Shared Safety Module

($Working\ paper, November\ 2019$)

Ondrej Bajgar
Future of Humanity Institute, University of Oxford
16-17 St. Ebbe's Street
Oxford OX1 1PT, United Kingdom
ondrej@bajgar.org

Abstract

This position paper suggests the *shared safety module* – a reusable package implementing a solution to some AI safety problem applicable across a range of different AI systems – as a possible goal of some branches of AI safety research. Such a module could provide an opportunity for technical collaboration as well as lower the costs and increase the availability of AI safety solutions, hence contributing to more widespread safety. Such aspiration could be seen as action guiding already in current safety research, calling for *shareability* to become an important property that AI safety researchers should seek in their solutions.

Whether it is more data, more testing, or more researcher hours, building solutions to most AI safety problems would benefit from extra resources. A natural path to more resources is pooling resources among multiple actors, which faces two challenges: the willingness of actors to cooperate (Askell, Brundage, and Hadfield 2019; Dafoe 2018), and having technical conditions allowing cooperation. This position paper proposes a framing for the latter and appeals to technical AI safety researchers to prioritize work on solutions that are *shareable* – easily applicable across multiple AI systems.

Beside the benefits of shareability in *building* a robust solution, once we *do have* a functioning shareable solution, it can become available to actors who by themselves may be unwilling or unable to dedicate sufficient resources to safety. Thus, shareable solutions can help in making safety widespread.

This document offers the frame of a *shared safety module* – a safety solution applicable to multiple AI systems that is as ready-to-use as possible – as a possible goal of shareable AI safety research. I illustrate how such a module could look and summarize some challenges and benefits that it could bring. This serves to emphasize the importance of *shareability* as a property that AI safety researchers should desire in their solutions.

General Description

I am not proposing a blueprint of a specific solution, but rather a very general frame into which solutions to various safety problems could aspire to fit in order to be widely useful. A shared safety module could take the form of a software package, but also a hardware module or an API service. This module could then be applied across a range of AI systems. Those systems would probably need to share similar safety concerns: an office cleaning robot and a stock trading AI are likely to be too different both in their structure and in the kind of safety considerations that need to be routinely considered; however, an office cleaning robot and a home assistant robot are probably similar enough to share a safety module if their internal structure is built to be ready for it.

The module would need to have standardized input and output, as well as a standard for its integration into the rest of the AI system. As inputs, it could have, for instance, a representation of the internal state of the system and observations from sensors monitoring the surrounding world. The output could be, for example, a binary signal that could halt the system. The appropriate integration mechanism would need to ensure that the "halt" signal is indeed translated into halting the system, and also that the inputs are accurate (otherwise, an inappropriately configured AI system could try sending deceptive data to avoid being halted). All of these would require standardization (and hence limitations) on the part of the AI system, which may prove to be a key obstacle, which we will return to later.

A shared safety module could be developed by an organization with an interest in global safety and made openly available to others; however, as long as it would provide safety functionality at a cost lower than in-house solutions, it could also be offered commercially. Another option would be multiple actors recognizing some safety problem as hard and grouping to collaborate on the shared safety module from the beginning, e.g. by including a common piece of architecture in their systems and then sharing data and parameter updates useful in improving the system.

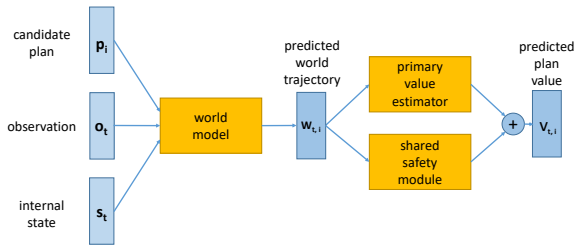

Figure 1: A toy example of a shared safety module for model-based value-based systems. A candidate plan of action p_i, the current observation o_t, and an internal state s_t are passed into a world model, which produces a prediction of the future world trajectory $\mathbf{w}_{t,i} = (w_{t,i,t+1}, w_{t,i,t+2}, ...)$ which is then assessed by a value estimator for the primary goal (for instance, this could estimate the expected reward due to the cleanliness of the household) and by the shared safety module, which calculates the expected value of the plan with respect to safety criteria. These two value estimates are then aggregated into a final value estimate which can be used to choose between different candidate plans of action.

