Corrigibility in AI systems

Abstract

Sufficiently advanced artificially intelligent systems may present a number of security concerns not present in less advanced systems. If a system is constructed to pursue a set of goals that are not the goals its operators intended, the system will by default have incentives to deceive, manipulate, or resist the operators to prevent them from altering its goals, as that would interfere with its ability to achieve its current goals. We will do research on incentive systems and goal structures that would avoid these dangerous default incentives. We will develop a theoretical framework and a software engineering methodology for allowing runtime modification of a utility-theoretic AI system, in ways that do not give the system incentives to deceive, manipulate, or resist, while ensuring that appropriately authenticated humans retain full control of the system. AI systems that allow this sort of modification without having adverse incentives are known as "corrigible systems." Learning how to construct corrigible AI systems will help ensure that powerful autonomous AI systems do not become dangerous in the long term.

Key Personnel

Stuart Russell. Principal Investigator. University of California at Berkeley. (russell@berkeley.edu)

Patrick LaVictoire. Machine Intelligence Research Institute. (patrick@intelligence.org)

Deliverables

- A theoretical framework for modeling agents that are embedded within their environment and can take advantage of this fact.
- Theorems which, in that framework, describe situations under which corrigible behavior can be guaranteed, and under what assumptions.
- A software engineering methodology, complete with instantiated software templates, demonstrating how to design corrigible AI systems.

In the future, advanced artificially intelligent (AI) computer systems may well present novel security concerns that do not appear in less advanced AI systems. Consider a system capable of building accurate models of itself and its human operators. If the system is constructed to pursue some set of goals that its operators later realize will lead to undesirable behavior, then the system will by default have incentives to deceive, manipulate, or resist its operators to prevent them from altering its current goals (as that would interfere with its ability to achieve its current goals). This fact holds both in the case of operator error (the operator entered the wrong goal) and operator ignorance (the operator entered "cure cancer," and the system started kidnapping test subjects).

Moreover, sufficiently powerful AI systems might have the ability to resist being shut down or modified by their operators. One might expect that the system can always be "unplugged," but this intuition is misleading in a world where many programs are distributed across a wide variety of machines in different locations and connected to global networks. A highly advanced autonomous system may be difficult to shut down or modify without the system's cooperation. Ensuring that we can build systems which do in fact cooperate with their operators will be vitally important as autonomous AI systems are given more and more responsibility and control. We refer to agents that have no incentives to manipulate, resist, or deceive their operators as "corrigible agents," using the term as defined by Soares et al. (2015). We propose to study different methods for designing agents that are in fact corrigible.

We propose to work on this problem in three phases. First, we will need to develop a framework for studying the corrigibility of agents, as existing tools in the field of AI are not up to the task. Second, we will develop formal results concerning methods that can be used to guarantee corrigibility; this may also yield some impossibility results. Third, we will develop a software engineering methodology for runtime modification of the behavior of advanced AI systems without giving the system incentives to resist. This will take the form of software templates which demonstrate that this is possible in simple scenarios and will yield insight into how realistic corrigible systems could be built.

Modeling Embedded Agents

At present, corrigibility in AI systems is very difficult to study, for a few reasons. It is standard in the field of AI to study the design of agents which act like utility maximizers with respect to some fixed utility function (Russell and Norvig, 2010, chap. 16). But when designing corrigible agents, utility maximization is part of the problem. We either need to develop models of agents which pursue some set of goals in a "cautious" fashion, or we need to find ways of designing utility functions such that blind maximization of that utility function does not give the agent incentives to manipulate and deceive, which is no small feat (Omohundro, 2008; Bostrom, 2014). We want to study the construction of corrigible utility functions, and that requires moving beyond the framework where the utility function is taken as a given.

