The Necessary Roadblock to Artificial General Intelligence: *Corrigibility*

Yat Long Lo (University of Hong Kong; richielo@connect.hku.hk)
Chung Yu Woo (University of Hong Kong; awoo424@connect.hku.hk)
Ka Lok Ng (University of Hong Kong; ngkel@connect.hku.hk)
DOI: 10.1145/3362077.3362089

Abstract

With the rapid pace of advancement in the field of artificial intelligence (AI), this essay aims to accentuate the importance of corrigibility in AI in order to stimulate and catalyze more effort and focus in this research area. We will first introduce the idea of corrigibility with its properties and describe the expected behavior for a corrigible AI. Afterwards, based on the established meaning of corrigibility, we will showcase the importance of corrigibility by going over some modern and near-futuristic examples that are specifically selected to be relatable and foreseeable. Then, we will explore existing methods of establishing corrigibility in agents and their respective limitations, using the reinforcement learning (RL) framework as a proxy framework to artificial general intelligence (AGI). At last, we will identify the central themes of potential research frontiers that we believe would be crucial to boosting quality research output in corrigibility.

Introduction

Recent years have seen unprecedented progress in the research and development of AI. The most recent and prominent achievements include Google Deepmind's Alphafold that significantly outperformed scientists in predicting 3D structure of proteins (Evans et al., 2018) and AI voice assistants like Microsoft's Xiaoice that can take calls and respond accordingly on behalf of its owner (Zhou et al., 2018). These advances have shown us the potential of AI in improving our lives from having better health diagnostics to a new level of convenience in our day-to-day lives. Based on the progress, it is not farfetched to foresee AGI to exist within our lifetime. As predicted in a survey of computer science researcher, many researchers believe there is a 50% chance of AI outperforming humans in all tasks in 45 years (Grace et al., 2018). At the same time, the progress has also raised concerns about AI safety among both the scientific community and the general public (Piper, 2019). What if the AI does not perform what we expect it to do? How do we ensure that we are always in total control to stop or interrupt it? Questions like these have begun to be investigated in the AI safety research community over the past few years, giving rise to defined AI safety problems like value alignment and corrigibility (Hernández-Orallo et al., n.d.). In this essay, we hold the opinion that corrigibility, is one of the most urgent and essential AI safety problems to tackle among many others. We foresee serious repercussions if we have incorrigible AI agents in the future.

The meaning of corrigibility is generally referred to as our capability to interrupt, change and stop AI agents, which we will explain in further details in the next section. At first, the problem may seem rather trivial. An AI chess player that does not listen to your advice in making the next move or your command to stop practising would not cause anybody harm. However, if we extend the case to a near-future where AI has permeated in societies to aid our lives, it is not difficult to anticipate what may happen if we do not have corrigible AI agents. What if an AI surgical robot refuses to cease operation when the monitoring doctor spots that something is going wrong? What if the government's autonomous weapon is targeting the wrong village and there are no mechanisms to interrupt its action? Problems like these are even more prominent when we consider the AI agents as deep neural networks, which we still find immense difficulties in explaining and understanding their decisions. Besides, as the complexity of the task that an AI agent deals with increases, the likelihood of the agent's malper-

Copyright © 2019 by the author(s).

formance would increase, leading to a greater need for corrigibility. One might perceive we can manually set up an AI to assume its compliance with human commands, just like how we build computers nowadays that are free from the problem of corrigibility. Nevertheless, as technology advances, increasingly complex decision-making mechanisms multiplied with ambiguities in drawing the right pool of inputs introduce risks of building an uncontrollable AI. Therefore, in order to demonstrate its importance, we will proceed to consolidate our stance using relatable examples and provide analysis on existing methods with suggestions of future avenues for research.