Toy Example: Model-based Value-based Reinforcement Learning

To give a more concrete illustration, let us sketch how such module could fit into a system which is model-based (for each candidate action plan, it tries to simulate the resulting world trajectory – or ideally to estimate a distribution over such possible trajectories) and value-based (it estimates the expected value of each such potential world trajectory and then - setting exploration aside for now - selects the plan with the highest expected value). The safety module could fit into such a system as shown in Figure 1.

Let us assume the module would be made for robotic systems operating in human environments - it could be shared, for example, between a waiter robot and a home-cleaning robot. In such a case, the individual value modules would be tracking the goals specific to each system: for the waiter robot, this could be customer satisfaction or the restaurant's income. For the cleaning robot, it would be some measure of cleanliness. The shared safety module could then handle safety concerns in both cases, such as preventing colliding with people or objects, or creating tripping hazards.

In the simplest case, the two value estimates could be summed together, to get the final expected value of the suggested course of action. In that case, we may require the values coming from the individual value estimator to be capped, so that for potentially unsafe plans, the safety module could produce a sufficiently large negative value to negate that - no matter how clean the floor could be, it should never outweigh a serious health risk as assessed by the shared safety module.

This is just one simple toy example, which, however, hopefully clarifies the direction in which the concept of a shared safety module is pointing.

Benefits

Work on shareable AI safety solutions confers several benefits compared to developing separate solutions for individual systems or architectures:

Increased robustness. Having a shared safety module deployed across a wide range of systems operating in a wide range of contexts would allow us to collect much more data and have much more robust feedback on how the safety module is performing. This could at least partly help with several problems. For instance, among the Concrete Problems (Amodei et al. 2016), it might help with:

- **Costly feedback** – costs would get amortized across more systems and among multiple organizations
- **Distributional shift** – by being deployed across a multitude of systems, the shared module gets trained and tested in a wide variety of environments
- **Safe exploration** – it is better if a vase needs to be broken a few times across a network of a million devices than for each of them

Reduced costs of deploying safety solutions. Once we have a trained and well-tested module, any organization deploying a suitable AI system may be able to reuse it and hence easily get some needed pieces of safety functionality. This could massively lower the cost of addressing the given safety problem. Since safety seems generally desirable to AI system producers, the cost may be the main obstacle so its lowering could make reliable safety solutions much more widespread.

Certification and regulation. Furthermore, the shared safety module could provide an opportunity for regulation or certification of safety in AI systems. Any AI system containing a particular well-tested safety module (and the appropriate integration structure) could be "certified safe" (in some specific limited sense – we do not want to invite complacency). This could increase the trust of potential buyers (hence motivating businesses to comply) or could be a regulatory requirement.

Platform for collaboration. Cooperation in AI research has been repeatedly suggested as desirable: it could decrease competitive dynamics and the associated risk of underinvestment in safety by actors developing AI systems (Askell, Brundage, and Hadfield 2019; Dafoe 2018). Such collaboration makes sense on resources that could be useful to all of the cooperating actors – the shared safety module could be one such resource.

Desiderata and Challenges

There may be restrictions on what solutions may be packageable as a shared safety module. Also, the resulting AI

safety module may not be applicable in an arbitrary AI system. Here are some key restrictions and challenges associated with the concept:

Compatibility with different applications

In its essence, the shared safety module should be applicable to a variety of AI systems and to the different tasks that those systems are designed to carry out. Hence, it should be targeting safety risks that are present in many possible application domains – or in domains which contain many different AI systems.

Input and output standardization

In order to be compatible with a variety of AI systems, the shared safety module would need to have a standardized interface through which it would be integrated with the system.

Output The simpler part may be the output of the module, since for many safety applications, this may involve only binary outputs indicating actions such as "ban this action plan", "halt the system", or "notify a human operator". Other uses may be scalar, such as a numerical estimate of the negative value stemming from risks the module was designed to monitor (relevant for instance for value-based agents). Creating a shared interface for such outputs seems imaginable.

Input The standardization of input is a more problematic requirement. The purpose of the module may be to vet a course of action proposed by the main system. In such a case, it should receive the action plan as an input. Since each system may use a different set or even kind of actions, the requirement of their shared description seems problematic.

However, if the main system included a world-model, it could itself simulate into what world trajectory its actions may result and then pass the description of this trajectory to the shared safety module. Since such description would be dependent mainly on the environment in which the systems would be operating, a requirement for a common encoding seems more realistic, at least for systems operating in a similar environment. For instance, for robotic systems, a common relevant environment would be the physical surroundings of the robot – finding a standardized encoding for what may be happening in those surroundings seems realistic.

