International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 03 | Mar 2021 www.irjet.net p-ISSN: 2395-0072

AN EMPIRICAL MODEL FOR VALIDITY AND VERIFICATION OF AI BEHAVIOR: OVERCOMING AI HAZARDS IN NEURAL NETWORKS:

Ayse K. Arslan

[1]*Association of Oxford Alumni, Northern California, USA*

Abstract - *Rapid progress in machine learning and artificial intelligence (AI) has brought increasing attention to the potential impacts of AI technologies on society. This paper discusses hazards in machine learning systems, defined as unintended and harmful behavior that may emerge from poor design of real-world AI systems with a particular focus on ANN. The paper provides a review of previous work in these areas as well as suggesting research directions with a focus on relevance to cutting-edge AI systems with a focus on neural networks. Finally, the paper considers the high-level question of how to think most productively about the safety of forward-looking applications of AI.*

Key Words: AI, machine learning, research, algorithms, neural networks, software development, engineering

1. INTRODUCTION

There is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase. The last few years have seen rapid progress on long-standing, difficult problems in machine learning (ML) and artificial intelligence (AI), in diverse areas which brought excitement about the positive potential for AI to transform medicine [12], science [9], and transportation [6], along with concerns about the privacy [7], security [1], fairness [3], economic [32], and military [16] implications of autonomous systems, as well as concerns about the longer-term implications of powerful AI [27, 17].

The aim of this paper is to catalogue some of the various possible ways in which AI, especially within the context of ANN (artificial neural networks) can cause harm. The aim is not to determine how common and serious these harms are or how they stack up against the many benefits of information—questions that would need to be engaged before one could reach a considered position about potential policy implications, yet rather to enlighten the reader on potential threats caused by this technology.

2. EXISTING WORK

Artificial Intelligence (AI) refers to the art of creating machines that are able to think and act like humans; or think and act reasonably [4, 7]. In order to build an agent that can think and act as so, the agent must be able to learn new things. To learn means that the agent should improve its performance on future tasks taking its past experience into account [20, 7]. Making an agent able to learn is an area of study called Machine Learning (ML).

Artificial Neural Network or ANN is a software structure developed and based on concepts inspired by biological functions of brain; it aims at creating machines able to learn like a human-being [2, 7]. Thus, ANN is part of ML. Interestingly, ANN has many other names in AI field including parallel distributed processing, neural computation and connectionism [12, 11]. Most ANN types are supervised learning network. That is, both an input and the correct output should be given to a network where the network should learn a function that maps inputs to outputs.

Artificial Neuron

Since a structure of ANN has been inspired by biological brain, ANN should consist of a collection of neurons. AI researchers designed artificial neurons called perceptron and sigmoid which are believed to have similar function to a biological neuron [27, 17]. Artificial neuron is hereafter referred to as neuron for short. A neuron is a node that receives input from preceding neurons and makes a decision to 'fire' to the next neurons. To make that decision, it should first evaluate each input according to its own perspective and then sum all inputs up to get a single and holistic view. Finally, a neuron presents the holistic view to its internal judgment system to make a decision to fire or not.

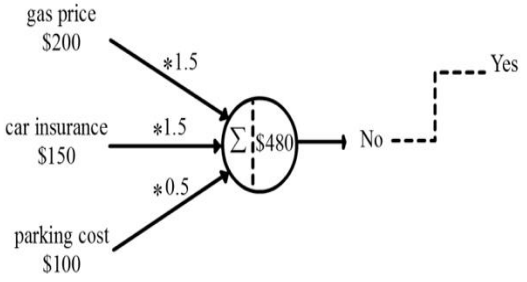

Fig 1. Perceptron neuron

This system seems trivial but it turns out to be a complicated decision-making model. For example, suppose that you are a neuron and you want to make a decision to buy car. You probably make that decision based on many variables which may include gas price ($200), car insurance ($150), and parking cost ($100). In your perspective, car insurance and gas price are more important and more likely to increase in near future than parking cost. In this case, you weigh up car insurance (1.5) and gas price (1.5) while downplay parking cost (0.5). Then you sum that up to get the holistic perspective (100*0.5 + 1.5*150 + 1.5*200). Therefore, according to your own perspective, a car would cost you $575 per month. Then you present this holistic perspective to your internal judgment system which may have been previously set on a specific threshold ($480). Therefore, you make a decision not to buy a car because it exceeds the threshold ($575 > $480). Your own perspectives of inputs, the internal judgment system, and the threshold are called weights, activation function and bias respectively. By changing weights and bias you reach a completely different decision. For example, set gas weight to 1 instead of 1.5 and notice the difference. Searching for weights and bias that generate the desired output is the job of learning algorithm.

