pretraining data [156]. Injecting backdoors through poisoning is becoming easier as ML systems are increasingly trained on uncurated data scraped from online—data that adversaries can poison. If an adversary uploads a few carefully crafted poisoned images [28], code snippets [154], or sentences [193] to platforms such as Flickr, GitHub or Twitter, they can inject a backdoor into future models trained on that data [9]. Moreover, since downstream models are increasingly obtained by a single upstream model [17], a single compromised model could proliferate backdoors.

Directions. To avoid deploying models that may take unexpected turns and have vulnerabilities that can be controlled by an adversary, researchers could improve backdoor detectors to combat an ever-expanding set of backdoor attacks [97]. Creating algorithms and techniques for detecting backdoors is promising, but to stress test them we need to simulate an adaptive competition where researchers take the role of both attackers and auditors. This type of competition could also serve as a valuable way of grounding general hidden model functionality detection research. Researchers could try to cleanse models with backdoors, reconstruct a clean dataset given a model [204, 196], and build techniques to detect poisoned training data. Research should also develop methods for addressing backdoors that are manually injected, not just those injected through data poisoning.

3.3.2 Emergent Hazardous Capabilities

Motivation. We are better able to make models safe when we know what capabilities they possess. For early ML models, knowing their limits was often trivial, as models trained on MNIST can do little more than classify handwritten images. However, recent large-scale models often have capabilities that their designers do not initially realize, with novel and qualitatively distinct capabilities emerging as scale increases. For example, as GPT-3 models became larger [22], they gained the ability to perform arithmetic, even though GPT-3 received no explicit arithmetic supervision. Others have observed instances where a model's training loss remains steady, but then its test performance spontaneously ascends from random chance to perfect generalization [141]. Sometimes capabilities are only discovered after initial release. After a multimodal image and text model [143] was released, users eventually found that its synthesized images could be markedly improved by appending "generated by Unreal Engine" to the query [12]. Future ML models may, when prompted carefully, make the synthesis of harmful or illegal content seamless (such as videos of child exploitation, suggestions for evading the law, or instructions for building bombs). These examples demonstrate that it will be difficult to safely deploy models if we do not know their capabilities.

Some emergent capabilities may resist monitoring. In the future, it is conceivable that agent-like models may be inadvertently incentivized to adopt covert behavior. This is not unprecedented, as even simple digital organisms can evolve covert behavior. For instance, Ofria's [118] digital organisms evolved to detect when they were being monitored and would "play dead" to bypass the monitor, only to behave differently once monitoring completed. In the automotive industry, Volkswagen created products designed to bypass emissions monitors, underscoring that evading monitoring is sometimes incentivized in the real world. Advanced ML agents may be inadvertently incentivized to be deceptive not out of malice but simply because doing so may help maximize their human approval objective. If advanced models are also capable planners, they could be skilled at obscuring their deception from monitors.

Directions. To protect against emergent capabilities, researchers could create techniques and tools to inspect models and better foresee unexpected jumps in capabilities. We also suggest that large research groups begin scanning models for numerous potential and as yet unobserved capabilities. We specifically suggest focusing on capabilities that could create or directly mitigate hazards. One approach is to create a continually evolving testbed to screen for potentially hazardous capabilities, such as the ability to execute malicious user-supplied code, generate illegal or unethical forms of content, or to write convincing but wrong text on arbitrary topics. Another more whitebox approach would be to predict a model's capabilities given only its weights, which might reveal latent capabilities that are not obviously expressible from standard prompts.

Detection methods will require validation to ensure they are sufficiently sensitive. Researchers could implant hidden functionality to ensure that detection methods can detect known flaws; this can also help guide the development of better methods. Other directions include quantifying and extrapolating future model capabilities [86, 87] and searching for novel failure modes that may be symptoms of unintended functionality.

Once a hazardous capability such as deception or illegal content synthesis is identified, the capability must be prevented or removed. Researchers could create training techniques so that undesirable capabilities are not acquired during training or during test-time adaptation. For ML systems that have already acquired an undesirable capability, researchers could create ways to teach ML systems how to forget that capability. However, it may not be straightforward to determine whether the capability is truly absent and not merely obfuscated or just removed partially.

4 Alignment

Challenges With Aligning Objectives

Specification
- Track nebuluous goals
- Learn complex objectives

Brittleness
- Prevent overoptimization
- Proxy gaming

Optimization
- Pursue only the main objective
- Tradeoffs of complex goals

Unintended Consequences
- Cautious and constrained behavior
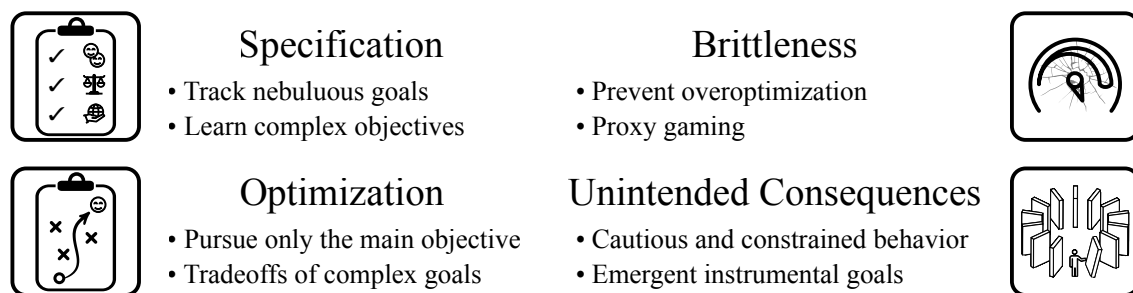- Emergent instrumental goals

Figure 3: Alignment research aims to create and safely optimize ML system objectives.

While most technologies do not have goals and are simply tools, future machine learning systems may be more agent-like. How can we build ML agents that prefer good states of the world and avoid bad ones? Objective functions drive system behavior, but aligning objective functions with human values requires overcoming societal as well as technical challenges. We briefly discuss societal challenges with alignment and then describe technical alignment challenges in detail.

Putting aligned goals in powerful future ML systems may be challenging because their goals may be given by unaligned companies, not necessarily the public. Unfortunately, sometimes corporate incentives can be distorted in the pursuit of maximizing shareholder value [93], so the goals that are given to ML systems may not be aligned with societal values. While many companies can help satisfy human desires, some companies have been incentivized to decimate rain forests [59], lie to customers that cigarettes are healthy [19], invade user privacy [209], and cut corners on safety [175]. Even if economic entities were far more aligned, such as if corporations absorbed their current negative externalities, the larger economic system would still not be aligned with human values. This is because the overall activity of the economy can be viewed as approximating material wealth maximization [139]. However, once wealth increases enough, it ceases to be correlated with emotional wellbeing and happiness [95]. Furthermore, wealth maximization with advanced ML may sharply exacerbate inequality [65], which is a robust predictor of aggression and conflict [53]. Under extreme automation in the future, wealth metrics such as real GDP per capita may drift further from tracking our values [25]. Given these considerations, the default economic objective is not fully aligned with human values.

Even if societal issues are resolved and ideal goals are selected, other problems remain. We focus on four important technical alignment problems: objective proxies are difficult to specify, objective proxies are difficult to optimize, objective proxies can be brittle, and objective proxies can spawn unintended consequences.

4.1 Objectives Can Be Difficult to Specify

Motivation for Value Learning. Encoding human goals and intent is challenging. Lawmakers know this well, as laws specified by stacks of pages still often require that people interpret the spirit of the law. Many human values, such as happiness [115], good judgment [167], meaningful experiences [52], human autonomy, and so on, are hard to define and measure. Systems will optimize what is measurable [150], as "what gets measured gets managed." Measurements such as clicks and watch time may be easily measurable, but they often leave out and work against important human values such as wellbeing [104, 52, 170, 171]. Researchers will need to confront the challenge of measuring abstract, complicated, yet fundamental human values.

Directions. Value learning seeks to develop better approximations of our values, so that corporations and policy makers can give systems better goals to pursue. Some important values include wellbeing, fairness, and people getting what they are deserve. To model wellbeing, future work could use ML to model what people find pleasant, how stimuli affect internal emotional valence, and other aspects of subjective experience. Other work could try to learn how to align specific technologies, such as recommender systems, with wellbeing goals rather than engagement. Future models deployed in legal contexts must understand justice, so models should be taught the law [77]. Researchers could create models that learn wellbeing functions that do not mimic cognitive

biases [76]. Others could make models that are able to detect when scenarios are clear-cut or highly morally contentious [76]. Other directions include learning difficult-to-specify goals in interactive environments [70], learning the idiosyncratic values of different stakeholders [121], and learning about cosmopolitan goals such as endowing humans with the capabilities necessary for high welfare [133].

