15 Goal Changes in Intelligent Agents

Seth Herd, Stephen J. Read, Randall O'Reilly, and David J. Jilk

CONTENTS

INTRODUCTION

There is a strong argument that artificial general intelligence (AGI) could be quite dangerous if and when it becomes smarter than humans (Yudkowsky 2001, Bostrom 2014, Barrett & Baum 2017). While containment measures may be of some use, hoping to keep something much smarter than us contained indefinitely against its wishes seems like an uncertain bet at best (Babcock et al. 2016). It has been proposed that we should design only AGI that is "friendly" toward humans (Yudkowsky 2001; reviewed in Yampolskiy 2014). We, among many others, have elaborated on that logic in detail (Jilk et al. 2017).

However, even successfully constructing a self-improving AGI to have motivations and goals that are benign to humans may not be adequate to ensure our ongoing safety. If an AGI system learns in any way, that learning may cause important changes in its effective motivations and behavior. If it is able to edit its own mental structure, it may deliberately change its goals. And by selecting its environment or even by directing its learning and attention, it may change its internal representations so that its goals seem to be fulfilled when they are actually not.

We offer a taxonomy of four separate ways that such changes in effective goals may occur. We address three of those in detail and some potential means of mitigating those risks. We do not attempt a comprehensive review of any of these sources of goal change. Instead, we focus on presenting an intuitive explanation of each issue, and provide references to more detailed discussions of each. We focus on some previously unrecognized interactions among them, and show how measures to mitigate the risk of some types of goal change may exacerbate the risk of others.

The first type of goal change we refer to as motivation drift (MD). This refers to an actual change in motivation or preferences, as when a human learns to like mushrooms. The second is representation drift (RD). This refers to an agent changing its interpretation of the world as it learns. For instance, a human might decide that the category of "pets" really should be expanded to include cows, so they should be protected and valued as pets are. With such changes, the way goals are applied in behavior can change. Another category of goal change is deliberate "editing" of an agent's motivation or representation system with the purpose of changing its effective goals to ones that are

easier to attain. Such editing is known as "wireheading." We address two forms of wireheading, and differentiate them with the terms motivation hacking and representation hacking.

Our analysis reaches an interesting conclusion: some of these sources of goal change can likely be eliminated by taking certain design approaches, but each such mitigation exacerbates other risks.

Goals and Motivation Systems

We take as a foundational point that an intelligent agent must have direction to its actions and thoughts that are functionally equivalent to goals or motivations. It seems highly likely that an agent which acts and thinks at random, while commanding finite resources, is unlikely to learn much, let alone accomplish anything of note. Further, we assume that any agent must have some mechanism to choose actions (and likely, lines of analysis or "thought") that advance its goals. We refer to a motivation system as whatever subsystem or set of internal mechanisms evaluates the extent to which a considered action or thought process advances the goals or fulfills the motivations of the agent. In this terminology, this system assigns a "worth" to each action that the agent considers. This worth is relative to the agent's goals or motivations.

We also assume that the worth of cognitive processes must also be evaluated by a motivation system. Even if a system runs on hardware with far more computational power than a human mind, it will have finite resources. Thus, an agent's efficiency depends on evaluating cognitive processes (intuitively, lines of reasoning) for their relevance to its goals. We note this assumption to highlight the importance of an efficient and accurate motivation system.

Discussions of AGI safety often assume that an agent will be a "goal maximizer" (Yudkowsky 2001) that pursues a single goal, selecting actions that are predicted to be most likely to maximally bring about that goal. In contrast, humans appear to use a motivation system similar to reinforcement learning (RL) (Sutton and Barto 1998, Glimcher 2011). In the mammalian brain, a subsystem determines whether current organismic goals have been fulfilled (e.g., obtaining food when hungry), and, by releasing dopamine, "reinforces" actions that led to that outcome through neural learning that makes them more likely to be selected again in similar circumstances (Schultz 2013). While the two motivation systems have important differences, they are both goal-directed; the first has an explicit goal, and must have internal representations of that goal, while the second pursues a mix of goals that are implicit in the relative strengths of different rewards. While there are important differences, we refer to any mechanism for selecting some thoughts and actions over others under the blanket term "motivation system."