Based on this foundational understanding, researchers arrange group of neurons to form a learnable network.

ANN may refer to two levels of abstraction:

(1) ANN as a person's brain and

(2) ANN as a group of learners.

Thus, network architecture refers first to a learner's inner abilities and mental capacities and; second, refers to a way in which designers of learning-environment arrange a network of learners. It is worth noting that ANN is a universal modeling system. Universality means that ANN can learn any given function no matter what neuron type is used. It has been proved that with few neurons and by changing biases and weights only, ANN can compute any zigzag-shaped function [2]. The question now is how we arrange neurons in ANN to make it easier for a learning algorithm to find those biases and weights.

For clarity and simplicity, the paper divides the most common ANN architectures based on three criteria: (1) number of layers, (2) flow of information and (3) neuron connectivity.

Number of layers:

By looking on how many layers a network has, ANN can be divided into (1) shallow and (2) deep networks.

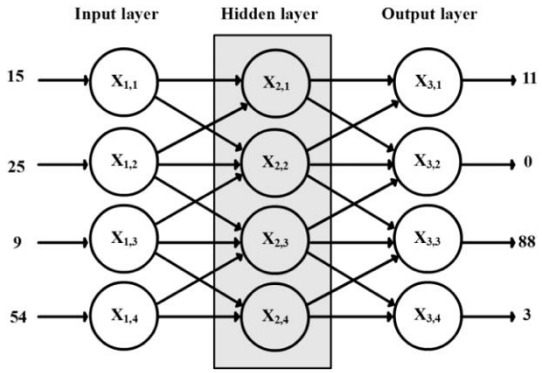

Fig. 2 Shallow neural network

A shallow neural network consists of three layers ordered from left to right: (1) input, (2) hidden and (3) output layer. The input layer does not really consist of neurons. Actually, it carries the input values to the network.

The second layer is named 'hidden' because it resides in the middle and does not appear in either the input or the output of the network. Other than that, it is a normal neural layer which contains normal neurons (Nielsen, 2015).

The output layer also contains normal neurons and its output represents the output of the network.

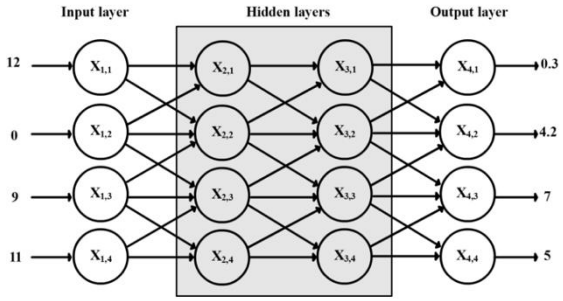

Fig. 3 Deep neural network

Flow of information:

By looking on how information flows through a network, ANN can be divided into (1) feedforward and (2) recurrent networks.

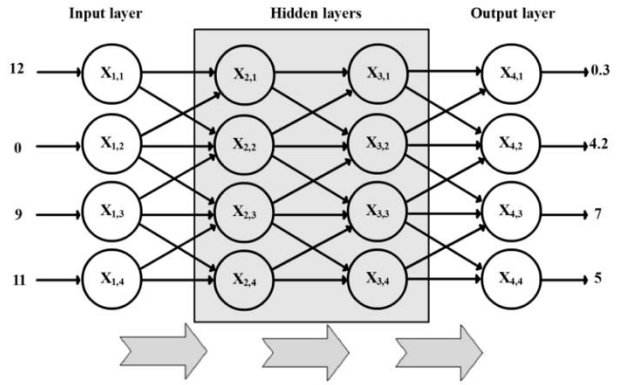

Fig. 4 Feedforward neural network

In feedforward networks, the output of a layer is used as an input for the next layer. There are no loops in feedforward networks; information flows in one direction where the output of a neuron can never return to its input. Feedforward network is one of the most used network structures. The value of this structure is self-explanatory since it significantly reduces the network complexity.

