AIR_5: Five Pillars of Artificial Intelligence Research

Yew-Soon Ong and Abhishek Gupta

***Abstract* – In this article, we provide an overview of what we consider to be some of the most pressing research questions currently facing the fields of *artificial and computational intelligence* (AI and CI). While AI spans a range of methods that enable machines to learn from data and operate autonomously, CI serves as a means to this end by finding its niche in algorithms that are inspired by complex natural phenomena (including the working of the brain). In this paper, we demarcate the key issues surrounding these fields using five unique *Rs*, namely, (i) *rationalizability*, (ii) *resilience*, (iii) *reproducibility*, (iv) *realism*, and (v) *responsibility*. Notably, just as *air* serves as the basic element of biological life, the term AIR_5 – cumulatively referring to the five aforementioned *Rs* – is introduced herein to mark some of the basic elements of artificial life, *for sustainable AI and CI*. A brief summary of each of the *Rs* is presented, highlighting their relevance as pillars of future research in this arena.**

I. Introduction

The original inspiration of *artificial intelligence* (AI) was to build autonomous systems capable of matching human-level intelligence in specific domains. Likewise, the closely related field of *computational intelligence* (CI) emerged in an attempt to artificially recreate the consummate learning and problem-solving facility observed in various forms in nature – spanning examples in cognitive computing that mimic complex functions of the human brain, to algorithms that are inspired by efficient foraging behaviors found in seemingly simple organisms like ants. Notwithstanding their (relatively) modest beginnings, in the present-day, the combined effects of (i) easy access to massive and growing volumes of data, (ii) rapid increase in computational power, and (iii) steady improvements in data-driven *machine learning* (ML) algorithms [1-3], have played a major role in helping modern AI systems vastly surpass humanly achievable performance across a variety of applications. In this regard, some of the most prominent success stories that have made international headlines include IBM's Watson winning Jeopardy! [4], Google DeepMind's AlphaGo program beating the world's leading Go player [5], their AlphaZero algorithm learning entirely via "self-play" to defeat a world champion program in the game of chess [6], and Carnegie Mellon University's AI defeating four of the world's best professional poker players [7].

Due to the accelerated development of AI technologies witnessed over the past decade, there is increasing consensus that the field is primed to have a significant impact on society as a whole. Given that much of what has been achieved by mankind is a product of human intellect, it is evident that the possibility of augmenting cognitive capabilities with AI (a synergy that is also referred to as *augmented intelligence* [8]) holds immense potential for improved *decision intelligence* in high-impact areas such as healthcare, environmental science, economics, governance, etc. That said, there continue to exist major scientific challenges that require foremost attention for the concept of AI to be more widely trusted, accepted, and seamlessly integrated within the fabric of society. In this article, we demarcate some of these challenges using five unique *R*s – namely, (i) R_1: *rationalizability*, (ii) R_2: *resilience*, (iii) R_3: *reproducibility*, (iv) R_4: *realism*, and (v) R_5: *responsibility* – which, in our opinion, represent five key pillars of AI research that shall support the sustained growth of the field through the 21st century and beyond. In summary, we highlight that just as *air* serves as the basic element of biological life, the term AIR_5 – cumulatively referring to the five aforementioned *R*s – is used herein to mark some of the basic elements of artificial life.

The remainder of the article is organized to provide a brief summary of each of the five *R*s, drawing attention to their fundamental relevance towards the future of AI.

II. R_1: Rationalizability of AI Systems

Currently, many of the innovations in AI are driven by ML techniques centered around the use of so-called *deep neural networks* (DNNs) [2, 3]. The design of DNNs is loosely based on the complex biological neural network that constitutes a human brain – which (unsurprisingly) has drawn significant interest over the years as a dominant source of intelligence in the natural world. However, DNNs are often criticized for being highly *opaque*. It is widely acknowledged that although these models can frequently attain remarkable prediction accuracies, their layered non-linear structure makes them exceedingly difficult to *interpret* (loosely defined as the science of comprehending what a model might have done [9]) and to draw *explanations* as to why certain inputs lead to the observed outputs / predictions / decisions. Due to the lack of transparency and causality, DNN models have come to be used mainly as *black-boxes* [10, 11].