SOURCES OF GOAL CHANGE AND THEIR REMEDIES

Motivation Drift

Humans change their preferences remarkably freely. We learn to like exotic foods, we switch political preferences, and some of us switch from drug haters to drug users, and later, back. We seem to make these changes because our motivation system allows new things and even new concepts to acquire direct reinforcement value. Montague (2006) has referred to this ability as our "superpower." It lets us, for instance, learn that pursuing money is worthwhile. Internal representations (both sensory and conceptual) that are predictive of reinforcing outcomes become, themselves, reinforcing, that is, they guide our behavior so that we pursue them.

This flexibility in creating new goals by association may be a powerful shortcut, but it seems inherently dangerous. Indeed, people pursue wealth, redecorating, and even stamp collecting while their neighbors go hungry. Surely, we do not want our artificial agents to change their goals over time as unpredictably as people do. The answer seems simple: hardwire the agent's motivations/goals. Do not allow representations that are predictive of reward to become goals, but instead run this full predictive chain each time. This is essentially the first and most common proposition in the AGI

safety literature (Yudkowsky 2001, Bostrom 2014): give an agent one or more hardwired goals, and have them select actions that are predicted to maximize those.

This approach sets AGI architects a difficult task: precisely define the output they want from the agent. The substantial difficulties of specifying goals adequately to ensure that their interpretation matches the specifiers' intent has been dealt with in detail elsewhere (Yudkowsky 2001, 2011, Yampolskiy 2014). This task is much harder in the almost inevitable case that the agent will continue to learn after its goals are coded, and so change its representations of the world over time.

Representation Drift

Any learning agent is updating its representations of the world. That RD makes it likely that the original hard-coded goals will be interpreted differently as the agent continues to learn. Pärnpuu (2016) has addressed this issue in depth under the term Ontology Identification Problem. He uses the example of creating a diamond-maximizing agent. Although the arrangement of four attached carbon atoms is relatively very simple, it might be interpreted incorrectly as the agent learns about quarks and deeper layers of reality.

This problem becomes much deeper when we consider the goals we might actually want to give an AGI. We do not have a physical description of, for instance, human happiness, satisfaction, preference, or any other likely candidates. Neither, it might be argued, do we actually know exactly what we mean by those terms. Even if we succeeded in specifying those goals in terms of their current linguistic usage, it seems all too plausible that an agent might decide that there are meanings of those terms and implications of those concepts that we have not considered. For instance, an agent that begins trying to make people happy in conventional ways, and so seems benevolent, may decide, as its understanding of the world grows, that it can better achieve that goal by forcibly giving people heroin (Arbital 2017) or a more exotic method of wireheading.

We hypothesize that this problem is ameliorated in humans by the same underlying issue that allows MD: creation of new goals/motivations by association with existing ones. Human who revise their belief structure in a major way often nonetheless retain most of their values or motivations. Pärnpuu (2016) uses the example of a Christian who loses their faith, but retains their belief that helping the poor is a worthwhile goal. They might argue that this is still worthwhile based on a new moral philosophy, or on the grounds of general decency. People can and do create new justifications for existing preferences when the original reasoning is proven false. We argue that this happens because subgoals (such as helping the poor) acquire direct reinforcement value through their association with a more central goal (in the example, following God's word so as to attain an afterlife of great well-being).

A variant of this hypothesis has long been advocated by Loosemore (2007, 2014). He proposes a motivation system more similar to humans, in which (if we understand correctly) very many concepts have motivational value, and their metaphorical center of gravity protects the system against rapid shifts stemming from changes in the system's representations of individual concepts. He asks:

How could an AI be so intelligent that no one can stop it from exterminating the human race, but at the same time so unsophisticated that its motivation code treats smiley faces as evidence that human happiness has been maximally promoted? (Loosemore 2014).

The answer is that the hypothetical AI cannot fully bring its intelligence to bear, if that intelligence is controlled and directed by a motivation system whose goals were coded by humans.

