Cognitive Systems Research 76 (2022) 63–77

Contents lists available at ScienceDirect

Cognitive Systems Research

journal homepage: www.elsevier.com/locate/cogsys

Where to from here? On the future development of autonomous vehicles from a cognitive systems perspective

Sara Mahmoud [a,*], Erik Billing [a], Henrik Svensson [a], Serge Thill [b,a]

[a] *Interaction Lab, University of Skövde, 54128 Skövde, Sweden*
[b] *Donders Institute for Brain, Cognition, and Behaviour, Radboud University, 6525 HR Nijmegen, Netherlands*

ARTICLE INFO

Keywords:
Artificial cognition
Self-driving cars
Cognitive paradigms

ABSTRACT

Self-driving cars not only solve the problem of navigating safely from location A to location B; they also have to deal with an abundance of (sometimes unpredictable) factors, such as traffic rules, weather conditions, and interactions with humans. Over the last decades, different approaches have been proposed to design intelligent driving systems for self-driving cars that can deal with an uncontrolled environment. Some of them are derived from computationalist paradigms, formulating mathematical models that define the driving agent, while other approaches take inspiration from biological cognition. However, despite the extensive work in the field of self-driving cars, many open questions remain. Here, we discuss the different approaches for implementing driving systems for self-driving cars, as well as the computational paradigms from which they originate. In doing so, we highlight two key messages: First, further progress in the field might depend on adapting new paradigms as opposed to pushing technical innovations in those currently used. Specifically, we discuss how paradigms from cognitive systems research can be a source of inspiration for further development in modelling driving systems, highlighting emergent approaches as a possible starting point. Second, self-driving cars can themselves be considered cognitive systems in a meaningful sense, and are therefore a relevant, yet underutilized resource in the study of cognitive mechanisms. Overall, we argue for a stronger synergy between the fields of cognitive systems and self-driving vehicles.

1. Introduction

Self-driving cars are a real world application of robotics with a significant societal impact (Hussain, Lee, & Zeadally, 2018). They operate in an uncontrolled environment that includes other agents (both human and non-human). A self-driving car therefore requires a range of capabilities such as predicting the intentions of pedestrians and other cars, or negotiating terms between several actors. Additionally, the vehicle is expected to provide a comfortable experience from the passenger's perspective. The development of self-driving cars is one of the most challenging research areas in robotics (Campbell, Egerstedt, How, & Murray, 2010), made more difficult by the fact that mistakes may cost lives.

Automation in self-driving cars is often defined by the Society of Automotive Engineers' (SAE) levels of autonomy (Committee, 2014) ranging from zero to five. Situations in which a car may operate autonomously under certain conditions start at level three. At this level, the human driver remains ready to take over when the system fails to proceed. Level four includes the ability to autonomously drive safely and operate well even when the system fails to hand control back to the human driver. However, level four vehicles are limited by infrastructure and legislation considerations that constrain them to operate within restricted areas. Level five represents fully autonomous operation in all situations. Despite being an active topic of industrial and academic research, there are presently no widely accepted solutions that reach this final level.

One of the earliest recognized achievements in developing self-driving cars was seen at the Defense Advanced Research Projects Agency (DARPA) Urban Challenge 2007, in which the top three teams – Boss (Urmson et al., 2008), Junior (Montemerlo et al., 2008) and Odin (Bacha et al., 2008) – used knowledge graphs with a search algorithm as a main method for developing their vehicle's control.

By the beginning of the 21st century with the advent of deep learning, neural networks gained significant popularity (Schmidhuber, 2015). The evolution of deep learning and convolutional neural networks allowed for great improvements in accuracy and performance, especially for pattern recognition tasks (Krizhevsky, Sutskever, & Hinton, 2012). However, the need for extraordinary amounts of training data and significant computational resources, in addition to issues such

[*] Corresponding author.
E-mail address: sara.mahmoud@his.se (S. Mahmoud).

https://doi.org/10.1016/j.cogsys.2022.09.005
Received 25 March 2021; Received in revised form 14 June 2022; Accepted 27 September 2022
Available online 1 October 2022

as transparency and interpretability of their functionality, raises challenges for their deployment in real-world applications in general (Rao & Frtunikj, 2018) and in self-driving cars in particular (Yaqoob et al., 2019).