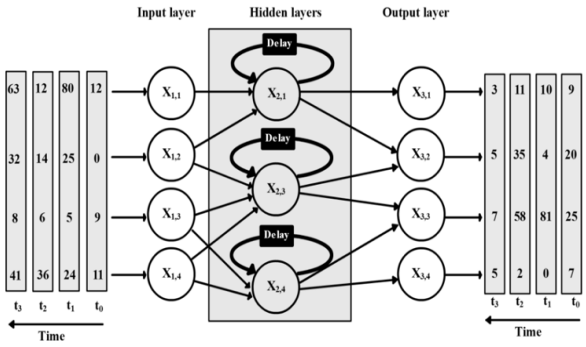

Fig. 5 Recurrent neural network

Recurrent network is a family of neural networks that processes the input sequentially and allows feedback connections [23]. Feedforward network structure assumes that all inputs are independent of each other [24]. It assumes that inputs order has no meaning. This, however, turns out to be false assumption for some tasks. For example, in natural language processing, the order of words makes a significant difference in meaning. Recurrent network tries to recover this issue by allowing feedback in a network. The feedback is allowed but with a delay constraint. That is, if the inputs are a sequence of A, B and C; then the output of hidden layer in step A can only be passed to the input of the hidden layer in step B, not the hidden layer in step A itself. To make a network simple, ANN researchers usually unfold the loop to see what it looks like on each step of the inputs. In Fig. 8, one can see that a loop allows information to flow from one step to another, and, therefore, acts as a memory [17, 10].

Learning Algorithm

Designing network architectures is a difficult task but training and teaching these networks are surely more difficult. To understand how ANN has been trained, it is better to start with a very simple one neuron example [26]. The principles which are used to teach a single neuron are also used to teach a whole network. However, a network level adds extra complexity which requires an additional step. Suppose you have a very simple neuron with one input and one output. You want to teach this neuron to do a certain task (for example to memorize a multiplication table for number 5). To teach this neuron, ANN researchers usually give it a so-called training set. A training set contains a number of different input values (1, 2, 3, 4, 5, 6 ...) paired with the correct output (5, 10, 15, 20, 25, 30 ...).

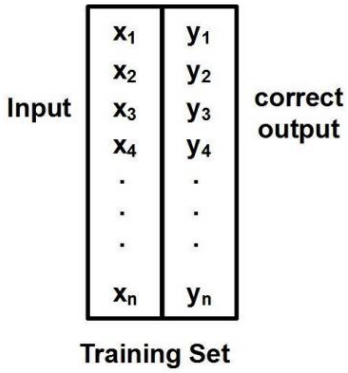

Fig. 6 Labeled training data

In the beginning, the neuron receives input and generates output according to its own weight and bias which were randomly selected. This means, the output of the neuron (as) would most probably differ from the correct output (ys).

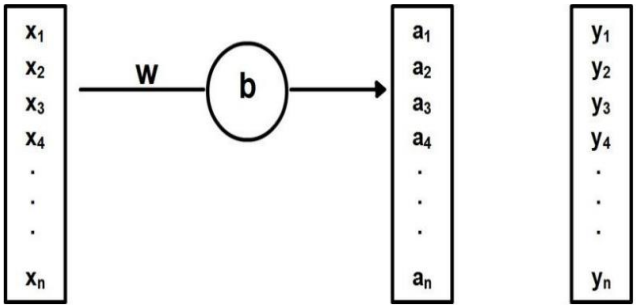

Fig. 7 Single neuron training

One note in ANN model of learning is how AI researchers are setting the value of learning rate. Actually, learning rate is one of many other parameters which are left free for human and outside of ANN's control. For example, (1) the number of layers, (2) the number of neurons in each layer, (3) the size of training set, (4) the activation function type, and (5) regularization parameter as well as (6) the learning rate are

some of those free parameters which are called hyperparameters [24, 18] Choosing the right values of hyper-parameters is left for a person who manages the ANN.

Bandura [18] criticizes those views of human learning which concentrate merely on neural patterns to interpret learning and argues that such views strip humans of agentic capabilities and a self-identity. In contrary, Bandura [18] conceives consciousness as an emergent property of brain activities which is not reducible solely to the property of neurons activity. In other words, the consciousness is higher-level force which is a result of lower-level neural activities but its properties are not limited to them. As clarified in this study, ANN design shows the need for consciousness force to manage and regulate ANN learning but this force does not occur as an emergent property of neural activity as Bandura proposes. Rather, it is a completely distinct entity which uses, guides and manages the neural activity and does not result from it. Therefore, overcoming hazards in the field AI becomes crucial to maximize societal benefit of AI given its significant expansion.

3. RESEARCH METHOD

The study provides a framework for conducting research to overcome AI hazards with a focus on ANN.

