Peer Review File

Manuscript Title: Outracing champion Gran Turismo drivers with deep 2 reinforcement learning

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1

This is an admirably well-written paper reporting a reinforcement learning approach to simulated car racing. The work is comprehensive and methodologically sound and the reporting is, as far as I can tell, complete. I am nevertheless not convinced the result is neither important, general, nor novel enough for publication in a top-tier journal such as Nature. It might be better suited to a more specialized journal.

While the results are indeed impressive, as the trained agent achieves either superhuman or champion-level performance, this was achieved through plenty of human guidance. Human expertise were seemingly used everywhere, including in designing the features, designing the rather multi-faceted reward function, curating the training courses and opponents, and ultimately selecting which version of the trained model to use. This does not rhyme well with the idea of training an agent from interaction with an environment only. It would have been more impressive if it could have been shown that the setup used for this particular task generalized to any other tasks, but, alas, this research concerns only Gran Turismo.

Further comments:

The motivation for this research ("progressing the technology to physical applications requires showing success in realistic, real-time control of physical systems with complex nonlinear dynamics") rings a little hollow, given that online reinforcement learning is very unlikely to be useful to train agents that can drive real cars: training with real cars is likely too slow and crashing too expensive. Now, if the motivation was that you could train agents in the game that could be used to drive real cars that's another thing, but this is not mentioned at all (and the transfer problem itself can be quite thorny). You could also have argued that the race car driving problem has unique features not found in other problems and so is principally interesting for RL research, but I don't really see such an argument. I think a more honest motivation would be the need for better agents as opponents in racing games, as well as for testing them.

The new algorithm presented is, as far as I can tell, a minor variation on soft actor-critic. Nothing wrong with that, and sometimes minor variations end up being very important. However, it is not clear to me how this particular algorithm was motivated by or helped for this particular task. The input and output space for this racing task is very standard, and the reward scheme seems frequent enough. There isn't anything in the problem formulation, as stated, that seems like it would present an obstacle to existing RL algorithms. Could this not have been solved with standard DRL approaches? I would like to see at least a convincing explanation of why the quantile regression was needed _for this specific problem_, and preferably some empirical comparison of this method to a standard DRL algorithm.

Much of the most-heralded research on RL for video games uses raw pixels as input. Instead, in this work the complete position of each other car on the road is used, and the car state information is arranged in a convenient array. This represents a significant amount of feature engineering, presenting the state information in a way that it is easily parsable by a neural network. Now, I do not think that feeding the network pixel data is typically a good engineering solution: generally, this just makes the task harder. The approach taken in this paper certainly makes sense for getting good results. It is just less impressive than working with raw pixels. It also means that the agent gets access to more

information than human players get (they cannot usually see all other cars), and that a lot of processing (calculating positions) is done somewhere else than in the network. Furthermore, it means that the approach here cannot easily be migrated to a physical robot or real car. It is not clear to every reader that changing the input representation fundamentally changes the problem, so this needs to be pointed out and discussed.

Even beyond the issue of pixels versus features, one might question how fair the comparison to human players really is. It is implied that the agent plays at 10 fps in order to not make decisions more frequently than a human would. But frequency in decisions is not the same as latency. An oft-quoted number for the fastest human reaction time is 250 ms, which would equal a delay of two and a half frame. Currently, it would seem the trained agent can react faster than a human, and it would be interesting to see how well it performs in a setting with a 250 ms delay between the game and the agent.

It is not clear why the bibliography contains so little previous work on this very problem. There is plenty of previous research on learning agents that play racing games, using both supervised/imitation learning and various forms of reinforcement learning, including evolutionary reinforcement learning. In particular, the Simulated Car Racing Competition/Championship, which in various forms ran for around a decade from 2007 onwards, was based on a fairly sophisticated 3D racing game (TORCS) and spawned dozens of papers from organizers and various contestants. If memory serves me right, the best agents there drove faster skilled human drivers. There was also earlier research on using neuroevolution for training race car drivers, and of course the various iterations of the Drivatar system for Forza Motorsport.

Referee #2

* Summary of the key results

The authors design a system to train AI driving policies for simulated racing, Gran Turismo (GT), with model-free deep reinforcement learning (RL). They propose to use a distributional variant of soft actor critic and design dense rewards and observational features to encourage learning fast racing skills. The system was tested in competitions against some of best human drivers of GT and the trained policies show competitive performance.

* Originality and significance: if not novel, please include reference

To my knowledge, this is the first paper that demonstrates model-free RL algorithms can compete head-to-head in simulated racing with expert human drivers. While this paper does not have much algorithmic novelty in my opinion (the proposed techniques, features, and rewards are fairly standard variations of modern deep RL algorithms and natural engineering choices), the paper has an impressive contribution in the system aspect: The authors built a working system and the experimental results with expert human drivers (including world champions) make this paper particularly valuable, as typical deep RL papers often have much smaller lab experiments. It is quite informative to know how well deep RL algorithms nowadays can perform. However, a limiting factor here is that the agent has access to some low dimensional hand coded features. It would be more impressive if similar results can be demonstrated by using raw video frames from the game, as these are the information human driver uses.

* Data & methodology: validity of approach, quality of data, quality of presentation

There are some concerns about the correctness of the algorithm.
1. In Eq. (1), why aren't the entropy terms for the first N steps included? Since the goal is to solve an entropy regularized MDP.

2. How was the replay buffer implemented? Standard ones sample transition tuples but it seems that the

algorithm here requires trajectories.

3. Since the codes are not released, I think it is important to release the experimental results in more details so the correctness and significance can be verified. For example, can all the training curves (of various metrics), and the full videos of the competitions be released some certain manner? The videos submitted were quite selective and were insufficient in my mind to validate the performance here.

There some important experiment details missing:

1. An ablation of using multi-table stratified sampling or not is needed.

2. Fairness of the game: Were all the players in each game using the same car? If so, did the human drivers have the chance to play with the designated cars on the correspondent tracks before the competition (if so, how long?). Since the agent here had been training for these environments specifically. I think, the results here would be fair only if the human players had the same chance.

3. How were the hyper-parameters chosen? What's the protocol and how much computes were used?

4. Was the training done in the frame rate of typical playing or was it accelerated? Can you provide some ideas on how long the learning would take (say for a single game) if it were to be trained with the standard frame rate that the playing on PS4 would use and without parallel data collection?

5. Can the policy trained for one track generalize to a different game? Can you include experimental results on how well they perform?

6. How were the policies in Fig 2 selected? Were they the last policies?

7. Can you verify that observation features that the learning agent uses can be derived from the video frames? So that we can establish that the learning agent doesn't have unfair advantage.

The important policy selection procedure information is missing in the main text:

1. In line 79, it writes "…over 25,000 driving hours—shaving off tenths of seconds, until its lap times stopped improving. With this training procedure, GT Sophy achieved superhuman performance on all three tracks." This is misleading, as it hints that the learning converges and the last policy was picked as the policy for the final competition. But this is not the case, according to the Policy Selection section on page 11. In fact, I think this multistage selection process is quite crucial to the success of the proposed system, and it is important that this information is moved to main text.

The writing of this paper can sometimes be hard to follow because it introduces some nomenclatures that do not have technical transparency. I list some examples below.

1. In line 67, it writes "This incremental reward function". Why is the reward called incremental?

2. In line 97, it writes "we let the agent practice against curated policies from prior experiments that were selected for not exhibiting unsportsmanlike behaviors." What does unsportsmanlike behaviors precisely mean?

3. Racing etiquette. The authors spend many paragraphs to talk about these and suggest they are difficult to quantify. But at then end, it seems that the authors simply mean that certain collisions are not allowed. I would suggest that the authors to be more direct and remove these paragraphs. It would be better to talk about these collision avoidance requirement along with other qualities that are desired in the reward design section. Perhaps better, a summary list of behaviors desired for racing can be provided in the reward design section in the main text.