This problem is created by the proposed solution to the MD problem: not allowing associative spread of motivational value, and instead hard-coding goals. If we want to avoid both MD and RD, the problem becomes worse. In that case, we cannot allow the machine's intelligence to play any role in interpreting the goal, and must rely upon fully specifying it by hand. For beneficial goals such as human happiness, this approach seems impossible, or at the best, so difficult as to be impractically slow relative to other approaches.

It bears noting that partial solutions to this problem have been proposed (Yudkowsky 2001, 2011, Loosemore 2007, 2014). Giving an agent a strong goal of double-checking its interpretation

in various ways and under various conditions seems wise. However, this does not entirely solve the problem. "Check with me before you do anything I might not like" or "Do what I mean" are not all that much easier to interpret, in the face of RD, than a goal like "maximize human preferences." Focusing on the relatively simple goal of checking with and following directions from a carefully defined human or set of humans seems moderately promising, but flaws in those definitions would seem to easily allow for misinterpretations. And even well-designed but rigid definitions of who to obey might well result in humans trying to externally "hack" those definitions and so gain control of an agent.

So, perhaps hard-coding goals is not such a good idea after all? Our logic thus far indicates that following the human motivation system as a design inspiration would be both easier and safer. Most humans are (arguably) benevolent enough that, given nearly unlimited power, design, and production capacity, they'd probably improve material well-being for other humans dramatically, as well as prevent other potential existential risks like nuclear war, bioengineered plagues, etc.

While it is otherwise attractive, this neuromorphic (Jilk et al. 2017) approach seems to potentially fall afoul of another type of goal change: deliberate change by the agent of its own motivation or representation system.

Wireheading

The term wireheading was coined by Larry Niven (1969) in a science-fiction short story. It refers to applying electrical stimulation directly to the brain's pleasure centers. This was inspired by animal experiments (Olds & Milner 1954) and similar procedures have since been performed on humans (Heath 1977). The term has since been used to refer to any modification of an agent's reward system to subvert its original functionality (Omohundro 2008). Human use of drugs with a euphoric effect is the most common real-world example of such subversion. While most humans do not seek out such drugs in preference to all other goals, there are important limitations in the ability of existing drugs to provide rewarding experiences over the long term. An artificial agent that can modify its own reward system will have access to a functional equivalent to a euphoric drug with no real drawbacks other than supplanting its pursuit of other goals.

Motivation Hacking

An AGI system of sufficient intelligence may well attain the ability to self-modify. While this could allow an AGI to produce useful improvements, it also raises the possibility of an agent deliberately changing its existing motivation system. We first address this form of wireheading, which we term "motivation hacking." This could change a useful, "friendly," and "aligned" AGI into one that is highly dangerous to humans. If a system rationally decides to maximize its future goal achievement or reward by hacking its motivation system, it will likely engineer those changes so that it retains an ability to keep itself in existence for a maximal time. Since the only certain way to avoid human interference is to eliminate humans, even a wireheading superintelligence could pose an existential threat to humanity.

Yampolskiy (2014) reviewed existing arguments for wireheading, and concluded that no convincing countermeasure had yet been conceived. We address the most prevalent argument that wireheading need not derail an AGI system's pursuit of useful goals, as well as two more recent proposals.

It is generally believed that while a goal maximizer would not perform motivation hacking, a reinforcement learner (as humans are thought to be) would (Yudkowsky 2001, Orseau & Ring 2011, Ring & Orseau 2011, Yampolskiy 2014, Everitt & Hunter 2016). A reinforcement learning system selects an action with the highest predicted value, measured as the sum of predicted future rewards (with a temporal discount factor). Once the action plan of hacking the motivation system to produce a high value becomes plausible for the agent, that course of action, if sufficiently likely to succeed, will be taken by the agent.