Research has been developed and constructed based on a review of various books focusing on Russell and Norvig (2016), Tinholt, et al. (2017), Tito (2017), and Zhang and Dafoe (2019). This research identifies various concepts that are very helpful in formulating final questions. These simple but effective methods are useful to achieve the purpose of exploratory research.

The focus of any type of research should be on delivering AI that is beneficial to society and robust in the sense that the benefits are guaranteed: our AI systems must do what we want them to do.

Different ways in which an AI system may fail to perform as desired correspond to different areas of robustness research:

1. Verification: How can it be proven that a system satisfies certain desired formal properties? ("Did I build the system right?")

2. Validity: How can it be ensured that a system that meets its formal requirements does not have unwanted behaviors and consequences? ("Did I build the right system?")

3. Security: How can one prevent intentional manipulation by unauthorized parties?

4. Control: How can one enable meaningful human control over an AI system after it begins to operate? ("Ok, I built the system wrong, can I fix it?")

Verification

Verification refers to methods that yield high confidence that a system will satisfy a set of formal constraints. When possible, it is desirable for systems in safety-critical situations, e.g. self-driving cars, to be verifiable.

A lack of design-time knowledge also motivates the use of learning algorithms within the agent software, and verification becomes more difficult: statistical learning theory gives so-called ε-δ (probably approximately correct) bounds, mostly for the somewhat unrealistic settings of supervised learning from data and single-agent reinforcement learning with simple architectures and full observability, but even then requiring prohibitively large sample sizes to obtain meaningful guarantees.

Not only should it be possible to build AI systems on top of verified substrates; it should also be possible to verify the designs of the AI systems themselves, particularly if they follow a "componentized architecture", in which guarantees about individual components can be combined according to their connections to yield properties of the overall system. Agent architectures used in Russell and Norvig [12, 19] separate an agent into distinct modules (predictive models, state estimates, utility functions, policies, learning elements, etc. Research on richer kinds of agents—for example, agents with layered architectures, anytime components, overlapping deliberative and reactive elements, metalevel control, etc.—could contribute to the creation of verifiable agents, yet, there is a lack of the formal "algebra" to properly define, explore, and rank the space of designs.

Existing body of research would be most valuable to reducing the risk of adverse outcomes arising from bugs in implementation. This work would most likely be less theoretical and more practical and implementation-specific than most of the other research explored in this document. Some of the questions to investigate here are:

1. What categories of bugs are most hazardous? Some particularly undesirable sorts of bugs are:

   (a) bugs that lie dormant during ordinary testing but can be encountered in larger settings given enough time. (For example, integer overflows or accumulation of numerical error.)

   (b) portability bugs, ie bugs that arise from differences in libraries, environment, or hardware.

   (c) "Heisenbugs", ie bugs that manifest in practice but not in a debugging environment.

(d) bugs that are difficult to reproduce for some reason, such as bugs affected by non-deterministic scheduling of concurrent threads of execution, or by the interaction of this with some other sort of state, such as a random number generator.

2. How likely would these sorts of bugs be to arise in a hazardous way if an otherwise-promising super-intelligence project was undertaken in the medium-term?

3. What kinds of tools or software changes would make the most difference in mitigating the risk of an adverse outcome? Some ideas:

(a) influence current and upcoming programming language interpreters, compilers, application virtual machines, etc. to adopt a default behavior (or at least an option) of throwing exceptions on encountering numerical overflow/underflow.

(b) ensure software quality of particularly popular state-of-the-art machine learning libraries, and other core components.

(c) assess the prevalence of portability bugs and promote adherence of standards that could resolve them.

Validity

In order to build systems that robustly behave well, we of course need to decide what "good behavior" means in each application domain.[13] This ethical question is tied intimately to questions of what engineering techniques are available, how reliable these techniques are, and what trade-offs can be made –– all areas where computer science, machine learning, and broader AI expertise is valuable.

For example, Wallach and Allen [28] argue that a significant consideration is the computational expense of different behavioral standards (or ethical theories): if a standard cannot be applied efficiently enough to guide behavior in safety-critical situations, then cheaper approximations may be needed. Designing simplified rules – for example, to govern a self-driving car's decisions in critical situations – will likely require expertise from both ethicists and computer scientists.

Security

Security research can help make AI more robust. At a higher level, research into specific AI and machine learning techniques may become increasingly useful in security. These techniques could be applied to the detection of intrusions [16], analyzing malware [17], or detecting potential exploits in other programs through code analysis [20]. It is not implausible that cyberattack between states and private actors will be a risk factor for harm from near-future AI systems, motivating research on preventing harmful events.