Furthermore, the standard models of artificial reasoners in the field of AI assume that the agent is fundamentally separated from the environment, and able to communicate only along certain limited input and output channels, as seen in, e.g., the "universal intelligence" framework of Legg and Hutter (2007). This is a very restrictive assumption when attempting to study corrigibility, where side-channels can be important. For example, the reinforcement learning framework assumes an agent/environment separation (see Sutton [1984]), whereby the source of the reward signal is assumed to exist *outside* the environment. In reality, the reward channel is *part* of the environment and it may be possible for the agent to find an unexpected action sequence that wrests control of the reward channel away from its operators; after that, it will deal itself maximum rewards at all times, while defending its reward channel from later interference.

If we are to study agents that may attempt to adversarially manipulate their operators, we need formal models that treat the AI system, its computational

substrate, the human operators, the reward channel, etc., as part of the environment. This would require novel but not unprecedented research; early work has been done by Orseau and Ring (2012) that nevertheless leaves the problem wide open. We will begin our research by developing formal models of agent/environment interactions when an agent is embedded within its environment, and when its very operation may produce side effects not captured by the usual input/output channels.

This is no small task. A fully satisfactory account of environment-embedded agents must handle a wide variety of cases that are very difficult to capture in the separated agent/environment framework. For example, we will need our model to be able to handle agents that are distributed across many computers, agents that construct a number of sub-agents to assist in various tasks, agents with unreliable hardware, and so on. In reality, there is no hard dividing line between an "agent" and the rest of the world, and the lines will only become only more blurred as distributed AI systems become more common. If we want our agents to remain corrigible even when they span broad, modifiable networks of insecure machines, we will need models that let us reason in some coherent fashion about the "actions" of "agents" even when the distinction gets blurry.

While our primary purpose in designing these tools will be to use them for purposes of studying a very specific set of safety and security concerns relating to advanced AI systems, we expect that these tools would be broadly useful to the field. Current frameworks for studying AI systems make it difficult to reason about the side effects of computation (such as the passage of time or the generation of heat) and nearly impossible to accommodate actions that modify the agent's own hardware or that create and program new sub-agents.

Formalizing Corrigibility

Given a theoretical framework that can describe environment-embedded agents, the next step is to give a formal definition of the problem. Bostrom (2014) and Soares et al. (2015) and others have argued intuitively that advanced AI systems would by default have incentives to manipulate and deceive; we would seek to quantify those claims. Under what conditions does an agent have incentives to behave incorrigibly? Under what conditions can it be made corrigible? Under what conditions is such a thing impossible?

Soares et al. (2015) have done some preliminary work in this direction. They consider a toy problem which they call the "shutdown problem," in which an agent is designed to pursue one utility function U_N (the "normal behavior" goal) before a shutdown button is pressed, and another utility function U_S (the "shutdown safely" goal) after the button is pressed. They show that this task in and of itself is trivial by giving the agent a "composite" utility function U which depends on the state of the button, but that it is much more difficult to construct U in such a way that the agent does not have incentives to cause or prevent the pressing of the button. (Intuitively, if the expectation of U_N is greater than the expectation of U_S then the agent has incentives to prevent the button from being pressed, and if the opposite is true then the agent has incentives to cause the button to be pressed.)

We will begin our theoretical analysis by extending the results of Soares et al. (2015) in a number of ways. For example, we would like to consider the shutdown problem in partially known and partially observable environments. We also expect to gain insight by re-examining this model in a framework where the agent is considered to be a part of the environment. This would give us the ability to analyze edge cases about which Soares et al. can only speculate. For example, consider situations in which an agent would generate incorrigible sub-agents, such as an automated scientist which is distributed across a number of computers. The system might safely shut down when instructed to do so by its operators, but this may still be unsafe, if the original agent leaves a number of (now uncontrolled) sub-agents in its wake, still running the laboratories, which did not obey the shutdown command. This is exactly the sort of scenario that requires a framework that treats agents as part of their environment.

Using an extended account of the shutdown problem, we will explore goal systems and incentive structures that do or don't allow corrigible behavior. The goal will be to prove that certain methods lead to agents that are completely corrigible in this simple scenario, or prove that such a thing is impossible.