* Appropriate use of statistics and treatment of uncertainties

The experimental results only have the average of 3 seeds. It's simply too few. Since this is a complex problem and the performance difference is less than a second, please consider running the more experiments for about 8-16 seeds (e.g. in Fig 2) and report different statistics of the results (such as mean, confidence interval, etc.). The current results in Fig 2 do not have any statistical significance to draw the conclusions that the paper is making.

For Fig 3 (b), please also provide the statistics of GT Sophy with many runs, since GT Sophy doesn't depend on the human driver, if I understand it correctly.

* Conclusions: robustness, validity, reliability

Overall, I think that this is a good system paper with impressive experimental results that show deep RL can train driving policies competitive with expert human drivers in simulated racing. However, for the scientific side, important information about the algorithms and experiments is missing in the current manuscript, and they are critical to establish the significance, correctness, and reproducibility. In addition, the writing can be improved to be more technically transparent and use more direct, precise wording choices.

* Suggested improvements: experiments, data for possible revision

Please run more experiments and report the associated statistics. In addition, please revise the rewriting to make the content more transparent. Please see the comments above for details.

* References: appropriate credit to previous work?

There're many recent RL+racing papers, which are relevant but not referred here, but understandably many of them are on arXiv only (some of them might have made into recent conferences). Inå addition, distributional RL and distributional soft actor critic exist, and the reward design here is not completely novel (many of them are more less common practice; the authors should include appropriate citations if possible.). I would suggest the authors to down weight the novelty factor of the algorithmic part, but to focus more on the system contribution, i.e. the design and combination of all these different pieces. These details are very valuable.

* Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

The writing of this paper is clear, though there are some minor grammatical errors and typos. I have some minor comments.

1. In line 57, it writes "and other relevant state information about itself and all opponents." Please be clearer.

2. In line 68, it writes "In contrast, the classic approach of minimizing a constant step cost did not learn to complete the track in our experiment, as shown in Figure 2(c)." But the proposed reward here (progress reward + off-course penalty) is quite standard too.

3. Also please give the unit in Fig 2 (a).

Referee #3

This was a fun and engaging paper that I spent a lot of time working through in combination with the video provided. I think the authors have made some interesting contributions demonstrating the power of neural networks for learning a very complex task that has previously been the sole domain of human experts. They have very clearly described their neural network structure, their training methods and the specific rewards incorporated in training, further including some details into the relationship between rewards and learning. The end result is an artificial agent that demonstrates very competitive performance in the racing game Gran Turismo. I believe that this represents a solid contribution to the field from which others can learn.

I do feel that the authors should more fully ground their discussion with prior results that have been obtained on full-scale autonomous race cars on a track and dig a little deeper into what racing skills their neural network has and has not learned. This will increase the value of the results to the research community by more clearly identifying the contributions and limitations. I describe these suggestions below in what evolved to be a very long review (which is a testament to the way the paper captured my attention).

The third paragraph of the introduction states that "With the exception of an aerial dogfight project, none of these autonomous vehicle projects have demonstrated expert-level performance in head-to-head competition with humans." While this statement is strictly correct since it uses the qualifier "head-to-head," it overlooks work on full-scale automated race cars and the knowledge that has come from this work. One particularly relevant example from our own research is "Neural network vehicle models for high-performance automated driving," by Spielberg et al. (Science Robotics. 2019 Mar 27;4(28)). In this paper, we demonstrate that a structure consisting of an optimized path based on a relatively simple vehicle model and an extremely simple linear feedback can achieve comparable performance to a champion amateur race car driver in a real car on a real track. This paper also demonstrates that the simple vehicle model can be replaced by a neural network.

The work in this paper is distinct from the approach taken by Spielberg et al. and novel in its own right. But since the paragraph and the approach in this paper start with time trials, the authors seem to suggest that achieving expert-level performance in racing is novel when that has been previously established. Arguably, the time trial scenario considered in this paper is simpler than that investigated by Spielberg et al. since it is virtual and the vehicle and track conditions can remain constant during training and racing (Gran Turismo does allow tire wear but I can't find in the paper whether or not this was enabled). Racing in a real car involves dealing with constant changes in track temperature, tire wear and temperature, brake temperatures, engine output and other factors. So there are challenges facing a policy in the real world that do not appear in simulation.

That being said, Gran Turismo does have a solid dynamics engine (I have personally used it to learn the racing line around a track before driving it in the real world) and the level of outperformance of the human (virtual) drivers in time trials is higher than what we initially achieved on the track. These results are not surprising, however, given the way the authors have structured the problem.

GT Sophy has a few inherent advantages over the human driver as this comparison has been set up. The authors took care to limit the frequency of Sophy's inputs to 10Hz which would correspond to a maximum bandwidth of about 5Hz. That is at the high end of human ability (for a Nature article it would be nice to reference some of the large body of human factors work supporting this number instead of an NBC News blog post) but seems reasonable for race car drivers or gamers. As the article notes, no real advantage exists for higher frequencies which makes perfect physical sense. Cars are designed for humans and mechanically filter out frequencies higher than this so there is no value to be gained by including them. On this basis, GT Sophy and the humans seem well matched.

As I understand the paper, however, GT Sophy can apply inputs instantly. In other words, the steering input can reflect the current state. Even fast humans have a reaction time on the order of 300ms so there is a delay between the visual stimulus and their reaction. They also have to extract information from visual, auditory and haptic cues, giving them more uncertainty in the vehicle state, and have to physically execute their desired commands, which entails some error (the same sort of error that

prevents basketball players from hitting free throws every time). Given that the inputs to the neural network are essentially the same as the raw inputs into Spielberg et al.'s work (track boundaries from which a path can be derived and the vehicle state), prior art suggests that this approach should be able to beat a human. None of these are criticisms but rather explanations for why a neural network set up this way should be expected to beat human performance. It would be nice to see this sort of information in the paper so it doesn't come across as an unexpected result or one without precedent.

So the time trial results obtained are interesting but not very surprising given the state of the art. The more interesting results and the larger contributions come in the head-to-head aspects of the work. As the authors mention, this area has seen much less attention. It is interesting that the authors chose a network structure that essentially enabled them to train these behaviors on top of the time trial training. The structure and its success are clear contributions to the state of the art in my mind. However, the depth of these contributions is a little hard to gauge given the results presented. Since GT Sophy has a large advantage over the humans in time trials on two of the tracks, it doesn't have to have much passing capability to win the race. Once it gets out front, it just stays out front. So I spent a fair amount of time trying to parse what GT Sophy has actually been able to learn about passing. The results seem to me a bit more mixed than the paper would suggest.

Figure 3(d) shows an interesting limitation of the training approach used for GT Sophy. In this race on the Seaside track, three Sophys end up running near each other. In the process, they fall back from the other cars until one manages to break free of the pack at around 350 seconds (I'd love to see the video of what is happening there. I have a hypothesis described later in the review). This behavior seems to be a direct result of the training process which rewards passing over a short time horizon. An experienced racer knows that it is often better to simply follow a car for a while instead of attempting to pass since passing and defending slow the vehicles. An experienced racer will therefore not engage in a dogfight for seventh place but rather follow the car to catch up with the pack, looking for corners where the car ahead is slower and planning a pass by setting up a higher exit speed for that corner. While the neural network has all of the information necessary to learn such strategies, the passing reward is not judged over a long enough time horizon to learn such approaches. That is an interesting result.