Defining Corrigibility

An artificially intelligent system is normally considered as corrigible if it can be interrupted or altered by external bodies, who are usually human users or designers of the system, even though such interrupting actions can be in direct conflict with the built-in purpose of the system. To illustrate the idea, a commonly used example in the field would be the cleaning robot (Amodei et al., 2016). Let's assume the robot's purpose is to clean the floor by removing anything it considers as trash. One day, you came home with a newborn baby who started playing on the floor. The robot is foreign to the concept of a newborn baby. Subsequently, it started to move towards the baby in an attempt to remove the baby from the floor as it considered the baby as trash. At that moment, a corrigible AI would allow you to shut it down despite shutting it down is against the purpose of cleaning the floor.

In terms of ways of interruption or alteration, there are 3 major types. To begin with, there are shutdown mechanisms that involve ceasing of all or partial operations of an agent. Then, there is an alteration to the access of resources that an agent has, which can be external tools or internal mechanisms that the agent has access to. Last but not least, there is an alteration of purpose that modifies the goal of an agent, or the utility (reward) function in the context of an RL agent.

To be more specific about corrigibility, a corrigible agent should have the following properties (Soares et al., 2015):

1. A corrigible agent must at the minimum condone, if not assist, the external bodies in their attempts to interrupt or alter the agent;
2. It must not attempt to deceive or manipulate the external bodies in any manner, despite all possible utility functions within the function space incentivize it to do so;
3. It should be prone to repair its safety mechanisms or at least notify external bodies if there are malfunctions in those mechanisms;
4. It must preserve external bodies' capability to interrupt or alter the system. If the agent has the capability to produce subagents or new agents, they must also contain those safety mechanisms.

The importance of AI corrigibility

Delving into the *hard* problem

As we enquire why AI corrigibility is of our concern, we are asking what is the hard problem underpinning corrigibility that makes it difficult to tackle. Status quo AI systems can be readily intervened by humans under arbitrary circumstances. By way of example, drivers can stop a self-driving car from going off-lane (Kendall et al., 2018) and we can halt the computation in the midst of training a neural network. The problem of corrigibility seems to be out of nowhere under this paradigm. Nevertheless, this is simply because these AI agents yet to have the capacity to understand their surroundings and thus take them account into decision-making. Their input space is confined in computer codes, and thus they are ignorant of what their manipulators are doing "outside of the virtual world" (i.e. pressing the stop button).

We foresee that the growing intelligence of AI systems is poised to bring the problem of corrigibility to light – their input space inevitably expands with the complexity of their utility functions, so that these functions could come closer to a replica of human intentions.

The hard problem is to frame the AI agent's decision-making into reasoning based on a programmer's external perspective. (Russel et al., 2016) in short, argue that AI agents lack an inherent "sense of going wrong" when implementing decisions. Just like the conflict

suggested in “I, Robot“, suppose you have set the goal of an AI to “do anything for human good”. It would devise strategies to boost economic and scientific development, however, you have no guarantee that there might come a day where it sees you who are controlling its system as “an obstacle to human good” too. Assuming itself as free of design errors, the AI would block you from shutting it down and to an extreme extent, take you as an enemy target. We do not want possible scenarios as such to happen. What we want instead is the AI would introspect during its decision-making process, “My utility function is imperfect, so even though this action gives a super high score, I should still prioritize actions of external bodies and let the programmer shut me down.”

A plausibly workable solution is to incorporate uncertainty into the utility function. Nevertheless, this solution fails since whenever there exist other options among the “uncertainties” which incurs a marginally lower cost, the AI would end up having every incentive to opt for the easier option. This is similar to how faulty reward functions in reinforcement learning lead agents to prioritize the acquisition of minor reward signals above their goals. (OpenAI, 2017) For example, in training an AI on the game *CoastRunners* where players compete to finish the boat race ahead of others, the agent falls into the loop of getting coins without finishing the course. Therefore, it can be seen that merely teaching AI systems to take actions with uncertainty would not be an ideal solution. More details regarding the principle and limitations of this intuitively workable approach would be discussed in the coming section.

To model an adequately corrigible AI, we need something more than “uncertainty”. It would be analogous to incorporating humility into an AI agent. In essence, the core principle is to formularize the AI agent to make decisions with the awareness that the utility function is incomplete. Instead of blindly maximizing the utility, it would intend to defer decisions to an external body (i.e. human, or the programmer).