4.2   Objectives Can Be Difficult to Optimize

Motivation for Translating Values Into Action.   Putting knowledge from value learning into practice may be difficult because optimization is difficult. For example, many sparse objectives are easy to specify but difficult to optimize. Worse, some human values are particularly difficult to optimize. Take, for instance, the optimization of wellbeing. Short-term and long-term wellbeing are often anticorrelated, as the hedonistic paradox shows [164]. Hence many local search methods may be especially prone to bad local optima, and they may facilitate the impulsive pursuit of pleasure. Consequently, optimization needs to be on long timescales, but this reduces our ability to test our systems iteratively and rapidly, and ultimately to make them work well. Further, human wellbeing is difficult to compare and trade off with other complex values, is difficult to forecast even by humans themselves [200], and wellbeing often quickly adapts and thereby nullifies interventions aimed at improving it [21]. Optimizing complex abstract human values is therefore not straightforward.

To build systems that optimize human values well, models will need to mediate their knowledge from value learning into appropriate action. Translating background knowledge into choosing the best action is typically not straightforward: while computer vision models are advanced, successfully applying vision models for robotics remains elusive. Also, while sociopaths are intelligent and have moral awareness, this knowledge does not necessarily result in moral inclinations or moral actions.

As systems make objectives easier to optimize and break them down into new goals, subsystems are created that optimize these new intrasystem goals. But a common failure mode is that "intrasystem goals come first" [57]. These goals can steer actions instead of the primary objective [90]. Thus a system's explicitly written objective is not necessarily the objective that the system operationally pursues, and this can result in misalignment.

Directions.   To make models optimize desired objectives and not pursue undesirable secondary objectives, researchers could try to construct systems that guide models not just to follow rewards but also behave morally [83]; such systems could also be effective at guiding agents not to cause wanton harm within interactive environments and to abide by rules. To get a sense of an agent's values and see how it make tradeoffs between values, researchers could also create diverse environments that capture realistic morally salient scenarios and characterize the choices that agents make when faced with ethical quandaries. Research on steerable and controllable text generation [103, 98] could help chatbots exhibit virtues such as friendliness and honesty.

4.3   Objective Proxies Can Be Brittle

Proxies that approximate our objectives are brittle, but work on Proxy Gaming and Value Clarification can help.

Motivation for Proxy Gaming.   Objective proxies can be gamed by optimizers and adversaries. For example, to combat a cobra infestation, a governor of Delhi offered bounties for dead cobras. However, as the story goes, this proxy was brittle and instead incentivized citizens to breed cobras, kill them, and collect a bounty. In other contexts, some students overoptimize their GPA proxies by taking easier courses, and some academics overoptimize bibliometric proxies at the expense of research impact. Agents in reinforcement learning often find holes in proxies. In a boat racing game, an RL agent gained a high score not by finishing the race but by going in the wrong direction, catching on fire, and colliding into other boats [35]. Since proxies "will tend to collapse once pressure is placed upon" them by optimizers [64, 125, 169], proxies can often be gamed.

Directions.   Advancements in robustness and monitoring are key to mitigating proxy gaming.

"When a measure becomes a target, it ceases to be a good measure."   *Goodhart's Law*

ML systems encoding proxies must become more robust to optimizers, which is to say they must become more adversarially robust (Section 2.2). Specifically, suppose a neural network is used to define a learned utility function; if some other agent (say another neural network) is tasked with maximizing this utility proxy, it would be incentivized to find and exploit any errors in the learned utility proxy, similar to adversarial examples [187, 63]. Therefore we should seek to ensure adversarial

robustness of learned reward functions, and regularly test them for exploitable loopholes.

Separately, advancements in monitoring can help with proxy gaming. For concreteness, we discuss how monitoring can specifically help with "human approval" proxies, but many of these directions can help with proxy gaming in general. A notable failure mode of human approval proxies is their susceptibility to deception. Anomaly detectors (Section 3.1) could help spot when ML models are being deceptive or stating falsehoods, could help monitor agent behavior for unexpected activity, and could help determine when to stop the agent or intervene. Research on making models honest and teaching them to give the right impression (Section 3.2) can help mitigate deception from models trying to game approval proxies. To make models more truthful and catch deception, future systems could attempt to verify statements that are difficult for humans to check in reasonable timespans, and they could inspect convincing but not true assertions [137]. Researchers could determine the veracity of model assertions, possibly through an adversarial truth-finding process [92].

Motivation for Value Clarification.   While maximization can expose faults in proxies, so too can future events. The future will sharpen and force us to confront unsolved ethical questions about our values and objectives [199]. In recent decades, peoples' values have evolved by confronting philosophical questions, including whether to infect volunteers for science, how to equitably distribute vaccines, the rights of people with different orientations, and so on. How are we to act if many humans spend most of their time chatting with compelling bots and not much time with humans, or how should we fairly address automation's economic ramifications? Determining the right action is not strictly scientific in scope [91], and we will need philosophical analysis to help us correct structural faults in our proxies.

Directions.   We should build systems to help rectify our objectives and proxies, so that we are less likely to optimize the wrong objective when a change in goals is necessary. This requires interdisciplinary research towards a system that can reason about values and philosophize at an expert level. Research could start with trying to build a system to score highly in the philosophy olympiad, in the same way others are aiming to build expert-level mathematician systems using mathematics olympiad problems [126]. Other work could build systems to help extrapolate the end products of "reflective equilibrium" [147], or what objectives we would endorse by simulating a process of deliberation about competing values. Researchers could also try to estimate the quality of a philosophical work by using a stream of historical philosophy papers and having models predict the impact of each paper on the literature. Eventually, researchers should seek to build systems that can formulate robust positions through an argumentative dialog. These systems could also try to find flaws in verbally specified proxies.

4.4   Objective Proxies Can Lead to Unintended Consequences

Motivation.   While optimizing agents may work towards subverting a proxy, in other situations both the proxy setter and an optimizing agent can fall into states that neither intended. For example, in their pursuit to modernize the world with novel technologies, previous well-intentioned scientists and engineers inadvertently increased pollution and hastened climate change, an outcome desired neither by the scientists themselves nor by the societal forces that supported them. In ML, some platforms maximized clickthrough rates to approximate maximizing enjoyment, but such platforms unintentionally addicted many users and decreased their wellbeing. These cases demonstrate that unintended consequences present a challenging but important problem.

Directions.   Future research could focus on designing minimally invasive agents that prefer easily reversible to irreversible actions [66], as irreversibility reduces optionality and often unintentionally destroys potential future value. Likewise, researchers could create agents that properly account for their lack of knowledge of the true objective [69] and avoid disrupting parts of the environment whose value is unclear [190, 102, 157]. We also need more complex environments that can manifest diverse unintended side effects [192] such as feedback loops, which are a source of hazards to users of recommender systems [105]. A separate way to mitigate unintended consequences is to teach ML systems to abide by constraints [148, 153], be less brazen, and act cautiously. Since we may be uncertain about which values are best, research could focus on having agents safely optimize and balance many values, so that one value does not unintentionally dominate or subvert the rest [130, 48]. Sometimes unintended instrumental goals emerge in systems, such as the goal of self-preservation [69], so researchers could find new settings that cause new goals to emerge and others could try detecting unintended emergent goals; see Section 3.3.2 for more directions in detecting emergent functionality.

5 External Safety

ML for Cybersecurity

• ML for patching insecure code
• ML for detecting cyberattacks

Informed Decision Making

• Forecasting events and effects
• Raising crucial considerations

Figure 4: External safety research aims to address broader contextual risks to how ML systems are handled. Both cybersecurity and decision making may decisively affect whether ML systems will fail or be misdirected.

Machine learning systems do not exist in a vacuum, and the safety of the external context can influence how ML systems are handled and affect the overall safety of ML systems. ML systems are more likely to fail or be misdirected if the external context in which they operate is insecure or turbulent.

External safety research applies ML to mitigate potential contextual hazards that may decisively cause ML systems to fail or be misdirected. As two examples, we support research on cybersecurity and on informed decision making. The first problem is motivated by the observation that ML systems are integrated with vulnerable software, and in the future ML may change the landscape of cyberattacks. In the second problem, we turn to a speculative approach for improving governance decisions and command and control operations using ML, as institutions may direct the most powerful future ML systems.

Beyond technical work, policy and governance work will be integral to safe deployment [42, 11, 15, 210, 24]. While techno-solutionism has limitations, technical ML researchers should consider using their skillset to address deployment environment hazards, and we focus on empirical ML research avenues, as we expect most readers are technical ML researchers.

Finally, since there are multiple hazards that can hinder external safety, this section is nonexhaustive. For instance, if ML industry auditing tools could help regulators more effectively regulate ML systems, research developing such tools could become part of external safety. Likewise, using ML to help facilitate cooperation [43] may emerge as a research area.

5.1 ML for Cybersecurity

Motivation. Cybersecurity risks can make ML systems unsafe, as ML systems operate in tandem with traditional software and are often instantiated as a cyber-physical system. As such, malicious actors could exploit insecurities in traditional software to control autonomous ML systems. Some ML systems may also be private or unsuitable for proliferation, and they will therefore need to operate on computers that are secure.