The video provided does show a textbook slingshot pass. The text mentions that Sophy was specifically trained for such situations (which is fine, so are human drivers). These passes are available because of aerodynamics and not because of the actions of the other driver so it is one of the more straightforward passing opportunities. Nevertheless, Sophy "sees" the situation in the real race and applies this maneuver perfectly – extremely cool! On the other hand, the video also shows Sophy attempting this at the end of a straight which I can't imagine a human driver attempting to do. This positions the car on the outside heading into the approaching corner. As a very experienced racer recently observed to me, "The sun rises in the east and sets in the west and the guy on the outside loses." However, there is a possibility that this was not a flaw, depending upon the level of Sophy's competition in training. Against an inexperienced driver, this might have been a legitimate opportunity to overtake and may have been perceived as such by the neural network. Yamanaka, however, knows exactly what to do and mounts a successful defense by braking early and owning the inside of the turn instead of taking his normal racing line. So GT Sophy has clearly learned the mechanics of the slipstream pass quite well but not necessary exactly when to apply it. Given Figure 3d, I would suspect that mistimed pass attempts are likely the cause of the multi-Sophy pack falling back in the Seaside race.

The video also shows a pass at Maggiore where two Sophy cars dive underneath their human competitors. I have watched this a number of times to tease out what it shows about GT Sophy's ability to capitalize on opportunities on the track. In the video, Kokubun had lost speed because of going over the orange curbing on entry and therefore was at a disadvantage relative to Miyazono, who closed on the straight with a large speed differential. I am assuming that Kokubun's decision to enter the turn towards the bottom of the track and drive off line was motivated by blocking Miyazono but this move clearly slowed both human cars down dramatically, enabling both Sophy cars to get by. They are clearly able to take advantage of this situation but don't seem to have had to make much of a modification to the normal racing line to do so (both cars seem to follow the normal racing line through mid corner without the need for having to brake early and apex late to increase exit speed). In other words, they

have been able to overtake but they have not had to alter their nominal approach much to do so. I suspect that the pass on Seaside that broke up the Sophy pack might have been similar (two cars slow dramatically racing each other and the third takes advantage). Skilled drivers need to be able to recognize and capitalize on these situations when presented and GT Sophy does so. At the same time, this is not a terribly difficult pass to make given the information available to GT Sophy.

The more common and more challenging passing scenario is when two drivers are relatively equally matched but the following driver sees an opportunity in the manner in which the lead driver takes particular curves. The following driver may observe this over multiple laps and then set up a passing opportunity, potentially over several turns, by modifying their racing line. I don't see any evidence presented that GT Sophy has this skill, though it might.

What in my mind would have been the best test of GT Sophy's passing abilities would have been to use earlier Sophy versions that had lap times on par with the human drivers. These cars would likely have been better at some corners and the human drivers better at others and it would have been interesting to see if either the humans or the Sophy cars could find passing opportunities in these cases. That would have truly shown that the neural network had learned skilled passing beyond taking advantage of opportunities presented by aerodynamics or aggressive defending by other cars. I don't know how feasible it would be to do this or if the races run to date have solid evidence of this already (for instance, GT Sophy taking one line on an open track and a very different line/apex to overtake a single vehicle, demonstrating true passing skills) but this would strengthen the paper greatly.

To summarize, I found this to be very interesting work and it certainly encouraged me to spend a lot of time unpacking the results. I recommend that the authors take a bit deeper of a dive and give some of this narrative to the reader up front. Published work with full-scale automated race cars seems to have been overlooked and I believe that the time trial results described here are more straightforward than the authors seem to realize based on that work. The head-to-head racing results are very interesting but merit being analyzed and discussed in more detail to separate out GT Sophy's passing capabilities from the inherent advantage of not having human reaction time, perception and actuation limitations. Finally, I think the authors should be a bit clearer about what they have done. The title (and sections of the paper) state that they are "Training champion-level race car drivers" but they have trained an agent to play a racing game. I have deep respect for Gran Turismo but there are differences from results in simulation and real-world testing. The approach here is not appropriate for training human drivers nor is it obvious that it can directly transfer or apply to real cars (I'm sure no race car ever built has been driven the 25,000 hours needed to train the network, much less with consistent tire properties). I think the title and paper should make this clear. I also hope that the authors will share full videos of the race from the perspectives of the different cars. That would be useful to really gain a full understanding of the contribution here.

J. Christian Gerdes
8/28/2021

Author Rebuttals to Initial Comments:

Dear Reviewers,

We are pleased to resubmit our revised manuscript, which you will find substantially rewritten from the original draft. We had a chance to improve the agent from the July exhibition and had a rematch with the human drivers in October. The improvements include:

1. GT Sophy easily won the rematch, making the overall claim that the agent is superhuman more strongly supported.
2. Through various improvements in the training scenarios and using a larger network, we found we could train a single policy that learned to be both tactical and fast. Thus, the

agent no longer needs a separate policy for time trial and traffic. The only exception to the single policy improvement was the start of Sarthe, which is discussed in the Methods section.

3. The text is more clear and more coherent.
4. The time trial results have been deemphasized and more evidence as to GT Sophy's tactical skills is provided.
5. References to related work are much more thorough.

We are grateful for your thoughtful comments and have done our best to ensure that they have been addressed in the paper or an explanation is given below.

We have submitted a new video showing highlights from the races, but to make sure that you have the resources to judge GT Sophy, we've made all of the videos from the October race available at:

https://drive.google.com/drive/folders/1PGpylzG0ORpz2aL1kFQ5AkY143DCne8L

Please let us know if you have any trouble accessing the videos. Please keep these video confidential. If the paper is accepted, Sony will release videos of the event.

Referees' comments:

Referee #1 (Remarks to the Author):

This is an admirably well-written paper reporting a reinforcement learning approach to simulated car racing. The work is comprehensive and methodologically sound and the reporting is, as far as I can tell, complete. I am nevertheless not convinced the result is neither important, general, nor novel enough for publication in a top-tier journal such as Nature. It might be better suited to a more specialized journal.

While the results are indeed impressive, as the trained agent achieves either superhuman or champion-level performance, this was achieved through plenty of human guidance. Human expertise were seemingly used everywhere, including in designing the features, designing the rather multi-faceted reward function, curating the training courses and opponents, and ultimately selecting which version of the trained model to use. This does not rhyme well with the idea of training an agent from interaction with an environment only. It would have been more impressive if

it could have been shown that the setup used for this particular task generalized to any other tasks, but, alas, this research concerns only Gran Turismo.

Thanks for this comment. While we share your interest in what can be achieved with end-to-end machine learning, we found that human judgement was necessary in this domain because the objective function incorporates an imprecisely specified concept of sportsmanship, and the objective was to race against humans who behave differently than the AIs we use training. We submit that one of the takeaways from this work is precisely that human involvement will be required to build AI systems that interact closely with humans and need to respect social norms.

Additionally, there is a very general learning method embedded within the system. But to us (and we expect to many readers), part of the interesting aspect of this work is 1) understanding/evaluating the power and limitations of this general method and 2) introducing techniques that succeed at overcoming the limitations. While some may see these techniques as “not AI”, we find that the application of AI concepts to complex, real-world problems requires a lot of AI systems engineering. By sharing our experiences building GT Sophy, we hope to be useful to other researchers trying to solve similar problems.

Further comments:

The motivation for this research ("progressing the technology to physical applications requires showing success in realistic, real-time control of physical systems with complex nonlinear dynamics") rings a little hollow, given that online reinforcement learning is very unlikely to be useful to train agents that can drive real cars: training with real cars is likely too slow and crashing too expensive. Now, if the motivation was that you could train agents in the game that could be used to drive real cars that's another thing, but this is not mentioned at all (and the transfer problem itself can be quite thorny). You could also have argued that the race car driving problem has unique features not found in other problems and so is principally interesting for RL research, but I don't really see such an argument. I think a more honest motivation would be the need for better agents as opponents in racing games, as well as for testing them.