More recently, the need for autonomous vehicles to be able to learn from unforeseen events and situations they may encounter, opened new perspectives with advances in the area for reinforcement learning (RL) and their application to autonomous vehicles (Marina & Sandu, 2017). In particular, RL-based approaches were able to demonstrate high performance in simulations (Grigorescu, Trasnea, Cocias, & Macesanu, 2020). At the same time, their application in real world environments remains challenging (Rao et al., 2020), despite improvements in knowledge transfer methods used for deploying agents trained in simulation into real-world environments (Zhao, Queralta, & Westerlund, 2020).

It has also become clear that there is not likely to be a gradual transition from driver assistance systems into a fully autonomous vehicles (Hars, 2016) because driver assistance and full autonomy have to tackle very different problems. Specifically, the former are designed for limited settings in well-defined scenarios while a fully autonomous cars operate in an unconstrained environment with a large degree of uncertainty. Overall, current paradigms for the control of autonomous cars rely either on a priori programmatic design or massive amounts of training data. Critically, both approaches assume that it is possible for developers and engineers to cover all possible driving situations, whether the driving model is expressed in the a priori rules or training datasets. Although these approaches may be sufficient for bounded tests and limited scenarios, it does not currently offer a path to level five of the SAE classification. As we will show in this paper, a level five autonomous car can be considered a cognitive system in a meaningful sense. Moreover, as others have argued (Thórisson, 2009; Thórisson & Helgasson, 2012), developing a cognitive system that can go beyond solving specific problems is not achieved by developing and improving certain cognitive abilities in isolation from each other. Rather, it requires a holistic view that takes into account underlying theoretical assumptions, inspiration, motivation, requirements, methodology, structure, and technology.

Interpreting an autonomous vehicle as a cognitive system opens the possibility of looking towards existing cognitive systems research and considering the degree to which progress and theoretical insights from that domain can be applied to state-of-the-art progress in work on autonomous cars.

More specifically, we can consider work on cognitive architectures (which can be understood as a framework that enables the cognitive abilities of an agent). In cognitive systems research, such architectures are typically divided into three different types (Vernon, 2014): cognitivist, emergent, and hybrid. Cognitivist architectures take a computationalist, symbolic based approach according to which information is processed according to some formal architecture *provided by the designer*. Emergent architectures, on the other hand, emphasize developmental processes that lead to the *emergence* of an appropriate architecture; in particular one that allows a cognitive agent to learn from the interaction with the environment and adapt to novel situations. *Hybrid* approaches, meanwhile, combine aspects of computationalist systems with aspects of emergent systems (the former often being symbolic while the latter are subsymbolic). Most modern cognitive architectures are hybrid in the sense that they rely on (subsymbolic) neural networks for some aspects of their functionality and computationalist approaches for others (Kotseruba & Tsotsos, 2020).

It is also worth highlighting *Enactivism*, a particular theory of cognition that emphasizes that cognition is *enacted* in the world, and therefore pays particular attention to the precise relationship between a cognitive agent, other agents, the environments, and the various interactions in between. While enactivism as a whole is a rather complex subject matter (on which whole books have been written, e.g. Maturana and Varela (1987)) that cannot be addressed in full here, we will show that this focus is relevant for understanding how a self-driving car interacts with its environment.

The core purpose of this paper is therefore to discuss the utility of examining autonomous vehicles as a cognitive system, and to identify what kind of inspiration cognitive systems studies can offer for the development of self-driving cars. We begin by analysing the state of the art in the self-driving cars literature to identify how progress has occurred so far. We highlight what kind of problems are solved (or solvable) and where current challenges lie. We then map this work onto different paradigms in cognitive systems research, noting that research on self-driving vehicles is largely driven by a cognitivist perspective. We then argue that alternative paradigms, e.g. emergent approaches, have the potential to contribute significantly to the further development of self-driving cars.

Lastly, we conclude by noting the potential of a symbiotic relationship between the field of self-driving cars and cognitive systems research. Specifically, while most of the paper focuses on the contributions that cognitive systems research can bring to the development of self-driving cars, we note that such cars can be meaningfully characterized as cognitive systems in the sense of cognitive systems research, and, as such, provide researchers within this domain with a platform on which to test their cognitive architectures and models. In particular, a self-driving vehicle retains the simplicity of wheeled robots in terms of degrees of freedom, but interacts with the rich and unpredictable real world.