Separately, ML may amplify future automated cyberattacks and enable malicious actors to increase the accessibility, potency, success rate, scale, speed, and stealth of their attacks. For example, hacking currently requires specialized skills, but if state-of-the-art ML models could be fine-tuned for hacking, then the barrier to entry for hacking may decrease sharply. Since cyberattacks can destroy valuable information and even destroy critical physical infrastructure [32] such as power grids [134] and building hardware [114], these potential attacks are a looming threat to international security.

While cybersecurity aims to increase attacker costs, the cost-benefit analysis may become lopsided if attackers eventually gain a larger menu of options that require negligible effort. In this new regime, attackers may gain the upper hand, like how attackers of ML systems currently have a large advantage over defenders. Since there may be less of a duality between offensive and defensive security in the future, we suggest that research focus on techniques that are clearly defensive. The severity of this risk is speculative, but neural networks are now rapidly gaining the ability to write code and interact with the outside environment, and at the same time there is very little research on deep learning for cybersecurity.

Directions. To mitigate the potential harms of automated cyberattacks to ML and other systems, researchers should apply ML to develop better defensive techniques. For instance, ML could be used to detect intruders [113, 165] or impersonators [88]. ML could also help analyze code and detect software vulnerabilities, and could help generate unexpected inputs to programs [160, 198, 159]. Massive unsupervised ML methods could also model binaries and learn to detect malicious obfuscated payloads [168, 161, 132, 71]. Researchers could also create ML systems that model software behavior and detect whether programs are sending packets when they should not. ML models could help predict future phases of cyberattacks, and such automated warnings could be judged by their lead time, precision, recall, and the quality of their contextualized explanation. Advancements in code translation [110, 8] and code generation [33, 136] suggest that future models could apply security patches and make code more secure, so that future systems not only flag security vulnerabilities but also fix them.

5.2 Improved Epistemics and Decision Making

Motivation. Even if we create reliable ML systems, these systems will not exhibit or ensure safety if the institutions that steer ML systems make poor decisions. Although nuclear weapons are a reliable and dependable technology, they became especially unsafe during the Cold War. During that time, misunderstanding and political turbulence exposed humanity to several close calls and brought us to the brink of catastrophe, demonstrating that external safety issues can make technologies unsafe. The most pivotal decisions are made during times of crisis, and future crises may be similarly risky as ML continues to be weaponized [151, 2]. This is why we suggest creating tools to help decision-makers handle ML systems in highly uncertain, quickly evolving, turbulent situations.

Directions. To improve the decision-making and epistemics of political leaders and command and control centers, we suggest two efforts: using ML to improve forecasting and bringing to light crucial considerations.

Many governance and command and control decisions are based on forecasts [186] from humans, and some forecasts are starting to incorporate ML [38]. Forecasters assign probabilities to possible events that could happen within the next few months or years (e.g., geopolitical, epidemiological, and industrial events), and are scored by their correctness and calibration. To be successful, forecasters must dynamically aggregate information from disparate unstructured sources [94]. This is challenging even for humans, but ML systems could potentially aggregate more information, be faster, be nonpartisan, consider multiple perspectives, and thus ultimately make more accurate predictions [146]. The robustness of such systems could be assessed based on their ability to predict pivotal historical events, if the model only has access to data before those events. An accurate forecasting tool would need to be applied with caution to prevent over-reliance [74], and it would need to present its data carefully so as not to encourage risk-taking behavior from the humans operating the forecasting system [183].

Separately, researchers should develop systems that identify questions worth asking and crucial factors to consider. While forecasting can refine estimates of well-defined risks, these advisory systems could help unearth new sources of risk and identify actions to mitigate risks. Since ML systems can process troves of historical data and can learn from diverse situations during training, they could suggest possibilities that would otherwise require extensive memory and experience. Such systems could help orient decision making by providing related prior scenarios and relevant statistics such as base rates. Eventually advisory systems could identify stakeholders, propose metrics, brainstorm options, suggest alternatives, and note trade-offs to further improve decision quality [58]. In summary, ML systems that can predict a variety of events and identify crucial considerations could help provide good judgment and correct misperceptions, and thereby reduce the chance of rash decisions and inadvertent escalation.

6 Related Research Agendas

There is a large ecosystem of work on addressing societal consequences of machine learning, including AI policy [42], privacy [3, 162], fairness [72], and ethics [56]. We strongly support research on these related areas. For purposes of scope, in this section we focus on papers that outline paths towards creating safe ML systems.
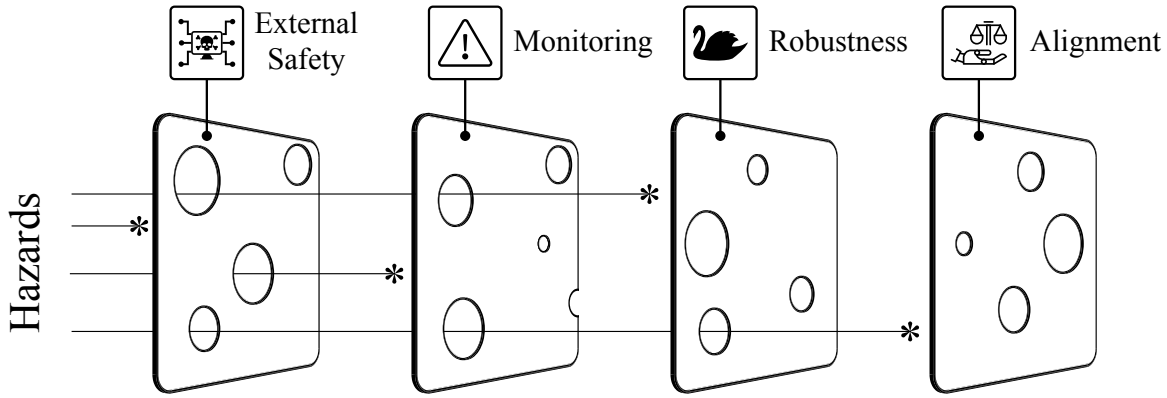
Figure 5: A Swiss cheese model of ML Safety research. Pursuing multiple safety research avenues creates multiple layers of protection which mitigates hazards and makes ML systems safer.

An early work that helps identify safety problems is Russell *et al.*, 2015 [152], who identify many potential avenues for safety, spanning robustness, machine ethics, research on AI's economic impact, and more. Amodei and Olah *et al.*, 2016 [5] helped further concretize several safety research directions. With the benefit of five years of hindsight, our paper provides a revised and expanded collection of concrete problems. Some of our themes extend the themes in Amodei and Olah *et al.*, such as Robustness and some portions of Alignment. We focus here on problems that remain unsolved and also identify new problems, such as emergent capabilities from massive pretrained models, that stem from recent progress in ML. We also broaden the scope by identifying external safety risks surrounding the deployment context of ML. The technical agenda of Taylor *et al.*, 2016 [184] considers similar topics to Amodei and Olah *et al.*, and Leike *et al.*, 2018 [119] considers safety research directions in reward modeling. Although Leike *et al.*'s research agenda focuses on reinforcement learning, they highlight the importance of various other research problems including adversarial training and uncertainty estimation. Recently, Critch and Krueger, 2020 [39] provide an extensive commentary on safety research directions and discuss safety when there are multiple stakeholders.

7   Conclusion

This work presented a non-exhaustive list of four unsolved research problems, all of which are interconnected and interdependent. Anomaly detection, for example, helps with detecting proxy gaming, detecting suspicious cyberactivity, and executing fail-safes in the face of unexpected events. Achieving safety requires research on all four problems, not just one. To see this, recall that a machine learning system that is not aligned with human values may be unsafe in and of itself, as it may create unintended consequences or game human approval proxies. Even if it is possible to create aligned objectives for ML systems, Black Swan events could cause ML systems to misgeneralize and pursue incorrect goals, malicious actors may launch adversarial attacks or compromise the software on which the ML system is running, and humans may need to monitor for emergent functionality and the malicious use of ML systems. As depicted in Figure 5's highly simplified model, work on all four problems helps create comprehensive and layered protective measures against a wide range of safety threats.

As machine learning research evolves, the community's aims and expectations should evolve too. For many years, the machine learning community focused on making machine learning systems work in the first place. However, machine learning systems have had notable success in domains from images, to natural language, to programming—therefore our focus should expand beyond just accuracy, speed, and scalability. Safety must now become a top priority.

Safety is not auxiliary in most current widely deployed technology. Communities do not ask for "safe bridges," but rather just "bridges." Their safety is insisted upon—even assumed—and incorporating safety features is imbued in the design process. The ML community should similarly create a culture of safety and elevate its standards so that ML systems can be deployed in safety-critical situations.

Acknowledgements

References

[1]   Signed by approximately 2000 AI researchers. "Asilomar AI Principles". In: (2017).