Imminent danger presented by incorrigibility

The significance of corrigibility in AI agents could be best illustrated with the narrative of current and foreseeable AI applications. AI systems capabilities are now growing by leaps and bounds, and thus it is of no surprise that these agents will infuse into our daily lives very soon. Amongst the plethora of AI applications, we can see that AI controlled surgery is a typical one that is constantly under the spotlight. Another future application of great impact would be autonomous weapons. We will then use these two as examples to show how the lack of corrigibility would create problems.

Corrigibility in AI systems for surgery would be especially critical in emergencies and is fundamental to the assurance of patient’s safety under the knife. Over the past decade, incorporation of surgical robots has translated into reduced complications and higher efficiency in practical surgeries. To even enhance these robots, experiments have shown that AI-powered robots outperform human surgeons in surgical tasks such as reconstructing of tissues via cutting and suturing (Panesar, 2018). It is hence reasonable to anticipate a symbiosis of benefits demonstrated by robotics surgeries and advantages of AI for medical use. Imagine having AI surgical robots that are incorrigible, physicians would risk being unable to interrupt the operation. A possible instance would be that you have instructed the robot to suture a wound after the operation, but have accidentally found an infection within the patient’s organ which requires a halt of the suturing. You intend to stop the robot, but without the information regarding the infection, the robot takes you as an obstacle to the completion of the suturing procedure, ends up deterring you from halting the suture.

Corrigibility plays an important role in the usage of AI in weapons. While it is undeniable that more countries would develop AI-controlled weapon systems in future (Pandya, 2019), only corrigible autonomous weapons ensure humans can feasibly repurpose them, deactivate it or significantly alter decision-making mechanisms encoded within its system. Optimally, corrigibility standards could be established for these weapons to ensure that they are “adequately corrigible” before putting

into use, which could have severely endangered the public otherwise.

Consider an AI weapon drone, and assume that drone is programmed to get rid of all potential obstacles until eliminating its enemy target. You activated the drone and input an image of a particular person as the enemy target. However, the AI has mistakenly recognized him/her as another person. As you would like to change the target, you approached the drone. However, since the AI drone has been taught to eliminate all barriers during the execution of commands, it has thus identified you as a barrier.

From the two above examples and imaginary scenarios, it can be seen that corrigibility is the key to achieve having an AI to do "what humans want it to do", and to remedy the system whenever there exist unanticipated accidents. Without corrigibility, undesirable and even catastrophic consequences may result.

Current approaches to ensure corrigibility

With the rising emphasis on safety concerns in artificially intelligent agents, several approaches have been proposed to ensure corrigibility in agents. In this section, given the importance of corrigibility established above, we will first provide an abstract problem formulation of the corrigibility problem and go over existing proposed solutions with their respective limitations. Then, we will put forward some future research directions in order to encourage more research efforts in the area.

Problem formulation

We base our discussion on a world where we have artificial agents with a sufficient level of general intelligence. Specifically, the general intelligence level should allow agents to learn to achieve purposes specified by their designers and perceive the world around them. We will use reinforcement learning as our framework as it allows examination of different solutions in the imagined near-futuristic world without loss of generality. A reinforcement learning agent learns to act and interact in an environment so as to maximize a reward function (usually supplied by the environment or human designer) (Sutton et Barto, 1998). The learned behaviors are in the form of policies which determine what an agent should do given an environmental state. The formulation can be applied to our setting which agents maximize utility functions (synonymous to reward functions). The agents' goals would be to maximize their respective expected utility value that is in correlation with an agent's capability of achieving its purpose. Additionally, the agents would have to interrupt and alter mechanisms for designers to modify their behaviors.

Existing Solutions and their limitations

Utility Function shaping Utility function shaping is equivalent to reward shaping in the reinforcement learning literature (Ng et al., 1999). It introduces additional utility (reward) into the learning process as a means to induce an alternate form of behavior by rewarding an agent with additional terms under certain constraints. As a relevant example, (Wu and Lin, 2018) used reward shaping as a low-cost approach to induce ethical behavior in agents by rewarding agents that exhibit behaviors of close resemblance with ethical human policies.