In this revision, we have changed the tone to be more focused on the contributions towards autonomous racing. Whether or not reinforcement learning will be part of future autonomous driving systems is an open question. It seems likely that neural networks will remain a critical component in the autonomous driving sensing pipeline, suggesting that the field will have to learn to deal with the black box nature ofneural networks.

In this work we tried to determine whether RL can be used to learn driving skills competitive with humans. We have succeeded, demonstrating skills well beyond the state of the art with other techniques.

The new algorithm presented is, as far as I can tell, a minor variation on soft actor-critic. Nothing wrong with that, and sometimes minor variations end up being very important. However, it is not

clear to me how this particular algorithm was motivated by or helped for this particular task. The

input and output space for this racing task is very standard, and the reward scheme seems frequent enough. There isn't anything in the problem formulation, as stated, that seems like it would present an obstacle to existing RL algorithms. Could this not have been solved with standard DRL approaches? I would like to see at least a convincing explanation of why the quantile regression was needed _for this specific problem_, and preferably some empirical comparison of this method to a standard DRL algorithm.

It is true that the algorithm is a relatively small enhancement over prior algorithms, and in this revision we have somewhat reduced that claim of novelty. However, we did find that the n-step returns and QR enhancements were critical to achieving superhuman time trial performance, which SAC alone did not achieve (Figure 2a). There is a good chance we would not have won the races without this seemingly minor change. While we have a hypothesis as to why that might be, investigating it remains a topic for future work.

Much of the most-heralded research on RL for video games uses raw pixels as input. Instead, in this work the complete position of each other car on the road is used, and the car state information is arranged in a convenient array. This represents a significant amount of feature engineering, presenting the state information in a way that it is easily parsable by a neural network. Now, I do not think that feeding the network pixel data is typically a good engineering solution: generally, this just makes the task harder. The approach taken in this paper certainly makes sense for getting good results. It is just less impressive than working with raw pixels. It also means that the agent gets access to more information than human players get (they cannot usually see all other cars), and that a lot of processing (calculating positions) is done somewhere else than in the network. Furthermore, it means that the approach here cannot easily be migrated to a physical robot or real car. It is not clear to every reader that changing the input representation fundamentally changes the problem, so this needs to be pointed out and discussed.

We agree that some of the major AI accomplishments used raw pixels (all of the Atari work), but point out that many others did not (Go, Starcraft, etc.). The reviewer raises two separate issues in the above comment: fairness to humans and value for real vehicles.

First, to address whether the format of the features gives the agent an advantage we have added a section to the Methods with a more detailed comparison of “fairness”. In short, the agent has some advantages and the humans have some advantages. We tried to make it as fair as possible while working within the engineering constraints of the systems we had. Using pixels would not be practical in this case because the game can only generate images for one car at a time, and only in real-time.

The second question is whether the features we’ve used would be available in an autonomous car. On thispoint, we respectfully disagree with the reviewer. The prior work on autonomous racing with real vehicles all start with a map of the track which they use to compute trajectories. Presenting the neural network with 60 equally spaced points along the track edges is arguably easier than the clothoid mechanisms that are currently used to compute trajectories. Furthermore, all autonomous vehicles employ a wide variety of sensors to locate the vehicles, pedestrians, and other obstacles around them. In the real world, sensors are

noisy and complicated, but reducing the vast amount of information they generate to a point, orientation, and velocity is feasible.

Even beyond the issue of pixels versus features, one might question how fair the comparison to human players really is. It is implied that the agent plays at 10 fps in order to not make decisions more frequently than a human would. But frequency in decisions is not the same as latency. An oft-quoted number for the fastest human reaction time is 250 ms, which would equal a delay of two and a half frame. Currently, it would seem the trained agent can react faster than a human, and it would be interesting to see how well it performs in a setting with a 250 ms delay between the game and the agent.

As mentioned, we have added a section in the Methods discussing the advantages and disadvantages the AI has in comparison to the human races. We've also investigated and included some results on how latency would affect the performance of GT Sophy, though it did not race with fabricated latency. While we agree that reaction time plays some role in racing, it is not the driving factor that it is in twitch games, and we don't know how to gauge its impact precisely. We submit that whatever advantage Sophy gains by not having perception and motor delays is offset by the fact that the humans have smooth 60 hz actions and Sophy is limited to 10hz actions.

It is not clear why the bibliography contains so little previous work on this very problem. There is plenty of previous research on learning agents that play racing games, using both supervised/imitation learning and various forms of reinforcement learning, including evolutionary reinforcement learning. In particular, the Simulated Car Racing Competition/Championship, which in various forms ran for around a decade from 2007 onwards, was based on a fairly sophisticated 3D racing game (TORCS) and spawned dozens of papers from organizers and various contestants. If memory serves me right, the best agents there drove faster skilled human drivers. There was also earlier research on using neuroevolution for training race car drivers, and of course the various iterations of the Drivatar system for Forza Motorsport.

We've worked to appropriately cite as much of this work as possible. In doing so, we have become more convinced that the work presented here is a significant advance over these prior studies, none of which advanced to real racing scenarios. We have cited prior work that uses a variety of other racing games, including GT, TORCS, World Rally, RARS, F1, and some homegrown games. Unfortunately, Drivetar is not well documented, but based on what we've been able to uncover, they use completely different techniques. We've also cited more of the work on autonomous racing with real vehicles.

Referee #2 (Remarks to the Author):

*** Summary of the key results**

The authors design a system to train AI driving policies for simulated racing, Gran Turismo (GT), with model-free deep reinforcement learning (RL). They propose to use a distributional variant of soft actor critic and design dense rewards and observational features to encourage learning fast racing skills. The system was tested in competitions against some of best human drivers of GT and the trained policies show competitive performance.

*** Originality and significance: if not novel, please include reference**

To my knowledge, this is the first paper that demonstrates model-free RL algorithms can compete head-to-head in simulated racing with expert human drivers. While this paper does not have much algorithmic novelty in my opinion (the proposed techniques, features, and rewards are fairly standard variations of modern deep RL algorithms and natural engineering choices), the paper has an impressive contribution in the system aspect: The authors built a working system and the experimental results with expert human drivers (including world champions) make this paper particularly valuable, as typical deep RL papers often have much smaller lab experiments. It is quite informative to know how well deep RL algorithms nowadays can perform. However, a limiting factor here is that the agent has access to some low dimensional hand coded features. It would be more impressive if similar results can be demonstrated by using raw video frames from the game, as these are the information human driver uses.

As mentioned above, a perfectly fair comparison to humans is difficult. We've added a section to the Methods detailing what we see as Sophy's advantages and disadvantages.

Like the reviewer, we are also interested in finding out whether an agent can be trained to drive this well using pixels as its input. Doing a study of end-to-end learning from raw pixels would be extremely resource intensive because the game can generate images for only one vehicle at a time; we'd need 10 to 20x the number of playstations to collect an equivalent amount of data. However, one of the more promising ways to train a vision-based agent is to let it learn to copy a competent state-based teacher.

We've had some early success in this direction, but it remains a topic for future work. And, naturally, it requires having a competent teacher, like GT Sophy.

*** Data & methodology: validity of approach, quality of data, quality of presentation**

There are some concerns about the correctness of the algorithm:
1. In Eq. (1), why aren't the entropy terms for the first N steps included? Since the goal is to solve an entropy regularized MDP.