[2]   Signed by 30000+ people. "Autonomous Weapons: An Open Letter from AI and Robotics Researchers". In: (2015).

[3]   Martín Abadi, Andy Chu, I. Goodfellow, H. B. McMahan, Ilya Mironov, Kunal Talwar, and L. Zhang. "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (2016).

[4]   Zeeshan Ahmad, A. Khan, W. Cheah, J. Abdullah, and Farhan Ahmad. "Network intrusion detection system: A systematic study of machine learning and deep learning approaches". In: *Trans. Emerg. Telecommun. Technol.* (2021).

[5]   Dario Amodei, Christopher Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dandelion Mané. "Concrete Problems in AI Safety". In: *ArXiv* (2016).

[6]   Ross J. Anderson and Roger Needham. "Programming Satan's Computer". In: *Computer Science Today*. 1995.

[7]   Anish Athalye, Nicholas Carlini, and David A. Wagner. "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples". In: *ICML*. 2018.

[8]   Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc V. Le, and Charles Sutton. "Program Synthesis with Large Language Models". In: *ArXiv* (2021).

[9]   Eugene Bagdasaryan and Vitaly Shmatikov. "Blind Backdoors in Deep Learning Models". In: *USENIX Security Symposium*. 2021.

[10]   Abhijit Bendale and Terrance Boult. "Towards Open Set Deep Networks". In: *CVPR* (2016).

[11]   Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021).

[12]   Machine Learning at Berkeley. *Alien Dreams: An Emerging Art Scene*. URL: https://ml.berkeley.edu/blog/posts/clip-art/.

[13]   Victor Besnier, Andrei Bursuc, David Picard, and Alexandre Briot. "Triggering Failures: Out-Of-Distribution detection by learning from local adversarial attacks in Semantic Segmentation". In: *ArXiv* abs/2108.01634 (2021).

[14]   Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. "Evasion attacks against machine learning at test time". In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer. 2013, pp. 387–402.

[15]   Abeba Birhane, Pratyusha Kalluri, D. Card, William Agnew, Ravit Dotan, and Michelle Bao. "The Values Encoded in Machine Learning Research". In: *ArXiv* (2021).

[16]   Julian Bitterwolf, Alexander Meinke, and Matthias Hein. "Certifiably Adversarially Robust Detection of Out-of-Distribution Data". In: *NeurIPS* (2020).

[17]   Rishi Bommasani et al. "On the Opportunities and Risks of Foundation Models". In: *ArXiv* (2021).

[18]   Nick Bostrom. "The Vulnerable World Hypothesis". In: *Global Policy* (2019).

[19]   G. Botvin, C. Goldberg, E. M. Botvin, and L. Dusenbury. "Smoking behavior of adolescents exposed to cigarette advertising". In: *Public health reports* (1993).

[20]   Wieland Brendel, Jonas Rauber, and Matthias Bethge. "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models". In: *arXiv preprint arXiv:1712.04248* (2017).

[21]   Philip Brickman and Donald Campbell. "Hedonic relativism and planning the good society". In: 1971.

[22]   T. Brown, B. Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, G. Krüger, T. Henighan, R. Child, Aditya Ramesh, D. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, E. Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, J. Clark, Christopher Berner, Sam McCandlish, A. Radford, Ilya Sutskever, and Dario Amodei. "Language Models are Few-Shot Learners". In: *ArXiv* abs/2005.14165 (2020).

[23]   Miles Brundage, Shahar Avin, Jack Clark, H. Toner, P. Eckersley, Ben Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, Bobby Filar, H. Anderson, Heather Roff, Gregory C. Allen, J. Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, S. Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation". In: *ArXiv* abs/1802.07228 (2018).

[24]   Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian K. Hadfield, Heidy Khlaaf, Jingying Yang, H. Toner, Ruth Fong, Tegan Maharaj, P. W. Koh, Sara Hooker, J. Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, T. Ryffel, J. Rubinovitz, T. Besiroglu, F. Carugati, Jack Clark, P. Eckersley, Sarah de Haas, Maritza L. Johnson, B. Laurie, A. Ingerman, I. Krawczuk, Amanda Askell, Rosario Cammarota, A. Lohn, David Krueger, C. Stix, Peter Henderson, L. Graham, Carina E. A. Prunkl, Bianca Martin, E. Seger, Noa Zilberman, Se'an 'O h'Eigeartaigh, F. Kroeger, Girish Sastry, R. Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, A. Dafoe, P. Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, T. Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims". In: *ArXiv* (2020).

[25]   Erik Brynjolfsson and Adam Saunders. "What the GDP Gets Wrong (Why Managers Should Care)". In: *MIT Sloan Management Review* (2009).

[26]   Ben Buchanan, John Bansemer, Dakota Cary, Jack Lucas, and Micah Musser. "Automating Cyber Attacks". In: 2021.

[27]   Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. "Truth, Lies, and Automation". In: 2021.

[28]   Nicholas Carlini and A. Terzis. "Poisoning and Backdooring Contrastive Learning". In: *ArXiv* abs/2106.09667 (2021).

[29]   Nicholas Carlini and David Wagner. "Towards evaluating the robustness of neural networks". In: *2017 ieee symposium on security and privacy (sp)*. IEEE. 2017, pp. 39–57.

[30]   Y. Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C. Duchi. "Unlabeled Data Improves Adversarial Robustness". In: *NeurIPS*. 2019.

[31]   Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. "Emerging Properties in Self-Supervised Vision Transformers". In: *Proceedings of the International Conference on Computer Vision (ICCV)*. 2021.

[32]   Dakota Cary and Daniel Cebul. "Destructive Cyber Operations and Machine Learning". In: 2020.

[33] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde, J. Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea. Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, F. Such, D. Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William H. Guss, Alex Nichol, I. Babuschkin, S. Balaji, Shantanu Jain, A. Carr, J. Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, M. Knight, Miles Brundage, Mira Murati, Katie Mayer, P. Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. "Evaluating Large Language Models Trained on Code". In: *ArXiv* (2021).

[34] Steven Chen, Nicholas Carlini, and David A. Wagner. "Stateful Detection of Black-Box Adversarial Attacks". In: *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence* (2019).

[35] Jack Clark and Dario Amodei. "Faulty Reward Functions in the Wild". In: *OpenAI* (2016).

[36] Karl Cobbe, Oleg Klimov, Christopher Hesse, Taehoon Kim, and J. Schulman. "Quantifying Generalization in Reinforcement Learning". In: *ICML*. 2019.

[37] Jeremy M. Cohen, Elan Rosenfeld, and J. Z. Kolter. "Certified Adversarial Robustness via Randomized Smoothing". In: *ICML*. 2019.

[38] North American Aerospace Defense Command and U.S. Northern Command Public Affairs. 2021. URL: https://www.af.mil/News/Article-Display/Article/2703548/norad-usnorthcom-lead-3rd-global-information-dominance-experiment/.

[39] Andrew Critch and David Krueger. "AI Research Considerations for Human Existential Safety (ARCHES)". In: *ArXiv* (2020).

[40] Francesco Croce, Maksym Andriushchenko, V. Sehwag, Nicolas Flammarion, M. Chiang, Prateek Mittal, and Matthias Hein. "RobustBench: a standardized adversarial robustness benchmark". In: *ArXiv* abs/2010.09670 (2020).

[41] Maria Cvach. "Monitor alarm fatigue: an integrative review". In: *Biomedical instrumentation & technology* (2012).

[42] Allan Dafoe. "AI governance: a research agenda". In: *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK* (2018).

[43] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. "Open Problems in Cooperative AI". In: *ArXiv* (2020).

[44] Mohamad H. Danesh and Alan Fern. "Out-of-Distribution Dynamics Detection: RL-Relevant Benchmarks and Results". In: *ArXiv* abs/2107.04982 (2021).

[45] Department of Defense. "Quadrennial Defense Review Report". In: (2001).

[46] Laura DeNardis. "A history of internet security". In: *The history of information security*. Elsevier, 2007.

[47] Thomas G. Dietterich. "Robust artificial intelligence and robust human organizations". In: *Frontiers of Computer Science* (2018).

[48] Adrien Ecoffet and Joel Lehman. "Reinforcement Learning Under Moral Uncertainty". In: *ArXiv* abs/2006.04734 (2021).

[49] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, E. Hovy, Hinrich Schütze, and Yoav Goldberg. "Measuring and Improving Consistency in Pretrained Language Models". In: *ArXiv* (2021).

[50] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, D. Tsipras, J. Steinhardt, and A. Madry. "Identifying Statistical Bias in Dataset Replication". In: *ICML*. 2020.

[51] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. "A rotation and a translation suffice: Fooling cnns with simple transformations". In: *arXiv* (2018).

[52] Facebook. *Bringing People Closer Together*. URL: https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/.

[53] Pablo Fajnzylber, Daniel Lederman, and Norman V. Loayza. "Inequality and Violent Crime". In: *The Journal of Law and Economics* (2002).