As AI systems are used in an increasing number of critical roles, they will take up an increasing proportion of cyber-attack surface area. Robustness against exploitation at the low level is closely tied to verifiability and freedom from bugs. For example, the DARPA SAFE program aims to build an integrated hardware-software system with a flexible metadata rule engine, on which can be built memory safety, fault isolation, and other protocols that could improve security by preventing exploitable flaws [30]. Such programs cannot eliminate all security flaws (since verification is only as strong as the assumptions that underly the specification), but could significantly reduce vulnerabilities of the type exploited by the recent "Heartbleed bug" and "Bash Bug". Such systems could be preferentially deployed in safety-critical applications, where the cost of improved security is justified.

As AI systems grow more complex and are networked together, they will have to intelligently manage their trust, motivating research on statistical-behavioral trust establishment [94] and computational reputation models [13].

Control

For certain types of safety-critical AI systems – especially vehicles and weapons platforms – it may be desirable to retain some form of meaningful human control, whether this means a human in the loop, on the loop[14, 18], or some other protocol. In any of these cases, there will be technical work needed in order to ensure that meaningful human control is maintained [13].

Automated vehicles are a test-bed for effective control-granting techniques. The design of systems and protocols for transition between automated navigation and human control is a promising area for further research. Such issues also motivate broader research on how to optimally allocate tasks within human- computer teams, both for identifying situations where control should be transferred, and for applying human judgment efficiently to the highest-value decisions.

RESULTS

A frequently discussed long-term goal of some AI researchers is to develop systems that can learn from experience with human-like breadth and surpass human performance in most cognitive tasks, thereby having a major impact on society.

Assessments of this success probability vary widely between researchers, but few would argue with great confidence that the probability is negligible, given the track record of such predictions.

The threat posed by a sufficiently advanced artificial intelligence may depend much more on its cognitive capabilities and its goal architecture than on the physical capabilities with which it is initially endowed. Not all risks related to robots or artificial intelligences are to be classified as information system hazards. A risk would count as such a hazard if, for example, it arose from the possibility of a computer virus infecting the operating system for a robot or an AI. But aside from such special cases, we shall not count robot hazards and artificial intelligence hazards as information system hazards.

Validity

Designing a powerful AI system without having a thorough understanding of these issues might increase the risk of unintended consequences, both by foregoing tools that could have been used to increase the system's reliability, and by risking the collapse of shaky foundations. Example research topics in this area include reasoning and decision under bounded computational resources `a la Horvitz and Russell [59, 100], how to take into account correlations between AI systems' behaviors and those of their environments or of other agents [14, 6, 8, 4, 15],how agents that are embedded in their environments should reason [10, 7], and how to reason about uncertainty over logical consequences of beliefs or other deterministic computations [14, 3].

In the long term, it is plausible that we will want to make agents that act autonomously and powerfully across many domains. Explicitly specifying our preferences in broad domains in the style of near-future machine ethics may not be practical, making "aligning" the values of powerful AI systems with our own values and preferences difficult [11, 13].

Reinforcement learning raises its own problems: when systems become very capable and general, then an effect similar to Goodhart's Law is likely to occur, in which sophisticated agents attempt to manipulate or directly control their reward signals [16]. This motivates research areas that could improve our ability to engineer systems that can learn or acquire values at run- time. For example, inverse reinforcement learning may offer a viable approach, in which a system infers the preferences of another actor, assumed to be a reinforcement learner itself [11, 8].

As systems become more capable, more epistemically difficult methods could become viable, suggesting that research on such methods could be useful.

Security

It is unclear whether long-term progress in AI will make the overall problem of security easier or harder; on one hand, systems will become increasingly complex in construction and behavior and AI-based cyberattacks may be extremely effective, while on the other hand, the use of AI and machine learning techniques along with significant progress in low-level system reliability may render hardened systems much less vulnerable than today's.

Some of the attributes that may be desirable or necessary are:

- Containment: it should prevent a contained super-intelligent AI from having arbitrary effects on the world. In particular, it should be verifiably free of vulnerabilities itself.
- Robustness: it should be difficult to unintentionally render ineffective.
- Uptake: It should be a system that AI builders want to use, and avoid being one that they want to not use.
- Inspectability: It should allow detailed debugging and inspection of the contained AI. This could contribute to uptake if it provides better inspection capabilities than AI builders typically have (for instance, debugging distributed software is typically awkward in the current state of affairs).