With the above in mind, it is argued that for humans to cultivate greater acceptance of modern AI systems, their workings and the resultant outputs need to be made more *rationalizable – i.e., possess the ability to be rationalized (interpreted and explained)*. Most of all, the need for rationalizability cannot be compromised in safety critical applications where it is imperative to fully understand and verify what an AI system has learned before it can be deployed in the wild; illustrative applications

This work was supported in part by the Data Science and Artificial Intelligence Research Centre of the School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore, and in part by the SIMTech-NTU Joint Lab on Complex Systems. *(Corresponding author: Yew-Soon Ong)*

Yew-Soon Ong is Chief Artificial Intelligence Scientist with the Agency for Science, Technology and Research (A*STAR), Singapore, and is also with the Data Science and Artificial Intelligence Research Centre, School of Computer Science and Engineering, NTU, Singapore. E-mail: asysong@ntu.edu.sg

Abhishek Gupta is a Scientist with the Singapore Institute of Manufacturing Technology (SIMTech), A*STAR, Singapore. E-mail: abhishek_gupta@simtech.a-star.edu.sg

include medical diagnosis, autonomous driving, etc., where peoples' lives are immediately at stake. For example, a well-known study revealing the threat of opacity in *neural networks* (NNs) is the prediction of patient mortality in the area of community-acquired pneumonia [12]. While NNs were *seemingly* the most accurate model for this task (when measured on available test data), an alternate (less accurate but more interpretable) rule-based system was found to uncover the following rule from one of the pneumonia datasets: $HasAsthma(\boldsymbol{x}) \Rightarrow LowerRiskOfDeath(\boldsymbol{x})$ [13]. By being patently dubious, the inferred rule shed light on a definite (albeit grossly misleading) pattern in the data that was used to train the system – a pattern that may have hampered the NN as well. Unfortunately, the inability to examine and verify the correctness of trained NNs in such delicate situations often tends to preclude their practical applicability; this turned out to be the case for the patient mortality prediction problem. Similar situations may be encountered in general scientific and engineering disciplines as well, where an AI system must at least be consistent with the fundamental laws of physics for it to be considered trustworthy. The development of rationalizable models, which are grounded in established theories, can thus go a long way in protecting against potential mishaps caused by the inadvertent learning of spurious patterns from raw data [14, 15].

It is contended that although interpretable and explainable AI are indeed at the core of rationalizability, they are not the complete story. Given previously unseen input data, while it may be possible to obtain explanations of a model's predictions, the *level of confidence* that the model has in its own predictions may not be appropriately captured and represented; it is only rational for such uncertainties to exist, especially for cases where an input point is located outside the regime of the dataset that was used for model training. Probability theory provides a mathematical framework for representing this uncertainty, and is thus considered to be another important facet of AI rationalizability – assisting the end-user in making more informed decisions by taking into account all possible outcomes. In this regard, it is worth noting that although DNNs are (rightly) considered to be state-of-the-art among ML techniques, they do not (as of now) satisfactorily represent uncertainties [16]. This sets the stage for future research endeavors in probabilistic AI and ML, with some foundational works in developing a principled Bayesian interpretation of common deep learning algorithms recently presented in [17, 18].

III. R_2: Resilience of AI Systems

Despite the spectacular progress of AI, latest research has shown that even the most advanced models (e.g., DNNs) have a peculiar tendency of being easily fooled [19]. Well-known examples have surfaced in the field of computer vision [20], where the output of a trained DNN classifier is found to be drastically altered by simply introducing a small additive perturbation to an input image. Generally, the added perturbation (also known as an *adversarial attack*) is so small that it is completely imperceptible to the human eye, and yet causes the DNN to misclassify. In extreme cases, attacking only a single pixel of an image has been shown to suffice in fooling various types of DNNs [21]. A particularly instructive illustration of the overall phenomenon is described in [22], where, by adding a few black and white stickers to a "Stop" sign, an image recognition AI was fooled into classifying it as a "Speed Limit 45" sign. It is worth highlighting that similar results have been reported in speech recognition applications as well [23].