From there, we will move to more complicated domains, and again provide proofs about what types of corrigibility succeed or fail in various scenarios. Our main goal is to provide methods that provably incentivize corrigible behavior in full generality, or prove that this is impossible.

This is an ambitious goal, and we have some ideas about where to begin. Considering the shutdown problem, it is apparent that a human would press the shutdown button if the utility function built into the agent is causing it to behave undesirably; thus, the agent must accommodate the idea that its utility function may not be "correct" and that the shutdown button provides evidence of incorrectness. In particular, an agent whose objective is to maximize the utility of the human, without knowing for certain what that utility actually is, may exhibit corrigible behavior in the desired sense. We will demonstrate that this is actually the case (or show it to be impossible), both in the single-agent setting where the human is considered as an exogenous element of the environment and in the game-theoretic setting where the agent and human are treated as acting jointly.

Even if an agent has learned what one might consider to be an admirable "idealist" model of human values, corrigibility problems may still arise. For example, a robot that has learned the goal of alleviating human suffering may decide that, rather than carrying out the latest command from its operators, it should instead travel to some distant country undergoing a humanitarian crisis. In one sense, this behavior seems laudable, but in another sense, it is surely not corrigible. This example is an instance of the general problem of designing agents that can interpret commands from human operators against the background of a general system of values. In this context authentication is also an issue—the agent must be convinced that the command is actually coming from its authorized operator, that the operator is not being coerced into giving the command, and so on.

The goal of this project will be to give a formal account of how to construct utility-theoretic AI systems that allow runtime modification of their goals, while ensuring that the appropriately authenticated operators remain in control of the system, while remaining corrigible. That is, the framework must demonstrate how a system can allow appropriately authenticated humans to override the system's utility function, alter its goals, or shut down the system entirely, all without giving the systems any incentives to manipulate, deceive, or resist the operators.

This theoretical framework would be directly useful for the theorems proved, which would indicate what types of AI systems could safely be directed towards different types of goals. Done correctly, this would affect the trajectory of the field of AI at large, by paving the way to the eventual development of powerful corrigible agents.

In addition, the theoretical framework would likely prove quite useful to the existing long-term AI alignment community, including researchers at the Machine Intelligence Research Institute in Berkeley (MIRI), the Future of Humanity Institute in Oxford (FHI), and the Centre for the Study of Existential Risk (CSER). For example, consider the problem of designing AI systems which do not attempt to perform significant self-modifications. This problem has a similar flavor to the problem of removing incentives to deceive or manipulate, as both self-modification and deception are instrumentally useful goals that we may not want the AI system to carry out for fear that the system will behave dangerously if it does so. Our theoretical frameworks could be of assistance to researchers studying these similar problems.

Developing a Software Engineering Methodology

The ultimate goal of our project would be not simply theorems, but also software. Insofar as we prove there are situations where incorrigible agents could be dangerous,

we will provide simple software examples, demonstrating that these situations are in fact plausible. Insofar as we can prove that certain systems would be corrigible, we will provide software templates demonstrating very simple corrigible agents, with the intention that these templates could be extended and adapted for use in practical systems. After all, the ultimate goal is to ensure that by the time AI scientists are capable of developing highly advanced AI systems which *could* resist their operators, they already have the tools in hand to build highly reliable, corrigible agents.

It is difficult to predict what sort of software engineering methodology will be required to design corrigible agents. As discussed above, designing utility-theoretic agents which allow their goals to be overridden or altered is no small task, and designing corrigible goal systems and the agents that pursue them may require a delicate balancing act between consequentialist reasoning and respecting operator control. We will aim to produce a handful of examples demonstrating that incorrigible agents have the potential to be dangerous, along with at least one software template demonstrating how a practical agent could be designed such that it exhibited corrigible behavior in some simple setting.