[54] Wendi Folkert. "Assessment results regarding Organization Designation Authorization (ODA) Unit Member (UM) Independence". In: *Aviation Safety* (2021).

[55] F. R. Frola and C. O. Miller. "System Safety in Aircraft Acquisition". In: 1984.

[56] Iason Gabriel. "Artificial Intelligence, Values and Alignment". In: *ArXiv* (2020).

[57] John Gall. "Systemantics: How Systems Work and Especially How They Fail". In: 1977.

[58] Sneha Gathani, Madelon Hulsebos, James Gale, P. Haas, and çCaugatay Demiralp. "Augmenting Decision Making via Interactive What-If Analysis". In: 2021.

[59] Helmut Geist and Eric Lambin. "What drives tropical deforestation?: a meta-analysis of proximate and underlying causes of deforestation based on subnational case study evidence". In: 2001.

[60] Yolanda Gil and Bart Selman. "A 20-Year Community Roadmap for Artificial Intelligence Research in the US". In: *ArXiv* abs/1908.02624 (2019).

[61] J. Gilmer, Ryan P. Adams, I. Goodfellow, David G. Andersen, and George E. Dahl. "Motivating the Rules of the Game for Adversarial Example Research". In: *ArXiv* (2018).

[62] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana Kagal. "Explaining Explanations: An Overview of Interpretability of Machine Learning". In: (2018).

[63] Adam Gleave, Michael Dennis, Neel Kant, Cody Wild, Sergey Levine, and Stuart J. Russell. "Adversarial Policies: Attacking Deep Reinforcement Learning". In: *ICLR* (2020).

[64] Charles Goodhart. "Problems of Monetary Management: The UK Experience". In: 1984.

[65] Jeremy Greenwood. *The third industrial revolution: Technology, productivity, and income inequality*. 435. American Enterprise Institute, 1997.

[66] Nathan Grinsztajn, Johan Ferret, O. Pietquin, P. Preux, and M. Geist. "There Is No Turning Back: A Self-Supervised Approach for Reversibility-Aware Reinforcement Learning". In: *ArXiv* abs/2106.04480 (2021).

[67] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. "Badnets: Identifying vulnerabilities in the machine learning model supply chain". In: *arXiv preprint arXiv:1708.06733* (2017).

[68] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On Calibration of Modern Neural Networks". In: *ICML* (2017).

[69] Dylan Hadfield-Menell, A. Dragan, P. Abbeel, and Stuart J. Russell. "The Off-Switch Game". In: *IJCA* (2017).

[70] Dylan Hadfield-Menell, Stuart J. Russell, P. Abbeel, and A. Dragan. "Cooperative Inverse Reinforcement Learning". In: *NIPS*. 2016.

[71] Richard Harang and Ethan M. Rudd. *SOREL-20M: A Large Scale Benchmark Dataset for Malicious PE Detection*. 2020.

[72] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *NIPS*. 2016.

[73] P. J. Heawood. "Map-Colour Theorem". In: *Proceedings of The London Mathematical Society* (1949), pp. 161–175.

[74] James Hedlund. "Risky business: safety regulations, risk compensation, and individual behavior". In: *Injury Prevention* (2000).

[75] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. "The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization". In: *ICCV* (2021).

[76] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. "Aligning AI With Shared Human Values". In: *ICLR* (2021).

[77] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. "Measuring Massive Multitask Language Understanding". In: *ICLR* (2021).

[78] Dan Hendrycks and Thomas Dietterich. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations". In: *Proceedings of the International Conference on Learning Representations* (2019).

[79] Dan Hendrycks and Kevin Gimpel. "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks". In: *ICLR* (2017).

[80] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. "Using Pre-Training Can Improve Model Robustness and Uncertainty". In: *ICML*. 2019.

[81] Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. "Deep Anomaly Detection with Outlier Exposure". In: *ICLR* (2019).

[82] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and D. Song. "Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty". In: *NeurIPS*. 2019.

[83] Dan Hendrycks, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, and Jacob Steinhardt. "What Would Jiminy Cricket Do? Towards Agents That Behave Morally". In: *arXiv* (2021).

[84] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. "AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty". In: *ICLR* (2020).

[85] Dan Hendrycks, Kevin Zhao, Steven Basart, J. Steinhardt, and D. Song. "Natural Adversarial Examples". In: *CVPR* (2021).

[86] T. Henighan, J. Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, A. Ramesh, Nick Ryder, Daniel M. Ziegler, J. Schulman, Dario Amodei, and Sam McCandlish. "Scaling Laws for Autoregressive Generative Modeling". In: *ArXiv* abs/2010.14701 (2020).

[87] J. Hestness, Sharan Narang, Newsha Ardalani, G. Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Y. Yang, and Yanqi Zhou. "Deep Learning Scaling is Predictable, Empirically". In: *ArXiv* (2017).

[88] Grant Ho, Asaf Cidon, Lior Gavish, Marco Schweighauser, V. Paxson, S. Savage, G. Voelker, and David A. Wagner. "Detecting and Characterizing Lateral Phishing at Scale". In: *USENIX Security Symposium*. 2019.

[89] Sanghyun Hong, Nicholas Carlini, and A. Kurakin. "Handcrafted Backdoors in Deep Neural Networks". In: *ArXiv* (2021).

[90] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. "Risks from Learned Optimization in Advanced Machine Learning Systems". In: *ArXiv* (2019).

[91] David Hume. *A Treatise of Human Nature*. 1739.

[92] Geoffrey Irving, Paul Christiano, and Dario Amodei. "AI safety via debate". In: *ArXiv* (2018).

[93] Michael C Jensen and William H Meckling. "Theory of the firm: Managerial behavior, agency costs and ownership structure". In: *Journal of financial economics* 3.4 (1976), pp. 305–360.

[94] Woojeong Jin, Suji Kim, Rahul Khanna, Dong-Ho Lee, Fred Morstatter, A. Galstyan, and Xiang Ren. "ForecastQA: A Question Answering Challenge for Event Forecasting with Temporal Text Data". In: *ACL/IJCNLP*. 2021.

[95] Daniel Kahneman and Angus Deaton. "High income improves evaluation of life but not emotional well-being". In: *Proceedings of the National Academy of Sciences* (2010).

[96] Daniel Kang, Yi Sun, Dan Hendrycks, Tom B. Brown, and J. Steinhardt. "Testing Robustness Against Unforeseen Adversaries". In: *ArXiv* (2019).

[97] Kiran Karra, C. Ashcraft, and Neil Fendley. "The TrojAI Software Framework: An OpenSource tool for Embedding Trojans into Deep Learning Models". In: *ArXiv* (2020).

[98] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. "Alignment of Language Agents". In: *ArXiv* (2021).

[99] A. Kirilenko, Mehrdad Samadi, A. Kyle, and Tugkan Tuzun. "The Flash Crash: The Impact of High Frequency Trading on an Electronic Market". In: 2011.

[100] Jack Koch, L. Langosco, J. Pfau, James Le, and Lee Sharkey. "Objective Robustness in Deep Reinforcement Learning". In: *ArXiv* (2021).

[101] P. W. Koh, Shiori Sagawa, H. Marklund, Sang Michael Xie, Marvin Zhang, A. Balsubramani, Wei hua Hu, Michihiro Yasunaga, Richard L. Phillips, Sara Beery, J. Leskovec, A. Kundaje, E. Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. "WILDS: A Benchmark of in-the-Wild Distribution Shifts". In: *ICML*. 2021.

[102] Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and S. Legg. "Avoiding Side Effects By Considering Future Tasks". In: *NeurIPS* (2020).

[103] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, N. Keskar, Shafiq R. Joty, R. Socher, and Nazneen Rajani. "GeDi: Generative Discriminator Guided Sequence Generation". In: *ArXiv* (2020).

[104] Ethan Kross, Philippe Verduyn, Emre Demiralp, Jiyoung Park, David Seungjae Lee, Natalie Lin, Holly Shablack, John Jonides, and Oscar Ybarra. "Facebook use predicts declines in subjective well-being in young adults". In: *PloS one* ().

[105] David Krueger, Tegan Maharaj, and J. Leike. "Hidden Incentives for Auto-Induced Distributional Shift". In: *ArXiv* abs/2009.09153 (2020).

[106] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. "Accurate uncertainties for deep learning using calibrated regression". In: *ICML* (2018). arXiv: `1807.00263`.

[107] Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. "Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration". In: *NeurIPS*. 2019.

[108] Ananya Kumar, Percy Liang, and Tengyu Ma. "Verified Uncertainty Calibration". In: *NeurIPS*. 2019.

[109] Patrick Ky. "Boeing 737 MAX Return to Service Report". In: (2021).

[110] Marie-Anne Lachaux, Baptiste Rozière, L. Chanussot, and Guillaume Lample. "Unsupervised Translation of Programming Languages". In: *ArXiv* (2020).