The reviewer is correct that a more direct extension of SAC to N-step transitions would include the entropy reward for each of the intermediate steps. The primary reason we excluded this reward is for computational and simplicity reasons: including it would require an additional forward model computation on the intermediate time steps. Although this lack of intermediate entropy reward diverges from the strictest version of SAC, the primary motivation for SAC including it is to encourage exploration

and the degree to which SAC succeeds at that is somewhat unclear. In the first version of the SAC paper (although not appearing in the latest version on Arxiv) the authors investigated how much impact the entropy rewards had on learning. They found that the benefit of the entropy reward was minimal, but that the KL-divergence policy loss (which we maintain) was critical. You can find the first version of the paper here https://arxiv.org/pdf/1801.01290v1.pdf. On page 8 they say:

> "The entropy appears in both the policy and value function. In the policy, it prevents premature convergence of the policy variance (Equation 9). In the value function, it encourages exploration by increase [sic] the value of regions of state space that lead to high entropy behavior (Equation 5). Figure 3a compares how inclusion of this term in the policy and value updates affects the performance, by removing the entropy from each one in turn. As evident from the figure, including the entropy in the policy objective is crucial, while maximizing the entropy as part of the value function is less important, and can even hinder learning in the beginning of training."

Ultimately, our experiments showed that performance improved substantially when using N-step transitions despite the lack of intermediate entropy rewards (Figure 2c), so we stuck with the simpler and more computationally efficient approach. We have added text to the paper to highlight this omission andreasoning.

2. How was the replay buffer implemented? Standard ones sample transition tuples but it seems that the algorithm here requires trajectories.

Yes, in this case we stored trajectories as sequences of transition tuples. The sampling pulls out

sub-trajectories of length *n*. Some engineering was required to make sure that dropped frames in the trajectories were handled properly with respect to the *n*-step returns.

3. Since the codes are not released, I think it is important to release the experimental results in more details so the correctness and significance can be verified. For example, can all the training curves (of various metrics), and the full videos of the competitions be released some certain manner? The videos submitted were quite selective and were insufficient in my mind to validate the performance here.

We tried to keep the original videos short based on the submission guidelines, but we have made the entire race videos from the October event available for the reviewers. With respect to the learning curves,we have tried to provide enough information to give the reviewers confidence that the process works as described. In addition to the learning curve for the lap time for 15 random seeds, we have added a curve (Figure 2h) showing the behavior and retention of skills within a training run. We've also documented the policy selection process which helps us weed down thousands of policies into a single selection to race against the human drivers (Figure 5).

There some important experiment details missing:

1. An ablation of using multi-table stratified sampling or not is needed.

An analysis of the impact of multi-table buffers has been included in this revision. We are working on a deeper study of the multi-table approach, but in this article we now show evidence that it allowed the agent to better hold onto the skills it learned (Figure 2g).

2. Fairness of the game: Were all the players in each game using the same car? If so, did the human drivers have the chance to play with the designated cars on the correspondent tracks before the competition (if so, how long?). Since the agent here had been training for these environments specifically. I think, the results here would be fair only if the human players had the same chance.

Yes, all of the players used the same car, and the humans knew well before the July race which cars andtracks would be used. Since the same cars and tracks were used in the October rematch, they had a lot of time to practice.

3. How were the hyper-parameters chosen? What's the protocol and how much computes were used?

Thanks for raising these important questions. Many of the hyperparameters were chosen through manual trial and error or through limited parameter sweeps. Their values are documented in the Methods section. Our access to PlayStations grew from only 10 at the beginning of the project to over 1000 at the end of the project. However, each experiment required 10–20 PlayStations for 7–12 days, which greatly limited the number of parameter sweeps we could afford to do, especially if we wanted to run enough random seeds to get statistically significant results. The most thoroughly studied parameters were the variations of the features and reward signals. We cannot say these were the best possible values for the hyperparameters; we can only say they were sufficiently good.

4. Was the training done in the frame rate of typical playing or was it accelerated? Can you provide some ideas on how long the learning would take (say for a single game) if it were to be trained with the standard frame rate that the playing on PS4 would use and without parallel data collection?

Good question. We added a sentence stating that the game runs only in real time. The policies that ran in the rematch took 20 PlayStations 12 days to train, so would have taken approximately 240 days if you had only one PlayStation.

5. Can the policy trained for one track generalize to a different game? Can you include experimental results on how well they perform?

That's a great question for future work, and we have started investigating this.

6. How were the policies in Fig 2 selected? Were they the last policies?

The process of selecting a policy to race was complex and hard to describe in this short of a document. We have added a graphic to cover the process at a high level. One of the current challenges applying RL to complex domains with multi-dimensional objectives is that learning converges to a region around thepareto frontier of the objectives, but there may not be a single point to converge to. In our case, the

policies became very good, but due to the stochasticity of the contents of the ERB, performance on different skills would wax and wane. Thus, the last policy trained was usually not the best. We built an evaluation framework to help us make the best tradeoff between these multiple objectives over the range of policies. We have added Figure 6 to illustrate this process.

7. Can you verify that observation features that the learning agent uses can be derived from the video frames? So that we can establish that the learning agent doesn't have unfair advantage.

The question of fairness is a crucial but difficult one. We have added a section in Methods describing the advantages that GT Sophy has over the humans, and the advantages that humans have over GT Sophy.

In response to this specific question, we have not proven that the observations the agent uses can be derived from the video frames with the same level of precision that the agent has. It is worth noting that the human drivers don't operate just from what they can see–they have trained on these tracks for many hours, and have built up a mental map of the track that extends beyond what is visible. For now, trying to train an agent that uses vision as an input remains for future work.

The important policy selection procedure information is missing in the main text:

1. In line 79, it writes "…over 25,000 driving hours—shaving off tenths of seconds, until its lap times stopped improving. With this training procedure, GT Sophy achieved superhuman performance on all three tracks." This is misleading, as it hints that the learning converges and the last policy was picked as the policy for the final competition. But this is not the case, according to the Policy Selection section on page 11. In fact, I think this multistage selection process is quite crucial to the success of the proposed system, and it is important that this information is moved to main text.

We clarified the text, but the paper size constraints forced us to reserve the detailed description of the policy selection process for the Methods section. The 25,000 hours (now 45,000 hours to beat Valerio on Maggiore) comment refers only to the time trial evaluation, and so the criteria for that evaluation was simply the lap time. That said, as the reviewer points out, it isn't always the final policy that reaches superhuman performance. Throughout the training, we periodically evaluate the current policy and keep track of the best lap time found so far. Once the experiment is terminated, we select the individual policy that recorded the best lap time. And different random initial seeds for the network can vary significantly in the length of time it takes to get to superhuman performance, as discusses in Figure 1c. Figure 6 illustrates the process of selecting a policy for the races.

The writing of this paper can sometimes be hard to follow because it introduces some nomenclatures that do not have technical transparency. I list some examples below.

1. In line 67, it writes "This incremental reward function". Why is the reward called incremental?

Changed to "shaping".

2. In line 97, it writes "we let the agent practice against curated policies from prior experiments that were selected for not exhibiting unsportsmanlike behaviors." What does unsportsmanlike behaviors precisely mean?

This sentence has been rewritten. That said, one of the challenges of this project is that "unsportsmanlike" isn't precisely defined. We came up with some proxy measures that were directionally indicative of good sportsmanship, and also ran a few tests with very good drivers who work for Polyphony to collect their impressions. But the criteria remains subjective.

3. Racing etiquette. The authors spend many paragraphs to talk about these and suggest they are difficult to quantify. But at then end, it seems that the authors simply mean that certain collisions are not allowed. I would suggest that the authors to be more direct and remove these paragraphs. It would be better to talk about these collision avoidance requirement along with other qualities that are desired in the reward design section. Perhaps better, a summary list of behaviors desired for racing can be provided in the reward design section in the main text.

This section has been reduced somewhat. However, the issue of etiquette is actually one of the hardest things to quantify and what makes this project different from so many other AI projects. It is not strictly true that collisions are not allowed, and the criteria for which types of collisions are allowed are impreciseand context dependent, a common feature of sports. If the agent's reward function penalizes them too strongly, the agent becomes afraid to pass, and will actually give up its driving line to get out of the way of opponents. None of the previous well-known AI milestones (e.g. winning at chess, go, or jeopardy) had to address these issues of sportsmanship and human judges.