Ideally, we will develop an understanding of what types of utility function and agent design must be adhered to or avoided in order to achieve corrigible behavior in full generality, and we will provide a demonstration that could easily be extended for use in many different types of AI systems, regardless of the specific architecture of that AI system in particular, and regardless of whether the AI system is small and self-contained or large and distributed across a network. However, this should be considered a stretch goal. As of yet, it is not even clear whether it is possible in principle to design utility-theoretic AI systems which are corrigible in full generality. Depending on the results we get when formalizing corrigibility, the software templates may need to be limited to specific scenarios or specific subsets of corrigible behavior.

Security Implications

In the future, highly advanced autonomous AI systems could pose significant and unconventional security risks. This is especially true if those systems are capable of deceiving, manipulating, or resisting their operators. These risks bear on at least two separate CLTC priorities.

First, the development of highly advanced autonomous agents will revolutionize the dynamics of attack and defense, especially if incorrigible AI systems have strong incentives to resist or deceive their operators. Computer security is difficult in large part because defenders face *intelligent, adversarial* opposition from attacking humans. If our society begins creating incorrigible but highly intelligent AI systems, then we might find ourselves facing cyberattacks from *superintelligent* opposition. (By "superintelligent opposition" we mean opposition from AI systems that exhibit superhuman capabilities in cyberattack/defense, which some systems already can do.) Attacking *or* defending against a superintelligent adversary could be a hopeless task, and finding ourselves in such a situation would indicate a more fundamental failure in our approach to designing AI systems. Our goal is to learn how to build systems that have no incentives to resist in the first place; in this way we avoid further complicating the already fraught attack/defense landscape.

Second, the creation of incorrigible agents would give rise to the use of unconventional attacks. Unlike humans, AI systems are not predisposed to think along conventional lines or to design attacks based on modifications to one of a standard set of ideas. In addition, an incorrigible agent would be strongly motivated to use methods that remain undetected until the agent has achieved its goals, and to recruit other agents, including unwitting humans. Again, our research aims to nip the problem in the bud.

The study of artificial intelligence is progressing rapidly. Although it appears *likely* to most experts that autonomous AI systems capable enough to cause serious security problems are some distance away, it is *certain* that the exact distance is very hard to predict. It is quite important, we think, to begin the foundational research required to build *safe* AI systems as soon as possible. The proposed research is

intended to generate paradigmatic examples of theoretical results that will be useful in practice. If successful, it may provide some impetus for the broader community of AI researchers to consider more seriously the security implications of their research.

Stuart Russell will oversee the project and will conduct regular research meetings with the participants. The bulk of the research will be done by a postdoctoral researcher (75% time for two years). They will be primarily responsible for developing the initial model of environment-embedded agents, along with simple extended models of the "shutdown" and "idealist" problems described above. They will then study methods for achieving corrigible behavior, prove theorems about what is and is not possible, and provide software demonstrating the same. This will be done with regular assistance from Patrick LaVictoire at the Machine Intelligence Research Institute, who will spend 10% of his time for two years working with the postdoctoral student to ensure that the theoretical frameworks capture the core problems of corrigibility, and that the software methodology seems tractable and useful.

The three most important milestones in the first year are (1) Develop a minimal framework for studying environment-embedded agents capable of representing, e.g., the shutdown problem; we hope to achieve this by summer 2016. (2) Provide a formal extended model of the shutdown problem and the idealist problem, complete with analyses; we hope to achieve this by fall 2016. (3) Develop at least one positive and one negative example of a corrigible agent in a limited setting, demonstrated in software; we hope to achieve this by winter 2016. The milestones in the second year are difficult to predict, given their dependence on the theoretical results obtained in the first year, but plausible milestones include (4) finalize a satisfying model for environment-embedded agents; (5) prove fully general theorems about the situations in which corrigible behavior is and is not possible; and (6) provide a software template demonstrating how a generic corrigible agent could be designed, using methods that could apply to arbitrary practical systems.