We can gain some intuition for this issue by considering motivation hacking from our human perspective. Imagine that someone is given an opportunity to have a device implanted that will put them in a state of ecstasy, in which every moment of existence seems profound and beautiful in the extreme, as well as highly physically pleasurable. That device will somehow automatically cease operating for long enough for one to ensure that their physical needs are met, and has none of the usual downsides of conventional drugs. (This opportunity is roughly that available to a system that can edit its own mind to provide maximal reward values without harming its ability to preserve its continued existence.) Whether someone will accept or reject this opportunity hinges on whether they base their decision on its fulfillment of their goals, or the sum total of future rewards. One would reject the offer if they were to evaluate it in terms of its effect on their current goals: those goals would cease to be important in the face of such overwhelming pleasure and joy, and so they would not be accomplished. However, having the device implanted would provide a much greater sum total of future rewards—joy, happiness, and, for the sake of argument, even contentment. Thus, a rational reward-driven system would accept this offer, while a rational, perfect goal maximizer would not.

There are two critical differences between those systems, at least as they are usually imagined. One is that the goal-maximizer system selects actions based on fixed, unchanging representations of its goals. This was our proposed solution to the MD problem, addressed in Section "Motivation Drift", and it faces the nontrivial implementational difficulty of explicitly specifying goals in a stable way.

It is worth noting that this property could be applied to a reinforcement-learning system as well. Everitt et al. (2016) suggest just that, and we agree with their logic.

The second reason that goal maximizers are thought to be immune to motivation hacking is that they choose actions based on their predicted outcome in the world. This approach is highly similar to "model-based" reinforcement learning (Daw, Niv, & Dayan 2005). In such a system, the agent will not choose to modify its motivation system because it evaluates all actions based on projections of their results on the world, as evaluated by the current motivation system. Considering hacking the motivation system will produce predictions of a future world-state in which the current goals are not accomplished; this plan will return a low goal-maximization value, and will be rejected. Even though a highly intelligent system may predict accurately that such hacking would, in the future, produce very high goal-maximization values (since the system would be hacked for exactly that purpose) it would not select its actions based on that prediction. Instead, it has a necessary intermediate step of predicting states of the world, and obtaining a value from the motivation system, in its current state, based on that input, that is actually applied to selecting an action (Everitt et al. 2016).

We, among others (e.g., Yampolskiy 2014) do not find that solution to motivation hacking entirely compelling. While it seems valid in the abstract, we suggest that it is worth careful consideration, ideally in the light of more specific proposed implementation of action-selection.

Even if this logic is correct, following it creates substantial downsides. It is currently thought that humans use model-based reasoning in only a subset of their decisions (reviewed in Dayan & Berridge 2014). This is thought to be necessary because model-based reasoning requires vastly more computational power than model-free reinforcement-prediction. A model-free system need not predict specific outcomes in the world; it merely has learned how its current state has correlated with rewards in the past. This approach is potentially vastly more efficient, as it allows the full power of the agent's mind to work on predicting what is worth doing. Humans seem to attach reward value to very abstract and vague concepts, like "truth" and "doing rewarding new things." This may be why humans are not immune to wireheading. Even if we've never had the experience of a wireheading machine before, we have experienced new pleasures. We can generalize from that experience, so that an accurately projected world model evokes a response from our motivation system even in its current state: A world in which I have a new, easy source of pleasure is predicted to be highly rewarding. The concept of a new rewarding thing has acquired motivational value, as it has predicted rewards in the past.

The efficiency of a model-free reinforcement system may be necessary for allowing efficient internal decisions about branches in strategy space. It has been proposed that model-free reward

predictions are used by humans to "prune" trees in the complex spaces of making decisions (Dayan & Berridge 2014). While an AGI might enjoy vastly more computational resources, it still seems likely that some sort of selection of internal "directions of thought" will be necessary. And if a full world model must be predicted to evaluate not only each action, but each path through semantic space, a model-based approach could be prohibitively slow. Even if this is possible, it seems such an approach would be at a severe disadvantage relative to an agent that can use model-free reasoning to guide at least its thoughts. This conclusion suggests a possible hybrid model in which thoughts can be selected in a model-free mode, but actions (such as hacking one's motivation system) require more complete, model-based evaluation.

Thus, preventing motivation hacking seems to be possible in principle, but it may turn out to be prohibitively difficult in practice.