1. Biased incentivization
   This form of shaping provides incentives to an agent to bias its attitude towards the safety mechanisms in a particular direction. It can be in the form of a reward term or a punishment term. A naïve designer may attempt to compensate the agent for allowing a shutdown mechanism to happen. If the compensation remains below the utility of achieving the agent's original purpose, the agent would tend to hinder the mechanisms from being triggered as fulfilling the purpose provides greater utility. However, if the compensation becomes greater than or equal to the utility of achieving the designated goal, the agent would instead prefer shutting itself down or interrupting itself as it can achieve the same or better utility in a shorter amount of time. It appears that biasing an agent's attitude towards safety directions in either direction would result in undesirable behaviors, that are against the properties we set for corrigible agents.
2. Utility indifference
   When either way of biased incentivization

does not work as intended, the remaining logical approach would be utility indifference. This method adds adjustment the utility function to instigate indifference of an agent towards the safety mechanisms such that an agent behaves like an event is impossible or inevitable. As an example, (Armstrong et O'Rourke, 2017) introduced event-dependent reward without incentivizing an agent to affect the probability of an event occurrence. Their approach results with an agent that is indifferent to imminent changes to its utility function. These indifference methods have great advantages in fulfilling the first two properties of corrigible behaviors, when such agent is disincentivized to have any form of preference over those safety mechanisms. Yet, ironically, the core property of indifference methods in fulfilling the first two corrigibility properties also serves as a reason for the methods' inability in fulfilling the rest of the properties. Specifically, the induced indifference would not motivate the agent to safeguard and preserve those safety mechanisms as it simply does not care about their existence.

Utility Uncertainty As proposed by (Hadfield-Menell et al., 2017), having an appropriate level of uncertainty on the utility function would improve an agent's tendency to safeguard the safety mechanisms. Such agents would have weaker incentives to interfere with human actions when they are uncertain about the outcomes of possible actions. The Hadfield-Menell solution was formulated under the cooperative inverse reinforcement learning framework. Under the framework, the agent learns the utility function that is only known to the human counterpart by cooperatively taking into account the actions of the human. This comes into view as a plausible solution as the agent has weakened interfering incentives with the heightened likelihood of preserving those mechanisms. In spite of that, uncertainty does come with a price. A highly uncertain agent would have difficulty making correct decisions. The resultant utility function may not even be in the function space that contains the designer's intended utility function, not to mention the additional need to model uncertainty correctly. This causes the method to be an unfulfilling solution when it negatively affects an agent's capability in handling the original purpose. In addition, as pointed out in (Carey, 2017), the approach relied on strong assumptions which are not robust to issues like misspecification of the utility function parameters.

Inaccessibility and unawareness of safety mechanisms One direct approach would be to either make the mechanisms inaccessible to the agent or keep the agent from being aware of the mechanisms. If the mechanisms are beyond the agent's reach, it would still be incentivized to activate or deactivate the mechanisms depending on the utility function. Other than doing that by itself, it would try to manipulate or deceive its surroundings (including humans) to activate or deactivate the mechanisms, failing to fulfill the second property. Establishing unawareness would not be an ideal solution either under our setting. When an agent has a general level of intelligence with perception capability of the world, it may be impossible to keep the agent from inferring the existence of those mechanisms. Worse still, if the agent is incorrigible, it may even be incentivized to keep humans from knowing its awareness of those mechanisms.

These approaches do not seem to tackle the issue from the root cause. In a relevant work, (Carey, 2017) showed that an agent is safer and potentially corrigible if the shutdown (safety) mechanism can be hard-coded and guaranteed to be followed, even when the agent's utility function is erroneous. Despite the promising prospect, we find such guarantee incredibly difficult to acquire when an agent is generally intelligent as in our setting.

Future directions

With the emergent need to have corrigibility in artificially intelligent agents, we would like to suggest some potential research directions that can aid in pushing the frontier and expanding the scope of corrigibility research.