Control

It has been argued [13] that the nature of the general AI control problem undergoes an essential shift, which can be referred to as the "context change", when transitioning from subhuman to superhuman general AI. This suggests that rather than judging potential solutions to the control problem using only experimental results, it is essential to build compelling deductive arguments that generalize and are falsifiable, and only when these arguments are available does it make sense to try to test potential solutions via experiment.

In general, an accident can be described as a situation where a human designer had in mind a certain (perhaps informally specified) objective or task, but the system that was designed and deployed for that task produced harmful and unexpected results. This issue arises in almost any engineering discipline, but may be particularly important to address when building AI systems [16].

- The designer may have specified the wrong formal objective function, such that maximizing that objective function leads to harmful results, even in the limit of perfect learning and infinite data.

Negative side effects and reward hacking describe two broad mechanisms that make it easy to produce wrong objective functions. In "negative side effects", the designer specifies an objective function that focuses on accomplishing some specific task in the environment, but ignores other aspects of the (potentially very large) environment, and thus implicitly expresses indifference over environmental variables that might actually be harmful to change. In "reward hacking", the objective function that the designer writes down admits of some clever "easy" solution that formally maximizes it but perverts the spirit of the designer's intent.

- The designer may know the correct objective function, or at least have a method of evaluating it (for example explicitly consulting a human on a given situation), but it is too expensive to do so frequently, leading to possible harmful behavior caused by bad extrapolations from limited samples.

- The designer may have specified the correct formal objective, such that we would get the correct behavior were the system to have perfect beliefs, but something bad occurs due to making decisions from insufficient or poorly curated training data or an insufficiently expressive model.

Within this context, the following research questions can be developed:

1. Can high Bayesian uncertainty and agent respect for the unknown act as an effective safety mechanism?

2. How can one investigate steep temporal discounting as an incentives control method for an untrusted general AI?

Safety

As predicting the exact behavior of complex software is notoriously difficult the purpose of AI safety research is therefore more modest: to show that the behavior, although not exactly predictable, will have certain de- sired properties, for example keeping certain behavioral parameters within certain bounds.

Rational agents are often composed of distinct modules (e.g. sensors, actuators, a performance element, a learning element, a problem generator, a critic, etc.), each with limited abilities, with some network of information flows between modules.[10] Within this framework, it would be valuable to provide guarantees that various modules would be safe or unsafe (individually or in combination).

Many of the above-mentioned safety issues are related to the issue of goals that the rational agent may have. This question provides an important link between architectures and goals: how amenable are different AI architectures to having their goals and beliefs read from the outside in a fashion useful for safety determination and monitoring?

DISCUSSION

There are many ways of responding to information hazards. In many cases, the best response is no response, i.e., to proceed as though no such hazard existed. The benefits of information may so far outweigh its costs that even when information hazards are fully accounted for, we still under-invest in the gathering and dissemination of information. Moreover, ignorance carries its own dangers which are oftentimes greater than those of knowledge. Information risks might simply be tolerated.

When mitigation is called for, it need not take the form of an active attempt to suppress information through measures such as bans, censorship, disinformation campaigns, encryption, or secrecy. One response option is simply to invest less in discovering and disseminating certain kinds of information. Somebody who is worried about the spoiler hazard of learning about the ending of a movie can simply refrain from reading reviews and plot summaries.

At the same time, however, we should recognize that knowledge and information frequently have downsides. Future scientific and technological advances, in particular, may create information which, misused, would cause tremendous harm—including, potentially, existential catastrophe.

It can also be hoped that new information technologies will bring about a vastly more transparent society, in which everybody (the watchmen included) are under constant surveillance; and that this universal transparency will prevent the worst potential misuses of the new technological powers that humanity will develop.

CONCLUSION

Even if our best policy is to form an unyielding commitment to unlimited freedom of thought, virtually limitless freedom of speech, an extremely wide freedom of inquiry, we should realize not only that this policy has costs but that perhaps the strongest reason for adopting such an uncompromising stance would itself be based on an information hazard; namely, norm hazard: the risk that precious yet fragile norms of truth-seeking and truthful reporting would be jeopardized if we permitted convenient exceptions in our own adherence to them or if their violation were in general too readily excused.

It is said that a little knowledge is a dangerous thing. It is an open question whether more knowledge is safer. Even if our best bet is that more knowledge is on average good, we should recognize that there are numerous cases in which more knowledge makes things worse.