*** Appropriate use of statistics and treatment of uncertainties**

The experimental results only have the average of 3 seeds. It's simply too few. Since this is a complex problem and the performance difference is less than a second, please consider running the more experiments for about 8-16 seeds (e.g. in Fig 2) and report different statistics of the results (such as mean, confidence interval, etc.). The current results in Fig 2 do not have any statistical significance to draw the conclusions that the paper is making.

We have been able to extend the number of samples to 5, but each one is very expensive (one playstation, one computer, and one GPU instance for 10+ days) and we cannot collect enough to reach that level of statistical confidence that we'd all like to have. We have added "error bars" showing the range of the 5 samples for each bar in the graph. While we would like to collect more data, it isn't feasible given our computing limitations.

For Fig 3 (b), please also provide the statistics of GT Sophy with many runs, since GT Sophy doesn't depend on the human driver, if I understand it correctly.

The information the reviewer is asking for is available in Figures 1e, 4a, and 4b. Those figures each show a little orange histogram representing 100 samples of the selected policy for GT Sophy.

*** Conclusions: robustness, validity, reliability**

Overall, I think that this is a good system paper with impressive experimental results that show deep RL can train driving policies competitive with expert human drivers in simulated racing. However, for the scientific side, important information about the algorithms and experiments is missing in the current manuscript, and they are critical to establish the significance, correctness, and reproducibility. In addition, the writing can be improved to be more technically transparent and use more direct, precise wording choices.

*** Suggested improvements: experiments, data for possible revision**

Please run more experiments and report the associated statistics. In addition, please revise the rewriting to make the content more transparent. Please see the comments above for details.

We appreciate the suggestions and have made as many of the changes as possible.

*** References: appropriate credit to previous work?**

There're many recent RL+racing papers, which are relevant but not referred here, but understandably many of them are on arXiv only (some of them might have made into recent conferences). Inå addition, distributional RL and distributional soft actor critic exist, and the reward design here is not completely novel (many of them are more less common practice; the authors should include appropriate citations if possible.). I would suggest the authors to down weight the novelty factor of the algorithmic part, but to focus more on the system contribution, i.e. the design and combination of all these different pieces. These details are very valuable.

We have made these changes.

*** Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions**

The writing of this paper is clear, though there are some minor grammatical errors and typos. I have some minor comments.

1. In line 57, it writes "and other relevant state information about itself and all opponents." Please be clearer.

For space reasons, we had to leave the details to the Methods section.

2. In line 68, it writes "In contrast, the classic approach of minimizing a constant step cost did not learn to complete the track in our experiment, as shown in Figure 2(c)." But the proposed reward here (progress reward + off-course penalty) is quite standard too.

We rewrote this section.

3. Also please give the unit in Fig 2 (a).

Fixed.

Referee #3 (Remarks to the Author):

This was a fun and engaging paper that I spent a lot of time working through in combination with the video provided. I think the authors have made some interesting contributions demonstrating the power of neural networks for learning a very complex task that has previously been the sole domain of human experts. They have very clearly described their neural network structure, their training methods and the specific rewards incorporated in training, further including some details into the relationship between rewards and learning. The end result is an artificial agent that demonstrates very competitive performance in the racing game Gran Turismo. I believe that this represents a solid contribution to the field from which others can learn.

I do feel that the authors should more fully ground their discussion with prior results that have been obtained on full-scale autonomous race cars on a track and dig a little deeper into what racing skills their neural network has and has not learned. This will increase the value of the results to the research community by more clearly identifying the contributions and limitations. I describe these suggestions below in what evolved to be a very long review (which is a testament to the way the paper captured my attention).

The third paragraph of the introduction states that "With the exception of an aerial dogfight project, none of these autonomous vehicle projects have demonstrated expert-level performance in head-to-head competition with humans." While this statement is strictly correct since it uses the qualifier "head-to-head," it overlooks work on full-scale automated race cars and the knowledge that has come from this work. One particularly relevant example from our own research is "Neural network vehicle models for high-performance automated driving," by Spielberg et al. (Science Robotics. 2019 Mar 27;4(28)). In this paper, we demonstrate that a structure consisting of an optimized path based on a relatively simple vehicle model and an extremely simple linear feedback can achieve comparable performance to a champion amateur race car driver in a real car on a real track. This paper also demonstrates that the simple vehicle model can be replaced by a neural network.

Thank you for these constructive suggestions. In this version, we have done a more thorough job citing related work, including the paper mentioned above.

The work in this paper is distinct from the approach taken by Spielberg et al. and novel in its own right. But since the paragraph and the approach in this paper start with time trials, the authors

seem to suggest that achieving expert-level performance in racing is novel when that has been previously established. Arguably, the time trial scenario considered in this paper is simpler than that investigated by Spielberg et al. since it is virtual and the vehicle and track conditions can remain constant during training and racing (Gran Turismo does allow tire wear but I can't find in the paper whether or not this was enabled). Racing in a real car involves dealing with constant changes in track temperature, tire wear and temperature, brake temperatures, engine output and other factors. So there are challenges facing a policy in the real world that do not appear in simulation.

Without question, racing with real vehicles is much more difficult than racing in a simulation, even one as advanced and realistic as GT. In this version of the paper, we have reduced the focus on time trial scenarios and increased the focus on the racing skills. As the reviewer points out, GT does have the option to turn on (accelerated) tire wear, but it was turned off for these races.

That being said, Gran Turismo does have a solid dynamics engine (I have personally used it to learn the racing line around a track before driving it in the real world) and the level of outperformance of the human (virtual) drivers in time trials is higher than what we initially achieved on the track. These results are not surprising, however, given the way the authors have structured the problem.

GT Sophy has a few inherent advantages over the human driver as this comparison has been set up. The authors took care to limit the frequency of Sophy's inputs to 10Hz which would correspond to a maximum bandwidth of about 5Hz. That is at the high end of human ability (for a Nature article it would be nice to reference some of the large body of human factors work supporting this number instead of an NBC News blog post) but seems reasonable for race car drivers or gamers. As the article notes, no real advantage exists for higher frequencies which makes perfect physical sense. Cars are designed for humans and mechanically filter out frequencies higher than this so there is no value to be gained by including them. On this basis, GT Sophy and the humans seem well matched.

We improved the phrasing and references. As the reviewer noted, the work on human performance in sports is a deep topic and doesn't reduce itself neatly to a frequency or latency measure. We were unable to find any prior work evaluating race car drivers in particular, but we did add better references for evaluating the performance of athletes and gamers.

As I understand the paper, however, GT Sophy can apply inputs instantly. In other words, the steering input can reflect the current state. Even fast humans have a reaction time on the order of 300ms so there is a delay between the visual stimulus and their reaction. They also have to extract information from visual, auditory and haptic cues, giving them more uncertainty in the vehicle state, and have to physically execute their desired commands, which entails some error (the same sort of error that prevents basketball players from hitting free throws every time). Given that the inputs to the neural network are essentially the same as the raw inputs into Spielberg et al.'s work (track boundaries from which a path can be derived and the vehicle state), prior art suggests that this approach should be able to beat a human. None of these are criticisms but rather explanations for why a neural network set up this way should be expected to beat human performance. It would

be nice to see this sort of information in the paper so it doesn't come across as an unexpected result or one without precedent.

Because the comparison to human performance is subtle, we added a section to the Methods detailing advantages and disadvantages that Sophy has versus humans. We have also tested the impact on latency in Sophy's ability to learn the time trial task. Because driving is not a twitch game, latency may not be the best dimension on which to handicap Sophy, but it does provide some evidence that the techniques can still be applied even if artificial latency is added to the perception pipeline.

We agree that it is less surprising that a neural network could learn the time trial task, as this is a clear optimization problem. We have reduced the discussion around time trial performance and focused more on the tactical performance, and put in more citations to prior art.