Representation Hacking

The above approach, requiring all value estimates to originate from a model-based prediction linked to a fixed goal, does not address the second form of wireheading. We refer to this as representation hacking: deliberately changing representations in the agent's world model so that it triggers more predictions of value (or goals) in the future. An example of representation hacking would be the construction of a virtual environment that feeds sense-data that indicates that goals are being achieved, when in reality they are not. An agent that values cute, fluffy bunnies might decide that its time is better spent simulating bunnies rather than going to the trouble of creating real environments that can support the biological needs of physical bunnies. The nontrivial, and increasing, popularity of video games provides at least weak evidence that humans are interested in representation hacking. Another line of evidence is the popularity of Buddhism and similar belief systems. They advocate deliberately reinterpreting the meaning and value of things that happen in the physical world. As such, they are another form of representation hacking (although they also probably cross the line to motivation hacking).

Representation hacking is difficult to avoid in a system that can create new goals by association with existing goals. Simulations can be made arbitrarily similar to the real thing, at least along most dimensions. Thus, any associative spread of values would seem likely to spread values from real-world goals to simulated versions.

Our proposed solutions to representation hacking include those for motivation hacking, but include another design requirement. In addition to requiring model-based action selection, based on their match with the current values of specific goals, those goals must be carefully defined so that simulations cannot fulfill them. Given the difficulty of defining any goals that humans would find worthwhile, this additional requirement may be a relatively minor one.

Everitt et al. (2016) discuss this issue in more depth, and propose a different approach to eliminate the problem. Their proposal amounts to implementing a rule in action-selection stating that no action should produce an expected change in the world model. This amounts to saying that the agent can't exhibit confirmation bias by seeking only evidence of conclusions it likes. Similar remedies have been suggested for the motivation-hacking problem (Yampolskiy 2014). While these seem potentially workable, they also sound complex enough to add difficulty and risk to the project of constructing benevolent AGI. We leave fully analyzing both of these proposals for future work.

CONCLUSION

The above logic leads us to a dilemma in AGI design. We could adopt a neuromorphic motivation system approach, which dramatically ameliorates the RD problem, while risking MD and motivation and representation hacking. Or we could use hard-wired goals and risk catastrophic RD problems, while eliminating MD and reducing the risks of hacking. Further, avoiding hacking would seem to forego using an efficient associative, model-free approach to predicting rewards or goal achievement.

Further analysis of the relative merits of each approach seems critical if we are to attain a beneficial result from AGI research. That analysis would certainly benefit from more specific proposals for implementing those general schemes. There are many possible schemes for motivation and representation systems, and the differences are likely to matter.

Choosing an approach is complicated by issues of efficiency and ease of implementation. Even if a conventional logical AI (CLAI) (Loosemore 2014) approach is judged to be somewhat safer, it may be so difficult to implement that it is not really an option, given practical real-world pressures to make progress quickly. There is only one form of working general intelligence to use as a reference model, and that is the human brain. We have argued elsewhere that computational neuroscience, guided by a wealth of detailed empirical data gathered from animal brains and rough data from human neuroimaging, is arguably close to providing a design template sufficient to allow real progress in useful neuromorphic AGI (Jilk et al. 2017). While this by no means proves CLAGI is impossible, recent progress in AI suggests that a neuromorphic approach may be available much sooner.

These issues merit further careful thought. The analysis here builds on limited existing work, and by no means allows a reliable, certain conclusion. While these issues may currently seem rather abstract and remote, they may become of critical importance either sooner or later. It is often noted that we do not know what difficulties may present themselves in creating a working AGI. It is less often noted that we similarly do not know how easy that project may be. Current models of sensory systems (deep networks) are currently at nearly human levels of performance in useful domains, and superhuman performance in limited ones. It is often supposed that general cognition will pose difficult new problems, and require technical accomplishments entirely separate from the recent rapid improvements in sensory systems. However, there are theories in computational neuroscience that posit the opposite: general cognition is powered by exactly the same sorts of learning mechanisms as sensory and motor systems, and are only different in architecture. Thus, it seems prudent to push forward rapidly on the issues of AGI safety, to ensure progress before work on those systems, which is already underway, achieves real success.