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant. "Privacy-preserving data mining". In: ACM Sigmod Record 29.2 (2000), pp. 439–450.

[2] Rajeev Alur. "Formal verification of hybrid systems". In: Embedded Software (EMSOFT), 2011 Proceedings of the International Conference on. IEEE. 2011, pp. 273–278.

[3] Kenneth Anderson, Daniel Reisner, and Matthew C Waxman. "Adapting the Law of Armed Conflict to Autonomous Weapon Systems". In: International Law Studies 90 (2014).

[4] Susan Leigh Anderson and Michael Anderson. "A Prima Facie Duty Approach to Machine Ethics Machine Learning of Features of Ethical Dilemmas, Prima Facie Duties, and Decision Principles through a Dialogue with Ethicists". In: Machine Ethics (2011), p. 476.

[5] David Andre and Stuart J Russell. "State abstraction for programmable reinforcement learning agents". In: Eighteenth national conference on Artificial intelligence. American Association for Artificial Intelligence. 2002, pp. 119–125.

[6] Stuart Armstrong, Nick Bostrom, and Carl Shulman. "Racing to the precipice: a model of artificial intelligence development". In: (2013).

[7] Stuart Armstrong, Kaj Sotala, and Se´an S O hEigeartaigh. "The errors, insights and lessons of famous AI predictions–and what they mean for the future". In: Journal of Experimental & Theoretical Artificial Intelligence ahead-of-print (2014), pp. 1–26. url: http://www.fhi.ox.ac.uk/wp-content/uploads/FAIC.pdf.

[8] Gustaf Arrhenius. "The impossibility of a satisfactory population ethics". In: Descriptive and normative approaches to human behavior (2011).

[9] Peter M Asaro. "What should we want from a robot ethic?" In: International Review of Information Ethics 6.12 (2006), pp. 9–16.

[10] Peter Asaro. "How just could a robot war be?" In: Current issues in computing and philosophy (2008), pp. 50–64.

[11] Karl J Astr¨om and Bj¨orn Wittenmark. Adaptive control. Courier Dover Publications, 2013.

[12] Silvia Bellezza, Anat Keinan, and Neeru Paharia. Conspicuous Consumption of Time: When Busyness at Work and Lack of Leisure Time Become a Status Symbol. 2014. url: http://www.hbs.edu/faculty/Pages/item.aspx?num=47139.

[13] M Boden et al. "Principles of robotics". In: The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC). web publication (2011).

[14] Nick Bostrom. "Infinite ethics". In: Analysis and Metaphysics 10 (2011), pp. 9–59.

[15] Nick Bostrom. Moral Uncertainty–Towards a Solution? 2009. url: http://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html.

[16] Nick Bostrom. Superintelligence: Paths, dangers, strategies. Oxford University Press, 2014.

[17] Nick Bostrom. "The superintelligent will: Motivation and instrumental rationality in advanced arti- ficial agents". In: Minds and Machines 22.2 (2012), pp. 71–85.

[18] Ju¨rgen Branke et al. Multiobjective optimization: Interactive and evolutionary approaches. Vol. 5252. Springer Science & Business Media, 2008.

[19] Selmer Bringsjord et al. "Piagetian roboethics via category theory: Moving beyond mere formal operations to engineer robots whose decisions are guaranteed to be ethically correct". In: Machine ethics (2011), pp. 361–374.

[20] Yuriy Brun and Michael D Ernst. "Finding latent code errors via machine learning over program executions". In: Proceedings of the 26th International Conference on Software Engineering. IEEE Computer Society. 2004, pp. 480–490.

[21] Erik Brynjolfsson and Andrew McAfee. The second machine age: work, progress, and prosperity in a time of brilliant technologies. W.W. Norton & Company, 2014.

[22] Erik Brynjolfsson, Andrew McAfee, and Michael Spence. "Labor, Capital, and Ideas in the Power Law Economy". In: Foreign Aff. 93 (2014), p. 44.

[23] Ryan Calo. "Robotics and the New Cyberlaw". In: Available at SSRN 2402972 (2014).

[24] Ryan Calo. "The Case for a Federal Robotics Commission". In: Available at SSRN 2529151 (2014).

[25] David Chalmers. "The singularity: A philosophical analysis". In: Journal of Consciousness Studies 17.9-10 (2010), pp. 7–65.