So the time trial results obtained are interesting but not very surprising given the state of the art. The more interesting results and the larger contributions come in the head-to-head aspects of the work. As the authors mention, this area has seen much less attention. It is interesting that the authors chose a network structure that essentially enabled them to train these behaviors on top of the time trial training. The structure and its success are clear contributions to the state of the art in my mind. However, the depth of these contributions is a little hard to gauge given the results presented. Since GT Sophy has a large advantage over the humans in time trials on two of the tracks, it doesn't have to have much passing capability to win the race. Once it gets out front, it just stays out front. So I spent a fair amount of time trying to parse what GT Sophy has actually been able to learn about passing. The results seem to me a bit more mixed than the paper would Suggest.

Figure 3(d) shows an interesting limitation of the training approach used for GT Sophy. In this race on the Seaside track, three Sophys end up running near each other. In the process, they fall back from the other cars until one manages to break free of the pack at around 350 seconds (I'd love to see the video of what is happening there. I have a hypothesis described later in the review). This behavior seems to be a direct result of the training process which rewards passing over a short time horizon. An experienced racer knows that it is often better to simply follow a car for a while instead of attempting to pass since passing and defending slow the vehicles. An experienced racer will therefore not engage in a dogfight for seventh place but rather follow the car to catch up with the pack, looking for corners where the car ahead is slower and planning a pass by setting up a higher exit speed for that corner. While the neural network has all of the information necessary to learn such strategies, the passing reward is not judged over a long enough time horizon to learn such approaches. That is an interesting result.

The new version of the paper shows the results of the rematch that was run on Oct 21st. In this rematch, Sophy won handily, placing 1,2,4,5 on the first two races, and 1,2,5,6 on Sarthe. A single race is not a statistically significant data point on which to draw conclusions, but from our internal testing we can say with certainty that the versions of the policies we ran in October were better than the ones in July. From our analysis of Sophy's behavior in July, we determined that the agent was timid around other cars (at one point in Sarthe, it slowed down and moved over to let Yamaka pass), was unable to leverage its speed

when a car was 50-70m ahead of it (and thus not able to catch up), and was relatively poor at starts (thus, starting off losing positions in the race). We adjusted the scenario training to improve on these specific cases, and we were also able to train a single policy instead of needing to combine two more specialized policies. In head-to-head testing, our October policies consistently beat our July policies.

In the October races, the humans had a harder time than in July. We can't claim Sophy's improved performance was the cause of that, but it seems likely it contributed. Although the combined policy wasn't quite as fast as the pure time trial policy, it was still essentially superhuman in time trial performance, and it was significantly faster than the versus policy from July. Along with the fixes mentioned above, this meant that the humans could not get away from Sophy; the agent was always right behind them, applying pressure and looking for passing opportunities. The speed, combined with the tactical passing behaviors, gave Sophy a very good chance of winning. In this version of the paper we have included examples of the generalization of the tactical skills, and more information about the training process for developing and retaining them.

Some of the examples the reviewer mentions we would classify as strategic behaviors. Sophy is a very tactical racer, and we completely agree that Sophy does not yet have the ability to think strategically about the race or keep track of the weakness of individual opponents.

The video provided does show a textbook slingshot pass. The text mentions that Sophy was specifically trained for such situations (which is fine, so are human drivers). These passes are available because of aerodynamics and not because of the actions of the other driver so it is one of the more straightforward passing opportunities. Nevertheless, Sophy "sees" the situation in the real race and applies this maneuver perfectly – extremely cool! On the other hand, the video also shows Sophy attempting this at the end of a straight which I can't imagine a human driver attempting to do. This positions the car on the outside heading into the approaching corner. As a very experienced racer recently observed to me, "The sun rises in the east and sets in the west and the guy on the outside loses." However, there is a possibility that this was not a flaw, depending upon the level of Sophy's competition in training. Against an inexperienced driver, this might have been a legitimate opportunity to overtake and may have been perceived as such by the neural network. Yamanaka, however, knows exactly what to do and mounts a successful defense by braking early and owning the inside of the turn instead of taking his normal racing line. So GT Sophy has clearly learned the mechanics of the slipstream pass quite well but not necessary exactly when to apply it. Given Figure 3d, I would suspect that mistimed pass attempts are likely the cause of the multi-Sophy pack falling back in the Seaside race.

This is a very insightful comment. The July version of Sophy was trained against policies that were slightly worse than the human drivers and therefore may have thought it could complete the pass on that first section. The October version of Sophy had the opportunity to train against better agents, so may have learned to recognize this situation better. At the same time, without tire wear or gas consumption turned on, the attempt on the outside may have been worth trying because it cost Sophy nothing and had a small chance to work.

While we acknowledge that Sophy has little in the way of strategic behaviors, it is possible that the agent has learned behaviors that work for it that human drivers don't do because it is "common knowledge" that they won't work. At the same time, we also see the agent trying things that are clearly ill-advised, like passing on the final chicane at Sarthe (which we tried very much to discourage) and getting a penalty, when it clearly has the tools it needs to pass later in the race.

The video also shows a pass at Maggiore where two Sophy cars dive underneath their human competitors. I have watched this a number of times to tease out what it shows about GT Sophy's ability to capitalize on opportunities on the track. In the video, Kokubun had lost speed because of going over the orange curbing on entry and therefore was at a disadvantage relative to Miyazono, who closed on the straight with a large speed differential. I am assuming that Kokubun's decision to enter the turn towards the bottom of the track and drive off line was motivated by blocking Miyazono but this move clearly slowed both human cars down dramatically, enabling both Sophy cars to get by. They are clearly able to take advantage of this situation but don't seem to have had to make much of a modification to the normal racing line to do so (both cars seem to follow the normal racing line through mid corner without the need for having to brake early and apex late to increase exit speed).

In other words, they have been able to overtake but they have not had to alter their nominal approach much to do so. I suspect that the pass on Seaside that broke up the Sophy pack might have been similar (two cars slow dramatically racing each other and the third takes advantage). Skilled drivers need to be able to recognize and capitalize on these situations when presented and GT Sophy does so. At the same time, this is not a terribly difficult pass to make given the information available to GT Sophy.

Respectfully, we interpret the particular situation described above quite differently. In fact, we have chosen to highlight it in the paper with a new figure because it shows exactly how Sophy is able to adapt its trajectory to the situation. In Figure 3d we illustrate how GT Sophy Grey follows the humans (who aretrying to protect the inside of the curve) into the curve but is able to execute a sharper turn and get on their inside coming out of the curve. GT Sophy Green, being on the outside of the other cars, actually moves farther to the outside so it can pull an outside-in maneuver and also beat the humans coming out of the corner. Finally, the diagram also shows a red line indicating GT Sophy's preferred trajectory when nobody else is around, which follows the more classic out-in-out pattern. It may be that mistakes by the humans made it possible for both GT Sophy's to pass them, but their trajectories clearly show the flexibility of the policy to generalize to the situation.

The more common and more challenging passing scenario is when two drivers are relatively equally matched but the following driver sees an opportunity in the manner in which the lead driver takes particular curves. The following driver may observe this over multiple laps and then set up a passing opportunity, potentially over several turns, by modifying their racing line. I don't see any evidence presented that GT Sophy has this skill, though it might.

We agree that this would be a fascinating strategic skill for Sophy to learn. We'd like to research this in future work.

What in my mind would have been the best test of GT Sophy's passing abilities would have been to use earlier Sophy versions that had lap times on par with the human drivers. These cars would likely have been better at some corners and the human drivers better at others and it would have been interesting to see if either the humans or the Sophy cars could find passing opportunities in these cases. That would have truly shown that the neural network had learned skilled passing beyond taking advantage of opportunities presented by aerodynamics or aggressive defending by other cars. I don't know how feasible it would be to do this or if the races run to date have solid evidence of this already (for instance, GT Sophy taking one line on an open track and a very different line/apex to overtake a single vehicle, demonstrating true passing skills) but this would strengthen the paper greatly.