While the consequences of such gross misclassification can evidently be dire, the aforementioned ("Stop" sign) case-study is especially alarming for industries like that of self-driving cars. For this reason, there have been targeted efforts over recent years towards attempting to make DNNs more *resilient – i.e., possess the ability to retain high predictive accuracy even in the face of adversarial attacks (input perturbations*). To this end, some of the proposed defensive measures include brute-force adversarial training [24], gradient masking / obfuscation [25], defensive distillation [26], and network add-ons [27], to name a few. Nevertheless, the core issues are far from being eradicated, and demand significant future research attention [28].

In addition to adversarial attacks that are designed to occur after a fully trained model is deployed for operation, *data poisoning* has emerged as a different kind of attack that can directly cripple the training phase. Specifically, the goal of an attacker in this setting is to *subtly* adulterate a training dataset – either by adding new data points [29] or modifying existing ones [30] – such that the learner is forced to learn a bad model. Ensuring performance robustness against such attacks is clearly of paramount importance, as the main ingredient of all ML systems – namely, the training data itself – is drawn from the outside world where it is vulnerable to intentional or unintentional manipulation [31]. Challenges are further exacerbated for modern ML paradigms such as *federated learning* that are designed to run on *fog networks* [32], where the parameters of a centralized global model are to be updated via distributed computations carried out using data stored locally across a federation of participating devices (e.g., mobile *edge devices* including hand phones, smart wearables, etc.); thus, making pre-emptive measures against malicious data poisoning attacks indispensable for *secure AI*.

IV. R_3: Reproducibility of AI Systems

An often talked about challenge faced while training DNNs, and ML models in general, is the replication crisis [33]. Essentially, some of the key results reported in the literature are found to be difficult to reproduce by others. As noted in [34], for any claim to be believable and informative, *reproducibility* is a minimum necessary condition. Thus, ensuring performance reproducibility of AI systems by creating and abiding by clear software standards, as well as rigorous system verification and validation on shared datasets and benchmarks, is vital for maintaining their trustworthiness. In what follows, we briefly discuss two other complementary tracks in pursuit of the desired outcome.

A significant obstacle in the path of successfully reproducing published results is the large number of *hyperparameters* – e.g., neural architectural choices, parameters of the learning algorithm, etc. – that must be precisely configured before training a model on any given dataset [35]. Even though these configurations typically receive secondary treatment among the core constituents of a model or learning algorithm, their setting can considerably affect the efficacy of the learning process. Consequently, the lack of expertise in optimal hyperparameter selection can lead to unsatisfactory performance of the trained model. Said differently, the model fails to live up to its true potential, as may have been reported in a scientific publication. With the above in mind, a promising alternative to hand-crafted hyperparameter configuration is to *automate* the entire process, by posing it as a *global optimization* problem. To this end, a range of

techniques, encompassing stochastic evolutionary algorithms [36, 37] as well as Bayesian optimization methods [38] have been proposed, making it possible to select near-optimal hyperparameters without the need for a human in the loop (thus preventing human inaccuracies). The overall approach falls under the scope of so-called *AutoML* (automated machine learning [39]), a topic that has recently been attracting much attention among ML practitioners.

At the leading edge of AutoML is an ongoing attempt to develop algorithms that can automatically transfer and reuse learned knowledge across datasets, problems, and domains [40]. The goal is to enhance the *generalizability* of AI, such that performance efficacy is not only confined to a specific (narrow) task, but can also be reproduced in other *related* tasks by sharing common building-blocks of knowledge. In this regard, promising research directions include *transfer* and *multitask learning* [41-43], and their extensions to the domain of global optimization (via transfer and multitask optimization [44-49]). An associated research theme currently being developed in the area of nature-inspired CI is *memetic computation* – where the sociological notion of a *meme* (originally defined in [50] as a basic unit of information that resides in the brain, and is replicated from one brain to another by the process of imitation) has been transformed to embody diverse forms of computationally encoded knowledge that can be learned from one task and transmitted to another, with the aim of endowing an AI with human-like general problem-solving ability [51].