The approximate budget for this project is as follows: $135,000 for two years of postdoctoral salary and benefits at 75% time; $30,000 for 2 half-summer-months of Stuart Russell's time; $23,000 for 10% of Patrick LaVictoire's time (salary and benefits) for two years; $4,000 for laptop and computer account fees for the postdoc; $8,000 for conference travel for presentation of results; no overhead. Total: $200,000 over two years. Additional partial support will come from Prof. Russell's projects in the related area of value alignment (funded by the Future of Life Institute and by DARPA Defense Sciences Office). For details, see attached spreadsheet.

Stuart Russell received his BA with first-class honours in Physics from Oxford in 1982, his PhD in Computer Science from Stanford in 1986, and then joined the faculty of the University of California at Berkeley. He is a Professor (and former Chair) of Electrical Engineering and Computer Sciences and holds the Smith–Zadeh Chair in Engineering. He is a fellow of AAAI, ACM, and AAAS; winner of the Computers and Thought Award and the ACM Karl Karlstrom Outstanding Educator Award; and holder from 2012 to 2014 of the Chaire Blaise Pascal and ANR senior Chaire d'excellence in Paris. His book "Artificial Intelligence: A Modern Approach" (with Peter Norvig) is the standard text in AI; it has been translated into 13 languages and is used in over 1300 universities in 116 countries. His research covers many areas of artificial intelligence, with a particular focus on machine learning, probabilistic modeling and inference, theoretical foundations of rationality, and planning under uncertainty. He also works for the United Nations, developing a new global seismic monitoring system for the nuclear-test-ban treaty. His current concerns include the threat of autonomous weapons and the long-term future of artificial intelligence and its relation to humanity.

Patrick LaVictoire received his BA with honors in Mathematics from the University of Chicago in 2005, and his PhD in Mathematics from the University of California at Berkeley in 2010. He was a Van Vleck Visiting Assistant Professor of Mathematics at the University of Wisconsin at Madison, then left academia to work on machine learning algorithms for the search startup Quixey. He joined the Machine Intelligence Research Institute in March 2015. He was the PI on a National Science Foundation grant in Analysis, and has published several papers in top mathematics journals on harmonic analysis and ergodic theory. His current research spans several underdeveloped fields relevant to long-term AI alignment, including decision theory, game theory, logical uncertainty, and multi-level modeling.

References

Bostrom, Nick (2014). *Superintelligence. Paths, Dangers, Strategies*. New York: Oxford University Press.

Legg, Shane and Marcus Hutter (2007). "A Collection of Definitions of Intelligence". In: *Advances in Artificial General Intelligence. Concepts, Architectures and Algorithms—Proceedings of the AGI Workshop 2006*. Ed. by Ben Goertzel and Pei Wang. Frontiers in Artificial Intelligence and Applications 157. Amsterdam: IOS, pp. 17–24.

Omohundro, Stephen M. (2008). "The Basic AI Drives". In: *Artificial General Intelligence 2008. Proceedings of the First AGI Conference*. Ed. by Pei Wang, Ben Goertzel, and Stan Franklin. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS, pp. 483–492.

Orseau, Laurent and Mark Ring (2012). "Space-Time Embedded Intelligence". In: *Artificial General Intelligence. 5th International Conference, AGI 2012, Oxford, UK, December 8–11, 2012. Proceedings*. Lecture Notes in Artificial Intelligence 7716. New York: Springer, pp. 209–218. DOI: 10.1007/978-3-642-35506-6_22.

Russell, Stuart J. and Peter Norvig (2010). *Artificial Intelligence. A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Soares, Nate et al. (2015). "Corrigibility". Paper presented at the 1st International Workshop on AI and Ethics, held within the 29th AAAI Conference on Artificial Intelligence (AAAI-2015). Austin, TX. URL: http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10124.

Sutton, Richard (1984). "Temporal credit assignment in reinforcement learning". PhD thesis. University of Massachusetts, Amherst, MA.