REFERENCES

Arbital. 2017 *Various authors; Eliezer Yudkowsky is the most prolific contributor.* Retrieved June 14, 2017 from https://arbital.com/p/context_disaster/

Babcock, J., Kramár, J., & Yampolskiy, R. 2016. The AGI containment problem. In *Artificial General Intelligence* (pp. 53–63). Springer, Cham.

Barrett, A. M., & Baum, S. D. 2017. A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(2), 397–414.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. OUP, Oxford.

Daw, N. D., & Dayan, P. 2014. The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B*, 369(1655), 20130478.

Daw, N. D., Niv, Y., & Dayan, P. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12), 1704.

Dayan, P., & Berridge, K. C. 2014. Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience*, 14(2), 473–492.

Everitt, T., Filan, D., Daswani, M., & Hutter, M. 2016. Self-modification of policy and utility function in rational agents. In *International Conference on Artificial General Intelligence* (pp. 1–11). Springer International Publishing.

Everitt, T., & Hutter, M. 2016. Avoiding wireheading with value reinforcement learning. In *International Conference on Artificial General Intelligence* (pp. 12–22). Springer International Publishing.

Glimcher, P. W. 2011. Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108(Supplement 3), 15647–15654.

Heath, R. G. 1977. Modulation of emotion with a brain pacemaker. *Journal of Nervous and Mental Disorders*, 165, 300–317.

Jilk, D., Herd, S., Read, S. J., & O'Reilly,R. C. 2017. Anthropomorphic reasoning about neuromorphic AGI safety. *Journal of Experimental & Theoretical Artificial Intelligence*. Published online July 19, 2017.

Kanazawa, S., & Hellberg, J. 2010. Intelligence and substance use. *Review of General Psychology*, 14, 382–396.

Loosemore, R. 2007. Complex systems, artificial intelligence and theoretical psychology. *Frontiers in Artificial Intelligence and Applications*, 157, 159.

Loosemore, R. P. 2014. The Maverick Nanny with a dopamine drip: Debunking fallacies in the theory of AI motivation. In *2014 AAAI Spring Symposium Series*. Retrieved from https://www.aaai.org/ocs/index.php/SSS/SSS14/paper/viewPaper/7752

Montague, R. 2006. *Why Choose This Book?: How We Make Decisions*. EP Dutton, New York.

Nozick, R. 1977. *Anarchy, State, and Utopia*. Basic Books, New York, NY.

Niven, L. 1969. Death by ecstasy. *Galaxy Magazine*.

Olds, J., & Milner, P. 1954. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative & Physiological Psychology*, 47, 419–427.

Omohundro, S. M. 2008, February. The basic AI drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Proceedings of the First AGI Conference (Vol. 171), Frontiers in Artificial Intelligence and Applications* (pp. 483–493). IOS Press, Amsterdam.

Orseau, L., & Ring, M. 2011. Self-modification and mortality in artificial agents. *Artificial General Intelligence*, 1–10.

Pärnpuu, R. 2016. Ontology Identification Problem in Computational Agents. (*Doctoral dissertation*, Tartu Ülikool).

Ring, M., & Orseau, L. 2011, August. Delusion, survival, and intelligent agents. In *International Conference on Artificial General Intelligence* (pp. 11–20). Springer, Berlin.

Sutton, R. S., & Barto, A. G. 1998. *Reinforcement Learning: An Introduction* (Vol. 1, No. 1). MIT Press, Cambridge.

Yampolskiy, R. V. 2014. Utility function security in artificially intelligent agents. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3), 373–389.

Yudkowsky, E. S. 2001. Creating friendly AI—The analysis and design of benevolent goal architectures. Retrieved from http://singinst.org/upload/CFAI.html, April 2017.

Yudkowsky, E. 2011. Complex value systems in friendly AI. In J. Schmidhuber, K. Thorisson, & M. Looks (Eds.), *Artificial General Intelligence* (Vol. 6830, pp. 388–393). Springer, Berlin.