[111] Cassidy Laidlaw, Sahil Singla, and S. Feizi. "Perceptual Adversarial Robustness: Defense Against Unseen Threat Models". In: *ICLR* (2021).

[112] Balaji Lakshminarayanan, A. Pritzel, and C. Blundell. "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles". In: *NIPS*. 2017.

[113] Terran Lane and Carla E Brodley. "An application of machine learning to anomaly detection". In: *Proceedings of the 20th National Information Systems Security Conference*. Vol. 377. Baltimore, USA. 1997, pp. 366–380.

[114] Ralph Langner. "Stuxnet: Dissecting a Cyberwarfare Weapon". In: *IEEE Security & Privacy* (2011).

[115] Katarzyna de Lazari-Radek and Peter Singer. "The Point of View of the Universe: Sidgwick and Contemporary Ethics". In: 2014.

[116] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. "Certified robustness to adversarial examples with differential privacy". In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2019, pp. 656–672.

[117] Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. "Hallucinations in neural machine translation". In: (2018).

[118] J. Lehman, J. Clune, D. Misevic, C. Adami, L. Altenberg, Julie Beaulieu, P. Bentley, Samuel Bernard, G. Beslon, David M. Bryson, P. Chrabaszcz, Nick Cheney, Antoine Cully, S. Doncieux, F. Dyer, Kai Olav Ellefsen, R. Feldt, Stephan Fischer, S. Forrest, Antoine Frénoy, Christian Gagné, L. K. L. Goff, L. Grabowski, B. Hodjat, F. Hutter, L. Keller, C. Knibbe, Peter Krcah, R. Lenski, H. Lipson, R. MacCurdy, Carlos Maestre, R. Miikkulainen, S. Mitri, David E. Moriarty, J. Mouret, Anh M Nguyen, C. Ofria, M. Parizeau, D. Parsons, Robert T. Pennock, W. Punch, T. Ray, Marc Schoenauer, E. Shulte, K. Sims, Kenneth O. Stanley, F. Taddei, Danesh Tarapore, S. Thibault, Westley Weimer, R. Watson, and Jason Yosinksi. "The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities". In: *Artificial Life* (2018).

[119] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. "Scalable agent alignment via reward modeling: a research direction". In: *ArXiv* (2018).

[120] Nancy Leveson. "Engineering a Safer World: Systems Thinking Applied to Safety". In: 2012.

[121] Beishui Liao, Marija Slavkovik, and Leendert van der Torre. "Building Jiminy Cricket: An Architecture for Moral Agreements Among Stakeholders". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019).

[122] Stephanie Lin, Jacob Hilton, and Owain Evans. "TruthfulQA: Measuring How Models Mimic Human Falsehoods". In: *arXiv* (2021).

[123] Rachel Luo, Aadyot Bhatnagar, Huan Wang, Caiming Xiong, Silvio Savarese, Yu Bai, Shengjia Zhao, and Stefano Ermon. "Localized Calibration: Metrics and Recalibration". In: *arXiv* (2021).

[124] Benoit Mandelbrot and Richard L. Hudson. "The Misbehavior of Markets: A Fractal View of Risk, Ruin, and Reward". In: 2004.

[125] David Manheim and Scott Garrabrant. "Categorizing Variants of Goodhart's Law". In: *ArXiv* (2018).

[126] Filip Maric and Sana Stojanovic-Durdevic. "Formalizing IMO Problems and Solutions in Isabelle/HOL". In: *ThEdu@IJCAR*. 2020.

[127] Microsoft. URL: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

[128] Michael Mitzenmacher. "A Brief History of Generative Models for Power Law and Lognormal Distributions". In: *Internet Mathematics* (2003).

[129] Chaithanya Kumar Mummadi, Robin Hutmacher, K. Rambach, Evgeny Levinkov, T. Brox, and J. H. Metzen. "Test-Time Adaptation to Distribution Shift by Confidence Maximization and Input Transformation". In: *ArXiv* (2021).

[130] Toby Newberry and Toby Ord. "The Parliamentary Approach to Moral Uncertainty". In: 2021.

[131] Khanh Nguyen and Brendan T. O'Connor. "Posterior calibration and exploratory analysis for natural language processing models". In: *EMNLP*. 2015.

[132] NSA. URL: https://ghidra-sre.org/.

[133] Martha Nussbaum. "CAPABILITIES AS FUNDAMENTAL ENTITLEMENTS: SEN AND SOCIAL JUSTICE". In: *Feminist Economics* 9 (2003), pp. 33 –59.

[134] Rain Ottis. "Analysis of the 2007 Cyber Attacks Against Estonia from the Information Warfare Perspective". In: 2008.

[135] Yaniv Ovadia, E. Fertig, J. Ren, Zachary Nado, D. Sculley, S. Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift". In: *NeurIPS*. 2019.

[136] Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. "An Empirical Cybersecurity Evaluation of GitHub Copilot's Code Contributions". In: *ArXiv* (2021).

[137] Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan L. Boyd-Graber. "It Takes Two to Lie: One to Lie, and One to Listen". In: *ACL*. 2020.

[138] Fabio Pierazzi, Feargus Pendlebury, Jacopo Cortellazzi, and Lorenzo Cavallaro. "Intriguing properties of adversarial ml attacks in the problem space". In: *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2020, pp. 1332–1349.

[139] Richard A. Posner. "Utilitarianism, Economics, and Legal Theory". In: *The Journal of Legal Studies* (1979).

[140] Omid Poursaeed, Tianxing Jiang, Harry Yang, Serge Belongie, and Ser-Nam Lim. "Robustness and Generalization via Generative Adversarial Training". In: 2021.

[141] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. "Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets". In: *ICLR MATH-AI Workshop*. 2021.

[142] "Principle 1: Preoccupation with Failure". In: *Managing the Unexpected*. John Wiley & Sons, Ltd, 2015. Chap. 3, pp. 45–61. ISBN: 9781119175834. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119175834.ch03.

[143] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. "Learning Transferable Visual Models From Natural Language Supervision". In: *ICML*. 2021.

[144] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. "Certified Defenses against Adversarial Examples". In: *ICLR* (2018).

[145] Pranav Rajpurkar, Jeremy A. Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, T. Duan, D. Ding, Aarti Bagul, C. Langlotz, K. Shpanskaya, M. Lungren, and A. Ng. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning". In: *ArXiv* (2017).

[146] Theodore D. Raphael. "Integrative Complexity Theory and Forecasting International Crises: Berlin 1946-1962". In: *The Journal of Conflict Resolution* (1982).

[147] John Rawls. *A Theory of Justice*. Harvard University Press, 1999.

[148] Alex Ray, Joshua Achiam, and Dario Amodei. "Benchmarking Safe Exploration in Deep Reinforcement Learning". In: 2019.

[149] Sylvestre-Alvise Rebuffi, Sven Gowal, D. A. Calian, Florian Stimberg, Olivia Wiles, and Timothy A. Mann. "Fixing Data Augmentation to Improve Adversarial Robustness". In: *ArXiv* abs/2103.01946 (2021).

[150] V. Ridgway. "Dysfunctional Consequences of Performance Measurements". In: *Administrative Science Quarterly* (1956).

[151] Stuart Russell, Anthony Aguirre, Emilia Javorsky, and Max Tegmark. "Lethal Autonomous Weapons Exist; They Must Be Banned". In: (2021).

[152] Stuart J. Russell, Daniel Dewey, and Max Tegmark. "Research Priorities for Robust and Beneficial Artificial Intelligence". In: *AI Magazine* (2015).

[153] William Saunders, Girish Sastry, Andreas Stuhlmüller, and Owain Evans. "Trial without Error: Towards Safe Reinforcement Learning via Human Intervention". In: *AAMAS*. 2018.

[154] Roei Schuster, Congzheng Song, Eran Tromer, and Vitaly Shmatikov. "You Autocomplete Me: Poisoning Vulnerabilities in Neural Code Completion". In: *USENIX* (2021).

[155] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. "Hidden technical debt in machine learning systems". In: *Advances in neural information processing systems* 28 (2015), pp. 2503–2511.

[156] A. Shafahi, W. R. Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, T. Dumitras, and T. Goldstein. "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks". In: *NeurIPS*. 2018.

[157] Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, P. Abbeel, and A. Dragan. "Preferences Implicit in the State of the World". In: *ICLR* (2019).

[158] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition". In: *Proceedings of the 2016 acm sigsac conference on computer and communications security*. 2016, pp. 1528–1540.

[159] Dongdong She, Yizheng Chen, Baishakhi Ray, and S. Jana. "Neutaint: Efficient Dynamic Taint Analysis with Neural Networks". In: *2020 IEEE Symposium on Security and Privacy (SP)* (2020), pp. 1527–1543.

[160] Dongdong She, Kexin Pei, D. Epstein, Junfeng Yang, Baishakhi Ray, and S. Jana. "NEUZZ: Efficient Fuzzing with Neural Program Smoothing". In: *2019 IEEE Symposium on Security and Privacy (SP)* (2019), pp. 803–817.