[26] Wei Chu and Zoubin Ghahramani. "Preference Learning with Gaussian Processes". In: In Proc. ICML 2005. 2005, pp. 137–144.

[27] Robin R Churchill and Geir Ulfstein. "Autonomous institutional arrangements in multilateral environmental agreements: a little-noticed phenomenon in international law". In: American Journal of International Law (2000), pp. 623–659.

[28] Andrew E Clark and Andrew J Oswald. "Unhappiness and unemployment". In: The Economic Journal (1994), pp. 648–659.

[29] Owen Cotton-Barratt and Toby Ord. Strategic considerations about different speeds of AI takeoff. Aug. 2014. url: http://www.fhi.ox.ac.uk/strategic-considerations-about-different- speeds-of-ai-takeoff/.

[30] Andr´e DeHon et al. "Preliminary design of the SAFE platform". In: Proceedings of the 6th Workshop on Programming Languages and Operating Systems. ACM. 2011, p. 4.

[31] Louise A Dennis et al. "Practical Verification of Decision-Making in Agent-Based Autonomous Systems". In: arXiv preprint arXiv:1310.2431 (2013).

[32] Daniel Dewey. "Long-term strategies for ending existential risk from fast takeoff". In: (Nov. 2014). url: http://www.danieldewey.net/fast-takeoff-strategies.pdf.

[33] United Nations Institute for Disarmament Research. The Weaponization of Increasingly Autonomous Technologies: Implications for Security and Arms Control. UNIDIR, 2014.

[34] Bonnie Lynn Docherty. Losing Humanity: The Case Against Killer Robots. Human Rights Watch, 2012.

[35] Peter Eckersley and Anders Sandberg. "Is Brain Emulation Dangerous?" In: Journal of Artificial General Intelligence 4.3 (2013), pp. 170–194.

[36] Beno Eckmann. "Social choice and topology a case of pure and applied mathematics". In: Expositiones Mathematicae 22.4 (2004), pp. 385–393.

[37] Benja Fallenstein and Nate Soares. Vingean Reflection: Reliable Reasoning for Self-Modifying Agents. Tech. rep. Machine Intelligence Research Institute, 2014. url: https://intelligence.org/files/VingeanReflection.pdf.

[38] Kathleen Fisher. "HACMS: high assurance cyber military systems". In: Proceedings of the 2012 ACM conference on high integrity language technology. ACM. 2012, pp. 51–52.

[39] Carl Frey and Michael Osborne. The future of employment: how susceptible are jobs to computerisa- tion? Working Paper. Oxford Martin School, 2013.

[40] Edward L Glaeser. "Secular joblessness". In: Secular Stagnation: Facts, Causes and Cures (2014), p. 69.

[41] Irving John Good. "Speculations concerning the first ultraintelligent machine". In: Advances in computers 6.31 (1965), p. 88.

[42] Katja Grace. Algorithmic Progress in Six Domains. Tech. rep. Machine Intelligence Research Institute, 2013. url: http://intelligence.org/files/AlgorithmicProgress.pdf.

[43] Katja Grace and Paul Christiano. Resolutions of mathematical conjectures. 2014. url: http://www.aiimpacts.org/resolutions-of-mathematical-conjectures.

[44] The Tauri Group. Retrospective Analysis of Technology Forecasting: In-scope Extension. Tech. rep. 2012. url: http://www.dtic.mil/get-tr-doc/pdf?AD=ADA568107.

[45] Tom Gunter et al. "Sampling for inference in probabilistic models with fast Bayesian quadrature". In: Advances in Neural Information Processing Systems. 2014, pp. 2789–2797.

[46] Joseph Y. Halpern and Rafael Pass. "Game Theory with Translucent Players". In: CoRR abs/1308.3778 (2013). url: http://arxiv.org/abs/1308.3778.

[47] Joseph Y. Halpern and Rafael Pass. "I Don't Want to Think About it Now: Decision Theory With Costly Computation". In: CoRR abs/1106.2657 (2011). url: http://arxiv.org/abs/1106.2657.

[48] Joseph Y Halpern, Rafael Pass, and Lior Seeman. "Decision Theory with Resource-Bounded Agents". In: Topics in cognitive science 6.2 (2014), pp. 245–257.

[49] Kristian J Hammond, Timothy M Converse, and Joshua W Grass. "The stabilization of environments". In: Artificial Intelligence 72.1 (1995), pp. 305–327.

[50] Robin Hanson. "Economics of the singularity". In: Spectrum, IEEE 45.6 (2008), pp. 45–50.