It is very difficult to run many tests against the top human drivers and be sure to collect interesting supporting examples of Sophy's skills. So, in addition to the evidence described above, we tested two similar GT Sophy policies against each other in 100 races on Sarthe and catalogued the number of passes that occurred (by either agent against any other car) across the whole breadth of the track. These results are shown in Figure 3e.

To summarize, I found this to be very interesting work and it certainly encouraged me to spend a lot of time unpacking the results. I recommend that the authors take a bit deeper of a dive and give some of this narrative to the reader up front. Published work with full-scale automated race cars seems to have been overlooked and I believe that the time trial results described here are more straightforward than the authors seem to realize based on that work. The head-to-head racing results are very interesting but merit being analyzed and discussed in more detail to separate out GT Sophy's passing capabilities from the inherent advantage of not having human reaction time, perception and actuation limitations. Finally, I think the authors should be a bit clearer about what they have done. The title (and sections of the paper) state that they are "Training champion-level race car drivers" but they have trained an agent to play a racing game. I have deep respect for Gran Turismo but there are differences from results in simulation and real-world testing. The approach here is not appropriate for training human drivers nor is it obvious that it can directly transfer or apply to real cars (I'm sure no race car ever built has been driven the 25,000 hours needed to train the network, much less with consistent tire properties). I think the title and paper should make this clear. I also hope that the authors will share full videos of the race from the perspectives of the different cars. That would be useful to really gain a full understanding of the contribution here.

Thank you very much for taking so much of your valuable time to fully understand our work, and also for agreeing to talk about it with us directly. Your feedback has helped us greatly improve the article.

We have improved the references to the prior work and deemphasized the time trial results.

In addition to the video with highlights, we have made the videos from the full race available to the reviewers and plan to release complete videos of the races once the paper is published.

J. Christian
Gerdes
8/28/2021

Reviewer Reports on the First Revision:

Referee #1

Thank you for taking the comments seriously. The new version of the manuscript is much improved, and I'm happy to see several substantial changes, including an expanded literature coverage.

My only substantial remaining request is that you offer some kind of explanation of _why_ QR-SAC was necessary to win the races, or at least why it helped. The ablation studies in Figure 2 are good and informative, but they don't offer much generalizable insight. As you designed this algorithm for the system you were working, and presumably tried a whole bunch of algorithm variations before you got things working as well as they do now, you should be able to comment on why the algorithm you have works. I presume that you did not introduce the algorithm modifications randomly, but had some kind of theory about why they would work on this problem?

Referee #2

The revised manuscript and the rebuttal have fully addressed the concerns I had. Especially, the inclusion of statistics, extra videos, and the discussion of pros and cons of GT Sophy make the paper more solid. Overall, I think this is a very impressive paper and engineering feat. Therefore, I would recommend acceptance.

Referee #3

Summary of the key results

The authors have demonstrated the power of neural networks for learning a very complex task that has previously been the sole domain of human experts. They have clearly described their neural network structure, their training methods and the specific rewards incorporated in training, further including some details into the relationship between rewards and learning. The end result is an artificial agent that demonstrates very competitive performance in the racing game Gran Turismo. This is a fun and engaging paper that I believe represents a solid contribution to the field from which others can learn. Relative to the original submission, the authors have provided more evidence demonstrating that the artificial agent has indeed learned passing techniques that can be contextually applied to overtake competitive human drivers.

Originality and significance: if not novel, please include reference

To the best of my knowledge, this is the first demonstration of an AI program capable of besting humans in a head-to-head competition on a racing simulator. In my review of a prior version of this work, I struggled to find any indication that the AI, GT Sophy, had learned much more than how to drive a fast racing line and how to make a slipstream pass on an open track. The results presented in this version clearly establish, in my view, that GT Sophy has indeed learned passing techniques from the training process and is able to apply these situationally in direct competition with a human sim racer. The results demonstrate respect for both the physical constraints of racing and the etiquette needed to drive safely with humans. In my opinion, this sets a new baseline of performance in terms of what can be expected from artificial agents competitively racing against humans and offers insight into methods appropriate for learning human-machine

interactions in complex dynamic environments.

As the authors detail, the reward functions necessary to produce these sorts of behaviors are far from trivial and the extended discussion of rewards will be useful to researchers working in fields where AI and collision avoidance are important (autonomous vehicles and robotics, for instance).

Data & methodology: validity of approach, quality of data, quality of presentation

Given the space constraints, there obviously have to be some choices about what to include. I liked the authors' decision to cover the network details briefly and focus more on the reward functions and specific issues that arose with training. I think that the data provided (including the video) are sufficient to back up the claims that the neural network has learned racing and passing techniques.

Relative to the original submission, I particularly like the addition of Figure 3(d). This puts the overtaking scenario at Maggiore into context and shows the (non-trivial) modifications the GT Sophy cars are making to their racing lines. After the last round of reviews, I spoke with some of the authors about possible ways to demonstrate evidence of passing ability beyond being able to drive a faster fixed line. The revised paper contains that evidence and presents it clearly.

Appropriate use of statistics and treatment of uncertainties

A head-to-head comparison with humans like this is obviously a limited set. Nevertheless I think it is appropriate for what the authors want to demonstrate. The paper does reference comparisons with a larger database of human trials from which the performance as a function of training time can be established statistically.

Conclusions: robustness, validity, reliability

The conclusions section nicely summarizes the contributions and all of the statements in this section are well supported by the text.

Suggested improvements: experiments, data for possible revision

• Line 219 mentions "g-forces." Do you just mean acceleration? It would be better to use a more precise term.
• On line 375 the paper states that GT Sophy had information about the distribution of mass. I don't see how that comes into the training features. Did I miss something?

References: appropriate credit to previous work?

The revised paper does a much better job of putting the work into the context of prior art. I thought the reference to the Indy Autonomous Challenge was timely but the phrasing These challenges may have contributed to the recent Indy Autonomous Challenge curtailing its planned head-to-head competition to time trials and simple obstacle avoidance was a bit awkward. Instead of speculating why the competition was moved from head-to-head to time trials, I think it makes the authors' point to simply point out that this move demonstrates the greater challenge of head-to-head competition.

Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

I do stand by my comment in the original review that the claim of training "race car drivers" seems a little overstated. I think the physics in Gran Turismo are excellent but they do tend to be a bit more forgiving than the real world. The passes provided in the video seem to show this in

some cases. I think the fact that the AI has learned to push the boundaries of the game is a testament to the structure chosen and the training process. I just think that there is a big step from training an agent to play a video game against human players and controlling an actual vehicle on track. The authors have made a real contribution here but I think it is important that the title and abstract not appear to be a broader contribution than what the paper supports. I believe they have developed a champion-level sim racer and presented the evidence to back up that claim. A champion-level race car driver is something a little different in my mind.

J. Christian Gerdes
11/29/2021

Author Rebuttals to First Revision:

Responses to reviewer/editor comments:

- The title changed to include Gran Turismo
- The abstract was expanded and some minor wording changed as requested
- Figure 2, caption, was significantly shortened
- Figure 2 (a–d) were reordered and their scales aligned
- All remaining references to “Sophy” were changed to “GT Sophy)
- Reference to Indy Autonomous Challenge was reworded to avoid supposition
- A pseudo code supplement was create, and hyperparameter values defined
- Reference to “g-forces” changed to acceleration
- Reference to “distribution of mass” changed to “tire loads”
- Table 1, which was in the Methods and showed the time trial results, was moved to Figure 3 in the main article because there was room
- Other minor typos and edits were made