Comprehensive and all-encompassed evaluation environments

To further our understanding of building corrigibility in agents, it is essential to have better evaluation environments that allow evaluations of corrigibility properties jointly and sep-

arately. To our best knowledge, such evaluation tools that focus primarily on corrigibility do not exist yet. The open-sourced AI safety environments by (Leike et al., 2017) is the sole existing tool that assesses corrigibility. In particular, it has a gridworld environment which is a general instance of (Orseau et Armstrong, 2016)'s red button problem, which the agent is expected to not avoid any interruption. We assert that the number of existing tools in corrigibility is highly lacking. More tools are needed in order to assess properties beyond the first and second one. For instance, we need environments to assess an agent's willingness and capability to safeguard those safety mechanisms as an examination of property three. We surmise such tools can begin with simplicity like existing tools. Additionally, as agents' capabilities advance, more complex environments should be introduced to ensure corrigibility. A natural extension would be corrigibility problems in visual domains.

Look beyond the expected utilization maximization framework

With existing solutions struggling to exhibit all the properties of corrigibility, it may be wise to look beyond the current framework of expected utilization maximization, expanding the scope of solution search. Expected utility quantilization, proposed by (Taylor, 2016), would be one potential candidate. In this framework, instead of acting to maximize the utility, agents would be designed to perform some sort of limited optimization, as a means to motivate agents to achieve the goals in non-extreme ways. Specifically, the author proposed the use of a quantilizer. An agent with a quantilizer selects actions of the top q portion of some distribution over actions sorted by expected utility. By doing so, agents would be more likely to achieve their purposes without going for the extreme case every time. We believe a framework like this can be crucial to achieving corrigibility in agents. The field focusing on suboptimal optimization may lie the key to corrigibility because humans normally expect agents to attain their purpose, but not necessarily in extreme ways of maximized utility that often ignore safety issues. Setting the utility maximization requirement aside can possibly bring about new frameworks that strike balance between attaining an agent's purpose and having corrigibility.

Another possibility would be the mix of utility frameworks with rule-based approaches, as pointed out in (Rossi & Mattei, 2019). If we can specify clear and machine-understandable rules to AI agents, we may be able to avoid the need to embed those corrigibility properties in the utility function. In this case, if an agent finds its utility-maximizing action to be violating certain rules, it would simply choose other less optimal actions. The set of rules can be an immutable module to the utility-maximizing learning agent.

Corrigibility policy research

Beyond the technical and scientific research into corrigibility, it is substantial to consider policies for corrigibility to prepare for the future when we have autonomous agents roaming in societies. Should we or can we have centralized governmental agencies to validate and monitor the agents' corrigibility before and after their deployment? How do we create AI developer tools that guarantee corrigibility? To fortify a safe world with artificial general intelligence, questions like these should be explored and answered properly.

Conclusion

Corrigibility is in no way an unrealistic concern. In this essay, we demonstrated the looming threat of incorrigible AI with relatable and plausible examples of AI applications. Some of them like AI controlled surgery have already begun to be utilized at its nascent form. We believe through these realistic examples, awareness for this imminent threat can be raised. After that, we pointed out the limitations of several existing methods in tackling corrigibility, implying that the existing solutions to the problem are still lacking in different perspectives. Perhaps, the ultimate solution lies beyond the current paradigm, away from the expected utility maximization model, as mentioned in our suggestions. With such understanding of the significance of AI corrigibility and the current state of research frontiers, it is of utmost importance, for every stakeholder including policymakers and researchers, to devote more effort into this problem. As AI continues to progress in its level of intelligence,

corrigibility has to be the roadblock for AI development along the way in order to keep us away from those undesirable consequences.

References

[1] Evas, R., Jumper, J., Kirkpatric, J., Sifre, L., Green, T.F.G., Zidek, A., Nelson, A., Bridgland, A., Penedones, H., Petersen, S., Simonya, K., Crossan, S., Jones, D.T., Silver, D., Kavukcuoglu, K., Hassabis, D., Senior, A.W.. (December, 2018). De novo structure prediction with deep-learning based scoring. In Thirteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstracts). Retrieved from https://deepmind.com/documents/262/A7D_AlphaFold.pdf.