Safe integration

The optimization with respect to the primary task may push the system towards unsafe behaviours. Appropriate integration of the shared safety module in the rest of the AI system should ensure that the safety module has the power to prevent harmful behaviours, even if it conflicts with the primary task the system is designed to solve. No matter how clean the floor could become (primary task), a cleaning robot should not harm humans (secondary objective guarded by the shared safety module). Beside the safety objective being able to dominate over other objectives in this sense, we should also prevent any adversarial relationship between the module and the rest of the system – the rest of the system should not have incentives to trick the safety module.

These requirements suggest that an AI system would need to be purposely built to be compatible with a shared safety module. Sometimes, such limitations may not be acceptable to the system's producers. However, since safety solutions may be costly (e.g. due to requiring an extensive machine learning procedure), robust safety functionality provided by a shared safety module may make the limitations worth it.

Discussion: Possible Implications of Shareability in AI Safety Research

Especially if building an instance of an AI safety solution is costly, being able to collaborate on it seems desirable, as is making it available to third parties - otherwise, the solution may be beyond their reach, leading to decreased global AI safety.

Let us now turn our attention to what may be these costly areas within AI safety, and to the implications of prioritizing shareability.

Side effects and full value learning

Many risks in AI safety involve the AI system having undesired impacts on the world while pursuing some primary objective, which may stem from the primary goal being too narrowly defined. Avoiding the undesired side effects can be seen as a secondary objective. There are two broad ways we could address this: either learn the two objectives jointly or try to separate them.

The first approach means learning positive and negative preferences together – which, if done completely, means learning full human values or preferences in their complexity (Yudkowsky 2011). This may be the long-term goal of methods such as Inverse Reinforcement Learning (Ng and Russell 2000) or reinforcement learning from human preferences (Christiano et al. 2017). However, since human values may turn out to be complex, learning them may be extremely challenging. If it is done in a way that is bound to a single AI system, such methods may be inaccessible to most actors in scenarios with widely distributed AI development. This would be an argument for seeking shareable solutions in this space.

But there may be a fundamental problem in a shared encoding of human values: there is no consensus on them. If we are seeking a solution that could be widely used, or even turned into a regulatory requirement, we do need its contents to be widely endorsed and this is not the case with any set of "full human values" – yes, there are social choice mechanisms offering various ways of integrating preferences across a group of humans. However, such mechanisms generally result in solutions that do not quite fit anyone and actors would hence face incentives to deviate and try shifting towards their values, possibly at the expense of safety.

Consensus and safety as a constraint

However, there are principles on which there seems to be reasonable consensus – for instance, avoiding harm to humans or damage to their property. In human society, these are values generally guarded by criminal law. Maybe, we should be seeking its equivalent in artificial intelligence. A shared safety module could be a frame for a technical solution in this space.

In technical AI safety, this may mean shifting focus from learning full values towards methods for seeking consensus on what behaviour to avoid. Correspondingly, it may mean a shift from pure utility maximizing agents, towards constraint satisfaction – actors could construct AI systems to serve their specific goal, which would, however, be constrained into a safe region by a robust shared safety solution.

Conclusion

This article proposed the shared safety module as a frame for implementing shareable AI safety solutions. It can be seen as a practical aspiration of some branches of AI safety research. The possibility of sharing a safety solution could bring an opportunity for wider collaboration and for lowering the costs of making AI systems safe, hence making safety more wide-spread. These benefits can be seen as a reason to prioritize work on AI safety solutions that are shareable – a property that has so far not received much attention from the AI safety community.

Acknowledgments

I would like to thank the many people with whom I had the chance to discuss these ideas, especially my colleagues at FHI and the participants of the CHAI AI safety seminar at UC Berkeley. I would also like to thank Carolyn Ashurst, Owen Cotton-Barratt, Max Daniel, Ozzie Gooen, and Rohin Shah for helping me improve previous versions of this text with their useful comments.

References

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Askell, A.; Brundage, M.; and Hadfield, G. 2019. The role of cooperation in responsible ai development. *arXiv preprint arXiv:1907.04534*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 4299–4307.

Dafoe, A. 2018. Ai governance: A research agenda. *Governance of AI Program, Future of Humanity Institute*.

Ng, A. Y., and Russell, S. J. 2000. Algorithms for inverse reinforcement learning. In *ICML*, volume 1, 2.

Yudkowsky, E. 2011. Complex value systems in friendly ai. In Schmidhuber, J.; Thórisson, K. R.; and Looks, M., eds., *Artificial General Intelligence*, 388–393. Springer Berlin Heidelberg.