Alongside the long-term development of algorithms that can automate the process of hyperparameter selection, a more immediate step for encouraging AI reproducibility is to inculcate the practice of sharing well-documented source codes and datasets from scientific publications. Although open collaborations and open-source software development are becoming increasingly common in the field of AI, a recent survey suggests that the current documentation practices at top AI conferences continue to render the reported results mostly irreproducible [52]. In other words, there is still a need for universally agreed software standards to be established – pertaining to code documentation, data formatting, setup of testing environments, etc. – so that the evaluation of AI technologies can be carried out more easily.

V. R_4: Realism of AI Systems

The three *R*s presented so far mainly focus on the performance efficacy and precision of AI systems. In this section, we turn our attention to the matter of instilling machines with a degree of *emotional intelligence*, which, looking ahead, is deemed equally vital for the seamless assimilation of AI in society.

In addition to being able to absorb and process vast quantities of data to support large-scale industrial automation and complex decision-making, AI has shown promise in domains involving intimate human interactions as well; examples include the everyday usage of smart speakers (like Google Home devices and Amazon's Alexa), the improvement of education through virtual tutors [53], and even providing psychological support to Syrian refugees through the use of chat-bots [54]. To be trustworthy, such *human-aware AI* systems [55] must not only be accurate, but should also embody human-like virtues of relatability, benevolence, and integrity. In our pursuit to attain a level of *realism* in intelligent systems, *a balance must be sought between the constant drive for high precision and automation, and the creation of machine behaviors that lead to more fulfilling human-computer interaction*. Various research threads have emerged in this regard.

On one hand, the topic of *affective computing* aims for a better understanding of humans, by studying the enhancement of AI proficiency in recognizing, interpreting, and expressing real-life emotions and sentiments [56]. One of the key challenges facing the subject is the development of systems that can detect and process *multimodal data streams*. The motivating rationale stems from the observation that different people express themselves in different ways, utilizing diverse modes of communication (such as speech, body-language, facial expressions, etc.) to varying extent. Therefore, in most cases, the fusion of visual and aural information cues is able to offer a more holistic understanding of a person's emotion, at least in comparison to the best unimodal analysis techniques that process separate emotional cues in isolation [57, 58].

In contrast to affective computing, which deals with a specific class of human-centered learning problems, *collective intelligence* is a meta-concept that puts forth the idea of explicitly tapping on the wisdom of a "crowd of people" to shape AI [54]. As a specific (technical) example, it was reported in [59] that through a *crowdsourcing* approach to feature engineering on big datasets, ML models could be trained to achieve state-of-the-art performance within short task completion time. Importantly, the success of this socially guided ML exercise shed light on the more general scope of combining human expertise (i.e., knowledge memes) into the AI training process, thus encouraging the participation of social scientists, behaviorists, humanists, ethicists, etc., in molding AI technologies. Successfully harnessing the wide range of expertise will introduce a more human element into the otherwise mechanized procedure of learning from raw data, thereby promising a greater degree of acceptance of AI in society's eye.

VI. R_5: Responsibility of AI Systems

Last but not least, we refer to the IEEE guidelines on Ethically Aligned Design, which states the following:

"As the use and impact of autonomous and intelligent systems become pervasive, we need to establish societal and policy guidelines in order for such systems to remain human-centric, serving humanity's values and ethical principles."

Thus, it is this goal of building ethics into AI [60, 61] that we subsume under the final *R*; the term "ethics" is assumed to be defined herein as *a normative practical philosophical discipline of how one should act towards others* [62]. We note that while the scope of *realism* emphasizes on intimate human and machine cooperation, *responsibility* represents an over-arching concept that must be integrated into all levels of an AI system.