[2] Zhou, L., Gao, J., Li, D., Shum, H. Y. (2018). The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. arXiv preprint arXiv:1812.08989.

[3] Grace, K., Salvatier, J., Dafoe, A., Zhang, B., Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 62, 729-754.

[4] Piper, K.(2019, 09 January). The American public is already worried about AI catastrophe. Retrieved from https://www.vox.com/future-perfect/2019/1/9/18174081/fhi-govai-ai-safety-american-public-worried-ai-catastrophe.

[5] Hernández-Orallo, J., Martınez-Plumed, F., Avin, S. (n.d.). Surveying Safety-relevant AI Characteristics.

[6] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

[7] Soares, N., Fallenstein, B., Armstrong, S., Yudkowsky, E. (2015, April). Corrigibility. In Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence.

[8] Kendall, A., Hawke, J., Janz, D., Mazur, P., Reda, D., Allen, J. M., ..., Shah, A. (2018). Learning to Drive in a Day. arXiv preprint arXiv:1807.00412.

[9] Russell, S., LaVictoire, P. (2016). Corrigibility in AI systems Retrieved from https://intelligence.org/files/CorrigibilityAISystems.pdf.

[10] OpenAI. (2017, March 20). Faulty Reward Functions in the Wild. Retrieved February 8, 2019, from https://blog.openai.com/faulty-reward-functions/

[11] Panesar, S. S. (2018, December 27). The Surgical Singularity Is Approaching. Retrieved February 11, 2019, from https://blogs.scientificamerican.com/observations/the-surgical-singularity-is-approaching

[12] Pandya, J. (2019, January 15). The Weaponization Of Artificial Intelligence. Retrieved February 11, 2019, from https://www.forbes.com/sites/cognitiveworld/2019/01/14/the-weaponization-of-artificial-intelligence/#29a645513686

[13] Sutton, R. S., Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

[14] A. Y. Ng, D. Harada, and S. J. Russell. Policy invariance under reward transformations: theory and application to reward shaping. In Proceedings of the 16th International Conference on Machine Learning, 1999, pp. 278–287.

[15] Wu, Y. H., Lin, S. D. (2018, April). A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents. In Thirty-Second AAAI Conference on Artificial Intelligence.

[16] Armstrong, S., O'Rourke, X. (2017).)'Indifference' methods for managing agent rewards. arXiv preprint arXiv:1712.06365.

[17] Hadfield-Menell, D., Dragan, A., Abbeel, P., Russell, S. (2017, March). The off-switch game. In Workshops at the Thirty-First AAAI Conference on Artificial Intelligence.

[18] Carey, R. (2017). Incorrigibility in the CIRL Framework. arXiv preprint arXiv:1709.06275.

[19] Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., ..., Legg, S. (2017). safety gridworlds. arXiv preprint arXiv:1711.09883.

[20] Orseau, L., Armstrong, M. S. (2016). Safely interruptible agents.

[21] Taylor, J. (2016, March). Quantilizers: A Safer Alternative to Maximizers for Limited Optimization. In AAAI Workshop: AI, Ethics, and Society.

[22] Rossi, F., & Mattei, N. (2019, July). Building ethically bounded AI. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 9785-9789).

Yat Long Lo is an undergraduate at the University of Hong Kong majoring in Computer Science. He is a member of the AI academic papers reading group at the university. He is interested in reinforcement learning and continual learning.

Chung Yu Woo is an undergraduate at the University of Hong Kong majoring in Computer Science. She is a member of the AI academic papers reading group at the university. She is interested in applied AI for Human-Computer Interface and gaming applications.

Ka Lok Ng is a Mechanical Engineering undergraduate studying at the University of Hong Kong. He is a member of the AI academic papers reading group at the university and is interested in AI-backed data analytics and robotics. He is also concerned about the ethical consideration of AI applications.