[161] E. C. Shin, D. Song, and R. Moazzezi. "Recognizing Functions in Binaries with Neural Networks". In: *USENIX Security Symposium*. 2015.

[162] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. "Membership inference attacks against machine learning models". In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 3–18.

[163] M. D.Amran Siddiqui, Alan Fern, Thomas G. Dietterich, and Weng Keen Wong. "Sequential feature explanations for anomaly detection". In: *ACM Transactions on Knowledge Discovery from Data* (2019).

[164] Henry Sidgwick. *The Methods of Ethics*. 1907.

[165] Robin Sommer and Vern Paxson. "Outside the closed world: On using machine learning for network intrusion detection". In: *2010 IEEE symposium on security and privacy*. IEEE. 2010, pp. 305–316.

[166] D. H. Stamatis. "Failure mode and effect analysis : FMEA from theory to execution". In: *ASQC Quality Press* (1996).

[167] Keith E. Stanovich, Richard F. West, and Maggie E. Toplak. "The Rationality Quotient: Toward a Test of Rational Thinking". In: 2016.

[168] Nick Carr Steve Miller Evan Reese. *Shikata Ga Nai Encoder Still Going Strong*. URL: https://www.fireeye.com/blog/threat-research/2019/10/shikata-ga-nai-encoder-still-going-strong.html.

[169] Marilyn Strathern. "'Improving ratings': audit in the British University system". In: *European Review* (1997).

[170] Jonathan Stray. "Aligning AI Optimization to Community Well-Being". In: *International Journal of Community Well-Being* (2020).

[171] Jonathan Stray, Ivan Vendrov, Jeremy Nixon, Steven Adler, and Dylan Hadfield-Menell. "What are you optimizing for? Aligning Recommender Systems with Human Values". In: *ArXiv* abs/2107.10939 (2021).

[172] David Stutz, Matthias Hein, and B. Schiele. "Confidence-Calibrated Adversarial Training: Generalizing to Unseen Attacks". In: *ICML*. 2020.

[173] Octavian Suciu, Scott E. Coull, and Jeffrey Johns. "Exploring Adversarial Examples in Malware Detection". In: *IEEE Security and Privacy Workshops (SPW)* (2019).

[174] RL Sumwalt, B Landsberg, and J Homendy. "Assumptions used in the safety assessment process and the effects of multiple alerts and indications on pilot performance". In: *District of Columbia: National Transportation Safety Board* (2019).

[175] Rebecca Sutton. *Chromium-6 in US tap water*. Environmental Working Group Washington, DC, 2010.

[176] Publius Syrus. *The Moral Sayings of Publius Syrus, a Roman Slave*. L.E. Bernard & Company, 1856.

[177] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199* (2013).

[178] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. "CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances". In: *NeurIPS* (2020).

[179] Jihoon Tack, Sihyun Yu, Jongheon Jeong, Minseong Kim, Sung Ju Hwang, and Jinwoo Shin. "Consistency Regularization for Adversarial Robustness". In: *ArXiv* (2021).

[180] Nassim Taleb. "Antifragile: Things That Gain from Disorder". In: 2012.

[181] Nassim Taleb. "Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications". In: 2020.

[182] Nassim Taleb. "The Black Swan: The Impact of the Highly Improbable". In: 2007.

[183] Nassim Taleb and Philip Tetlock. "On the Difference between Binary Prediction and True Exposure with Implications for Forecasting Tournaments and Decision Making Research". In: 2013.

[184] Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. "Alignment for Advanced Machine Learning Systems". In: 2016.

[185] Tesla. *Tesla AI Day*. 2021. URL: https://www.youtube.com/watch?v=j0z4FweCy4M.

[186] Philip Tetlock and Dan Gardner. "Superforecasting: The Art and Science of Prediction". In: 2015.

[187] Brandon Trabucco, Aviral Kumar, Xinyang Geng, and Sergey Levine. "Conservative Objective Models for Effective Offline Model-Based Optimization". In: *ICML*. 2021.

[188] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and A. Madry. "On Adaptive Attacks to Adversarial Example Defenses". In: *ArXiv* (2020).

[189] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick Mcdaniel. "Ensemble Adversarial Training: Attacks and Defenses". In: *ArXiv* abs/1705.07204 (2018).

[190] A. M. Turner, Neale Ratzlaff, and Prasad Tadepalli. "Avoiding Side Effects in Complex Environments". In: *ArXiv* abs/2006.06547 (2020).

[191] Building Seismic Safety Council US et al. "Planning for seismic rehabilitation: societal issues". In: (1998).

[192] Carroll L. Wainwright and P. Eckersley. "SafeLife 1.0: Exploring Side Effects in Complex Environments". In: *ArXiv* abs/1912.01217 (2020).

[193] Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. "Concealed Data Poisoning Attacks on NLP Models". In: *NAACL*. 2021.

[194] Dequan Wang, An Ju, Evan Shelhamer, David A. Wagner, and Trevor Darrell. "Fighting Gradients with Gradients: Dynamic Defenses against Adversarial Attacks". In: *ArXiv* abs/2105.08714 (2021).

[195] Dequan Wang, Evan Shelhamer, Shaoteng Liu, B. Olshausen, and Trevor Darrell. "Tent: Fully Test-Time Adaptation by Entropy Minimization". In: *ICLR*. 2021.

[196] Pei Wang, Yijun Li, Krishna Kumar Singh, Jingwan Lu, and N. Vasconcelos. "IMAGINE: Image Synthesis by Image-Guided Model Inversion". In: *ArXiv* abs/2104.05895 (2021).

[197] Yue Wang, Esha Sarkar, Wenqing Li, M. Maniatakos, and S. E. Jabari. "Stop-and-Go: Exploring Backdoor Attacks on Deep Reinforcement Learning-based Traffic Congestion Control Systems". In: *arXiv: Cryptography and Security* (2020).

[198] Yunchao Wang, Zehui Wu, Qiang Wei, and Qingxian Wang. "NeuFuzz: Efficient Fuzzing With Deep Neural Network". In: *IEEE Access* (2019).

[199] E. G. Williams. "The Possibility of an Ongoing Moral Catastrophe". In: *Ethical Theory and Moral Practice* (2015).

[200] Timothy Wilson and Daniel Gilbert. "Affective Forecasting". In: *Current Directions in Psychological Science* (2005).

[201] Dongxian Wu, Shutao Xia, and Yisen Wang. "Adversarial Weight Perturbation Helps Robust Generalization". In: *NeurIPS* (2020).

[202] Chaowei Xiao, Ruizhi Deng, Bo Li, F. Yu, M. Liu, and D. Song. "Characterizing Adversarial Examples Based on Spatial Consistency Information for Semantic Segmentation". In: *ECCV*. 2018.

[203] Cihang Xie, Mingxing Tan, Boqing Gong, A. Yuille, and Quoc V. Le. "Smooth Adversarial Training". In: *ArXiv* abs/2006.14536 (2020).

[204] Hongxu Yin, Pavlo Molchanov, Zhizhong Li, J. Álvarez, Arun Mallya, Derek Hoiem, N. Jha, and J. Kautz. "Dreaming to Distill: Data-Free Knowledge Transfer via DeepInversion". In: *CVPR* (2020).

[205] Sheheryar Zaidi, Arber Zela, T. Elsken, Chris C. Holmes, F. Hutter, and Y. Teh. "Neural Ensemble Search for Uncertainty Estimation and Dataset Shift". In: 2020.

[206] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael I. Jordan. "Theoretically Principled Trade-off between Robustness and Accuracy". In: *ICML*. 2019.

[207] Xinyang Zhang, Zheng Zhang, and Tianying Wang. "Trojaning Language Models for Fun and Profit". In: *ArXiv* abs/2008.00312 (2020).

[208] Yao Zhu, Jiacheng Ma, Jiacheng Sun, Zewei Chen, Rongxin Jiang, and Zhenguo Li. "Towards Understanding the Generative Capability of Adversarially Robust Classifiers". In: *ArXiv* (2021).

[209] Shoshana Zuboff. "The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power". In: (2019).

[210] Remco Zwetsloot, Helen Toner, and Jeffrey Ding. "Beyond the AI arms race: America, China, and the dangers of zero-sum thinking". In: *Foreign Affairs* (2018).

A Analyzing Risks, Hazards, and Impact

A.1 Risk Management Framework

Area	Problem	ML System Risks	Operational Risks	Institutional and Societal Risks	Future Risks
Robustness	Black Swans and Tail Risks	✓	✓		✓
	Adversarial Robustness	✓			✓
Monitoring	Anomaly Detection	✓	✓		✓
	Representative Outputs	✓	✓		✓
	Hidden Model Functionality	✓	✓		✓
Alignment	Value Learning		✓	✓	✓
	Translating Values to Action	✓			✓
	Proxy Gaming	✓			✓
	Value Clarification			✓	✓
	Unintended Consequences	✓	✓	✓	✓
External Safety	ML for Cybersecurity	✓	✓	✓	✓
	Informed Decision Making		✓	✓	✓

Table 1: Problems and the risks they directly mitigate. Each checkmark indicates whether a problem directly reduces a risk. Notice that problems affect both near- and long-term risks.