As previously mentioned, an astonishing outcome of modern AI technologies has been the ability to efficiently learn complex patterns from large volumes of data, often leading to performance levels that exceed human limits. However, not so surprisingly, it is their remarkable strength that has also turned out to be a matter of grave unease; dystopian scenarios of robots taking over the world are being frequently discussed nowadays [63]. Accordingly, taking inspiration from the fictional organizing principles of Isaac Asimov's robotic-based world, the present-day AI research community has begun to realize that machine ethics play a central role in the design of intelligent autonomous systems that are designed to be part of a larger ecosystem consisting of human stakeholders.

That said, clearly demarcating what constitutes ethical machine behavior, such that precise laws can be created around it, is not a straightforward affair. While existing frameworks have largely placed the burden of codifying ethics on AI developers, it was contended in [61] that ethical issues pertaining to intelligent systems may be beyond the grasp of the system designers. Indeed, there exist several subtle questions spanning matters of privacy, public policy, national security, etc., that demand a joint dialogue between, and the collective efforts of, computer scientists, legal experts, political scientists, and ethicists [64]. Issues that are bound to be raised, but are difficult (if not impossible) to objectively resolve, are listed below for the purpose of illustration.

(i) In terms of privacy, to what extent should AI systems be allowed to probe and access one's personal data from surveillance cameras, phone lines, or emails, in the name of performance customization?

(ii) How should policies be framed for autonomous vehicles to trade-off a small probability of human injury against near certainty of major material loss to private or public property?

(iii) In national security and defense applications, how should autonomous weapons comply with humanitarian law while simultaneously preserving their original design objectives?

Arriving at a consensus when dealing with issues of the aforementioned type will be a challenge, particularly because ethical correctness is often subjective, and can vary across societies and individuals. Hence, the vision of building ethics into AI is unquestionably a point of significant urgency that demands worldwide research investment.

In conclusion, it is important to note that the various concepts introduced from R_1 (rationalizability) to R_4 (realism) cumulatively serve as stepping stones to attaining greater responsibility in AI, making it possible for autonomous systems to function reliably and to explain their actions under the framework of human ethics and emotions. In fact, the ability to do so is necessitated by a "right to explanation", as is implied under the European Union's General Data Protection Regulation [65].

REFERENCES

[1] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117.

[2] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

[3] Stanley, K. O., Clune, J., Lehman, J., & Miikkulainen, R. (2019). Designing neural networks through neuroevolution. *Nature Machine Intelligence*, *1*(1), 24-35.

[4] Ferrucci, D. A. (2012). Introduction to "This is Watson". *IBM Journal of Research and Development*, *56*(3.4), 1-1.

[5] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484.

[6] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... & Lillicrap, T. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140-1144.

[7] Sandholm, T. (2017, August). Super-Human AI for Strategic Reasoning: Beating Top Pros in Heads-Up No-Limit Texas Hold'em. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 24-25).

[8] Szathmáry, E., Rees, M. J., Sejnowski, T. J., Nørretranders, T., & Arthur, W. B. (2018). 10. Artificial or Augmented Intelligence? The Ethical and Societal Implications. *Grand Challenges For Science In The 21st Century*, *7*, 51.

[9] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 80-89). IEEE.

[10] Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

[11] Zeng, Z., Miao, C., Leung, C., & Jih, C. J. (2018). Building More Explainable Artificial Intelligence with Argumentation. In *Proceedings of the Twenty-Third AAAI/SIGAI Doctoral Consortium* (pp. 8044-8045).

[12] Cooper, G. F., Aliferis, C. F., Ambrosino, R., Aronis, J., Buchanan, B. G., Caruana, R., ... & Janosky, J. E. (1997). An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial intelligence in medicine*, *9*(2), 107-138.

[13] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721-1730).

[14] Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., ... & Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318-2331.

[15] Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M., & Kumar, V. (2019, May). Physics Guided RNNs for Modeling Dynamical Systems: A Case Study in Simulating Lake Temperature Profiles. In *Proceedings of the 2019 SIAM International Conference on Data Mining* (pp. 558-566). Society for Industrial and Applied Mathematics.