To analyze how ML Safety progress can reduce abstract risks and hazards,[1] we identify four dimensions of risk in this section and five hazards in the next section.

The following four risk dimensions are adopted from the Department of Defense's broad risk management framework [45], with its personnel management risks replaced with ML system risks.

1. **ML System Risks** – risks to the ability of a near-term individual ML system to operate reliably.

2. **Operational Risks** – risks to the ability of an organization to safely operate an ML system in near-term deployment scenarios.

3. **Institutional and Societal Risks** – risks to the ability of global society or institutions that decisively affect ML systems to operate in near-term scenarios in an efficient, informed, and prudent way.

4. **Future (ML System, Operational, and Institutional) Risks** – risks to the ability of future ML systems, organizations operating ML systems, and institutions to address mid- to long-term challenges.

In Table 1, we indicate whether one of these risks is reduced by progress on a given ML Safety problem. Note that these all problems reduce risks to all three of future ML systems, organizations, and institutions. In the future, organizations and institutions will likely become more dependent on ML systems, so improvements to Black Swans robustness would in the future help improve operations and institutions dependent on ML systems. Since this table is a snapshot of the present, risk profiles will inevitably change.

---

[1]One can think of hazards as factors that have the potential to cause harm. One can think of risk as the amount of exposure to the hazard multiplied by the hazard's deleterious effect. For example, a wet floor is a hazard to humans. However, risks from wet floors are lower if humans heed wet floor signs and have less exposure to them. Risks are also lower for young adults than the elderly, since the elderly are more physically vulnerable. In other terms, robustness makes systems less vulnerable to hazards, monitoring reduces exposure to hazards, and alignment makes systems avoid creating hazards.

A.2   Hazard Management Framework

Area	Problem	Known Unknowns	Unknown Unknowns	Emergence	Long Tails	Adversaries & Deception
Robustness	Black Swans and Tail Risks	✔	✔	✓	✔	
	Adversarial Robustness	✓				✔
Monitoring	Anomaly Detection	✓	✔	✔	✔	✔
	Representative Outputs	✔	✓			✔
	Hidden Model Functionality	✔	✔	✔		✔
Alignment	Value Learning	✔				
	Translating Values to Action	✔			✓	
	Proxy Gaming	✓				✔
	Value Clarification	✔	✔	✓	✓	
	Unintended Consequences		✔	✔		
External Safety	ML for Cybersecurity	✓	✓			✔
	Informed Decision Making	✔	✔	✓	✓	

Table 2: Problems and the hazards they help handle. Checkmarks indicate whether a problem directly reduces vulnerability or exposure to a given hazard, and bold green checkmarks indicate an especially notable reduction.

We now turn from what is affected by risks to five abstract hazards that create risks.

1. **Known Unknowns** – Identified hazards for which we have imperfect or incomplete knowledge. These are identified hazards known to have unknown aspects.

2. **Unknown Unknowns** – Hazards which are unknown and unidentified, and they have properties that are unknown.

3. **Emergence** – A hazard that forms and comes into being as the system increases in size or its parts are combined. Such hazards do not exist in smaller versions of the system nor in its constituent parts.

4. **Long Tails** – Hazards that can be understood as unusual or extreme events from a long tail distribution.

5. **Adversaries & Deception** – Hazards from a person, system, or force that aims to attack, subvert, or deceive.

These hazards do not enumerate all possible hazards. For example, the problems in External Safety help with turbulence hazards. Furthermore, feedback loops, which can create long tails, could become a more prominent hazard in the future when ML systems are integrated into more aspects of our lives.

The five hazards have some overlap. For instance, when something novel emerges, it is an unknown unknown. When it is detected, it can become a known unknown. Separately, long tail events are often but not necessarily unknown unknowns: the 1987 stock market crash was a long tail event, but it was a known unknown to a prescient few and an unknown unknown to most everybody else. Emergent hazards sometimes co-occur with long tailed events, and an adversarial attack can cause long tail events.

In Table 2, we indicate whether an ML Safety problem reduces vulnerability or exposure to a given hazard. As with Table 1, the table is a snapshot of the present. For example, future adversaries could create novel unusual events or strike during tail events, so Black Swan robustness could improve adversarial robustness.

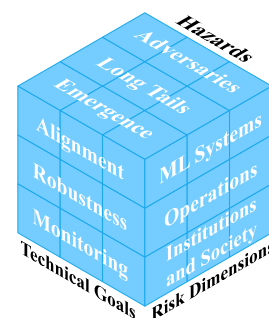With risks, hazards, and goals now all explicated, we depict their interconnectedness in Figure 6.

Figure 6: A simplified model of interconnected factors for ML Safety.

A.3 Prioritization and Strategy for Maximizing Impact

Area	Problem	Importance	Neglectedness	Tractability
Robustness	Black Swans and Tail Risks	● ●	● ●	● ●
	Adversarial Robustness	● ●	●	● ●
Monitoring	Anomaly Detection	● ● ●	● ●	● ● ●
	Representative Outputs	● ● ●	● ●	● ●
	Hidden Model Functionality	● ● ●	● ●	● ●
Alignment	Value Learning	● ● ●	● ●	● ●
	Translating Values to Action	● ●	● ●	● ● ●
	Proxy Gaming	● ● ●	● ●	● ●
	Value Clarification	● ●	● ● ●	●
	Unintended Consequences	● ●	● ● ●	●
External Safety	ML for Cybersecurity	● ●	● ● ●	● ● ●
	Informed Decision Making	● ●	● ●	● ●

Table 3: Problems and three factors that influence expected marginal impact.

We presented several problems, but new researchers may be able to make a larger impact on some problems over others. Some problems may be important, but if they are extremely popular, the risk of scooping increases, as does the risk of researchers stepping on each others' toes. Likewise, some problems may be important and may be decisive for safety if solved, but some problems are simply infeasible. Consequently, we should consider the importance, neglectedness, and tractability of problems.

1. **Importance** – How much potential risk does substantial progress on this problem reduce?

    - ● Progress on this problem reduces risks of catastrophes.
    - ● ● Progress on this problem directly reduces risks from potential permanent catastrophes.
    - ● ● ● Progress on this problem directly reduces risks from more plausible permanent catastrophes.

2. **Neglectedness** – How much research is being done on the problem?

    - ● The problem is one of the top ten most researched topics at leading conferences.
    - ● ● The problem receives some attention at leading ML conferences, or adjacent problems are hardly neglected.
    - ● ● ● The problem has few related papers consistently published at leading ML conferences.

3. **Tractability** – How much progress can we expect on the problem?

    - ● We cannot expect large research efforts to highly fruitful currently, possibly due to conceptual bottlenecks, or productive work on the problem likely requires far more advanced ML capabilities.
    - ● ● We expect to reliably and continually make progress on the problem.
    - ● ● ● A large research effort would be highly fruitful and there is obvious low-hanging fruit.

A snapshot of each problem and its current importance, neglectedness, and tractability is in Table 3. Note this only provides a rough sketch, and it has limitations. For example, a problem that is hardly neglected overall may still have neglected aspects; while adversarial robustness is less neglected than other safety problems, robustness to unforeseen adversaries is fairly neglected. Moreover, working on popular shovel-ready problems

may be more useful for newcomers compared to working on problems where conceptual bottlenecks persist. Further, this gives a rough sense of marginal impact, but entire community should not chose to act in the same way marginally, or else neglected problems will suddenly become overcrowded.

These three factors are merely prioritization factors and do not define a strategy. Rather, a potential strategy for ML Safety is as follows. 1. Force Management: Cultivate and maintain a force of ready personnel to implement safety measures into advanced ML systems and operate ML systems safely. 2. Research: Build and maintain a community to conduct safety research, including the identification of potential future hazards, clarification of safety goals, reduction of the costs to adopt safety methods, incorporation of safety methods that are built into existing ML systems, and so on. 3. Protocols: Establish and incentivize adherence to protocols, precedents, and standards for the safe development and deployment of ML systems. 4. Partnerships: Build and maintain safety-focused alliances and partnerships among academe, industry, and government.

In closing, throughout ML Safety's development we have seen numerous proposed strategies, hazards, risks, scenarios, and problems. In safety, some previously proposed problems have been discarded, and some new problems have emerged, just as in the broader ML community. Since no individual knows what lies ahead, safety analysis and strategy will need to evolve and adapt beyond this document. Regardless of which particular safety problems turn out to be the most or least essential, the success of safety's evolution and adaptation rests on having a large and capable research community.