[16] Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452.

[17] Gal, Y., & Ghahramani, Z. (2016, June). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning* (pp. 1050-1059).

[18] Gal, Y., & Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems* (pp. 1019-1027).

[19] Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 427-436).

[20] Akhtar, N., & Mian, A. (2018). Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access*, *6*, 14410-14430.

[21] Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*..

[22] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust Physical-World Attacks on Deep Learning Visual Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1625-1634).

[23] Carlini, N., & Wagner, D. (2018, May). Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)* (pp. 1-7). IEEE..

[24] Sankaranarayanan, S., Jain, A., Chellappa, R., & Lim, S. N. (2018). Regularizing deep networks using efficient layerwise adversarial training. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (pp. 4008-4015). AAAI Press.

[25] Ross, A. S., & Doshi-Velez, F. (2018). Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (pp. 1660-1669). AAAI Press.

[26] Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016, May). Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)* (pp. 582-597). IEEE.

[27] Akhtar, N., Liu, J., & Mian, A. (2018). Defense against Universal Adversarial Perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3389-3398).

[28] Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning* (pp. 274-283).

[29] Biggio, B., Nelson, B., & Laskov, P. (2012, June). Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning* (pp. 1467-1474).

[30] Zhao, M., An, B., Gao, W., & Zhang, T. (2017, August). Efficient label contamination attacks against black-box learning models. In *Proceedings of the IJCAI* (pp. 3945-3951).

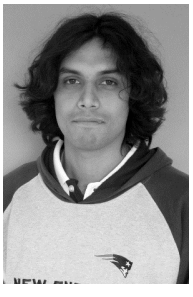
[31] Steinhardt, J., Koh, P. W. W., & Liang, P. S. (2017). Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems* (pp. 3517-3529).

[32] Smith, V., Chiang, C. K., Sanjabi, M., & Talwalkar, A. S. (2017). Federated multi-task learning. In *Advances in Neural Information Processing Systems* (pp. 4424-4434).

[33] Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science (New York, NY)*, *359*(6377), 725-726.

[34] Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., Olds, J. L., & Dean, H. (2015). Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science.
[35] Klein, A., Christiansen, E., Murphy, K., & Hutter, F. (2018). Towards Reproducible Neural Architecture and Hyperparameter Search. URL: https://openreview.net/pdf?id=rJeMCSnml7
[36] Loshchilov, I., & Hutter, F. (2016). CMA-ES for hyperparameter optimization of deep neural networks. In *Proceedings of ICLR 2016 Workshop*.
[37] Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., ... & Hodjat, B. (2019). Evolving deep neural networks. In *Artificial Intelligence in the Age of Neural Networks and Brain Computing* (pp. 293-312). Academic Press.
[38] Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & De Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE*, *104*(1), 148-175.
[39] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems* (pp. 2962-2970).
[40] van Rijn, J. N., & Hutter, F. (2018, July). Hyperparameter importance across datasets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2367-2376). ACM.
[41] Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*(1), 9.
[42] Da, B., Ong, Y. S., Gupta, A., Feng, L., & Liu, H. (2019). Fast transfer Gaussian process regression with large-scale sources. *Knowledge-Based Systems*, *165*, 208-218.
[43] Caruana, R. (1997). Multitask learning. *Machine learning*, *28*(1), 41-75.
[44] Gupta, A., Ong, Y. S., & Feng, L. (2018). Insights on transfer optimization: Because experience is the best teacher. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *2*(1), 51-64.
[45] Yogatama, D., & Mann, G. (2014, April). Efficient transfer learning method for automatic hyperparameter tuning. In *Artificial Intelligence and Statistics* (pp. 1077-1085).
[46] Da, B., Gupta, A., & Ong, Y. S. (2018). Curbing negative influences online for seamless transfer evolutionary optimization. *IEEE Transactions on Cybernetics*, (99), 1-14.
[47] Gupta, A., Ong, Y. S., & Feng, L. (2016). Multifactorial evolution: toward evolutionary multitasking. *IEEE Transactions on Evolutionary Computation*, *20*(3), 343-357.
[48] Bali, K. K., Ong, Y. S., Gupta, A., & Tan, P. S. (2019). Multifactorial Evolutionary Algorithm with Online Transfer Parameter Estimation: MFEA-II. *IEEE Transactions on Evolutionary Computation*.
[49] Swersky, K., Snoek, J., & Adams, R. P. (2013). Multi-task bayesian optimization. In *Advances in neural information processing systems* (pp. 2004-2012).
[50] Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press.
[51] Gupta, A., & Ong, Y. S. (2019). Memetic Computation: The Mainspring of Knowledge Transfer in a Data-Driven Optimization Era. Part of Springer's *Adaptation, Learning, and Optimization* book series (volume 21).
[52] Gundersen, O. E., & Kjensmo, S. (2017). State of the art: Reproducibility in artificial intelligence. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence and the Twenty-Eighth Innovative Applications of Artificial Intelligence Conference*.
[53] Kurshan, B. (2016). The future of artificial intelligence in education. *Forbes Magazine, New York Google Scholar*.
[54] Verhulst, S. G. (2018). Where and when AI and CI meet: exploring the intersection of artificial and collective intelligence towards the goal of innovating how we govern. *AI & SOCIETY*, *33*(2), 293-297.
[55] Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, *1*(1), 33-36.
[56] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98-125.
[57] D'mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, *47*(3), 43.
[58] Morency, L. P., Mihalcea, R., & Doshi, P. (2011, November). Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces* (pp. 169-176). ACM.
[59] Smith, M. J., Wedge, R., & Veeramachaneni, K. (2017, October). FeatureHub: Towards collaborative data science. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 590-600). IEEE.IEEE Global Initiative. (2016). Ethically Aligned Design. *IEEE Standards v1*.
[60] Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (pp. 5527-5533).
[61] Anderson, M., & Anderson, S. L. (2014, July). GenEth: A General Ethical Dilemma Analyzer. In *AAAI* (pp. 253-261).
[62] Cointe, N., Bonnet, G., & Boissier, O. (2016, May). Ethical judgment of agents' behaviors in multi-agent systems. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (pp. 1106-1114). International Foundation for Autonomous Agents and Multiagent Systems.
[63] Deng, B. (2015). Machine ethics: The robot's dilemma. *Nature News*, *523*(7558), 24.
[64] Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, *36*(4), 105-114.
[65] Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, *38*(3), 50-57.

Yew-Soon ONG received the PhD degree for his work on Artificial Intelligence in complex design from the Computational Engineering and Design Center, University of Southampton, UK in 2003. He is currently the President's Chair in Computer Science at Nanyang Technological University (NTU), and holds the position of Chief Artificial Intelligence Scientist of the Agency for Science, Technology and Research (A*STAR) Singapore. At NTU, he also serves as the Director of the Data Science and Artificial Intelligence Research Center, Director of the Singtel-NTU Cognitive & Artificial Intelligence Joint Lab, and Director of the A*STAR SIMTech-NTU Joint Lab on Complex Systems. His research interest in computational intelligence spans across memetic computing, complex design optimization, and big data analytics. He is the founding Editor-in-Chief of the IEEE Transactions on Emerging Topics in Computational Intelligence, Associate Editor of the IEEE Transactions on Evolutionary Computation, the IEEE Transactions on Neural Networks & Learning Systems, the IEEE Transactions on Cybernetics, and others.



Abhishek GUPTA received the PhD degree in Engineering Science from the University of Auckland, New Zealand, in 2014. He is currently a Scientist in the Singapore Institute of Manufacturing Technology, a research institute within the Agency for Science, Technology and Research (A*STAR) Singapore. He has previously served as a Research Scientist at the School of Computer Science and Engineering, Nanyang Technological University. Abhishek has diverse research experiences in computational science, ranging from numerical methods in engineering physics to topics in computational intelligence. Currently, his main research interests lie at the intersection of evolutionary computation and machine learning, with the aim of building efficient algorithms for design optimization.