2. Self-driving car systems

This section introduces self-driving car systems followed by a general description of an architecture for the perception-decision-making tasks that are inherent to self-driving car systems. Thus, this section provides a general overview of the challenges faced by self-driving car systems. More detailed discussions can be found in recent reviews (Badue et al., 2020; Paden, Čáp, Yong, Yershov, & Frazzoli, 2016; Pendleton et al., 2017).

As with all cognitive systems operating in the real world, it is neither feasible nor desirable to comprehensively enumerate all the situations they might encounter (Thill & Vernon, 2017), in particular when dealing with rare or unforeseen events. However, it is possible to categorize the most common tasks that a vehicle will encounter. Table 1 provides a brief overview of these tasks, representing primary problems that autonomous intelligent driving systems need to be able to solve.

When it comes to the design of intelligent driving systems, a commonly used architecture is the perception-decision-making architecture (Badue et al., 2020) (Fig. 1), which can also be seen as a kind of subsumption architecture in which decision making is divided into levels of individual sub-behaviours (Brooks, 1986). This kind of architecture was used by the Stanely team (Fig. 2), the first winner of DARPA challenge 2005 (Thrun et al., 2006), and in the "Junior" car that finished in the top three in the 2007 DARPA challenge (Montemerlo et al., 2008).

In a subsumption architecture, decisions are taken by hierarchically organized individual sub-systems based on the information provided by the perception system. Breaking down components into smaller sub-systems helps understanding the basic tasks and behaviours that are relevant for a self-driving car driving system. It also facilitates modelling driving tasks such as those summarized in Table 1.

Details of such a break-down of the perception-decision-making architecture into sub-systems can be found in thorough surveys (Levinson et al., 2011; Yurtsever, Lambert, Carballo, & Takeda, 2019). Briefly, however, the perception system links an agent to the environment through different types of sensors such as cameras, LIDAR, radar, GPS and odometers. It is thus composed of sub-systems determined by the collected data and its usage (summarized in Table 2). Although these sub-systems may seem well defined for a problem, many challenges may occur in the perception system. Weather conditions, for example,

Table 1
Examples of common driving tasks.

Task	Description	Includes
Lane keeping	Drive within the specified margins of the road.	Interpretation of road markings (Mathibela, Newman, & Posner, 2015). Estimate missing sidelines. Road speed limit.
Traffic	Adapt, and adequately respond to other vehicles.	Maintaining the safe distance with other vehicles. Response to other vehicles' behaviours (eg. sudden break or over take).
Intersection	Deal with signalized and unsignalized intersections.	Detect and identify traffic signs and rules. Negotiate with other road users in unsignalized intersections.
Roundabouts	Estimation of the situation to enter the roundabout.	Predicting the speed and the distance of incoming vehicles for precise judgments of when to enter the roundabout.
Pedestrians and other non-vehicle road participants	Deal with non-vehicle road participants such as cyclists, pedestrians or even animals.	Understand participant intentions. Knowledge of local conventions.

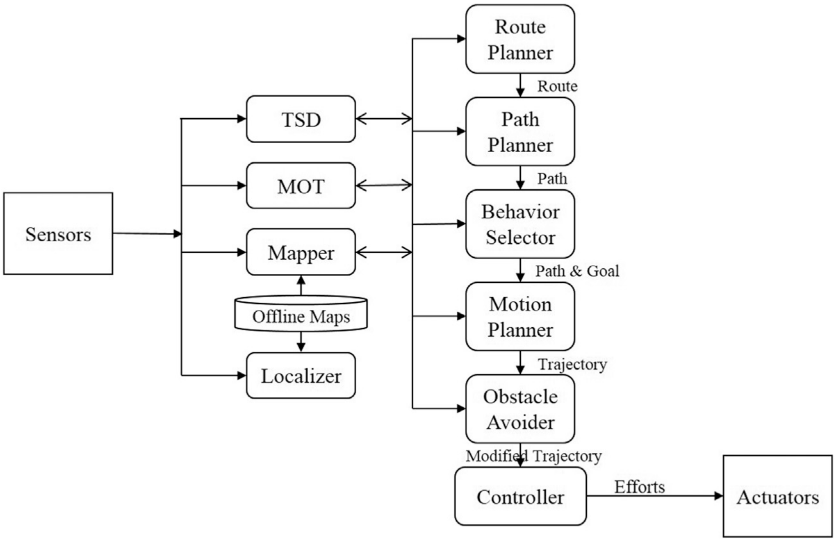

Fig. 1. Common architecture of self-driving cars.
Source: Simplified from Badue et al. (2020).

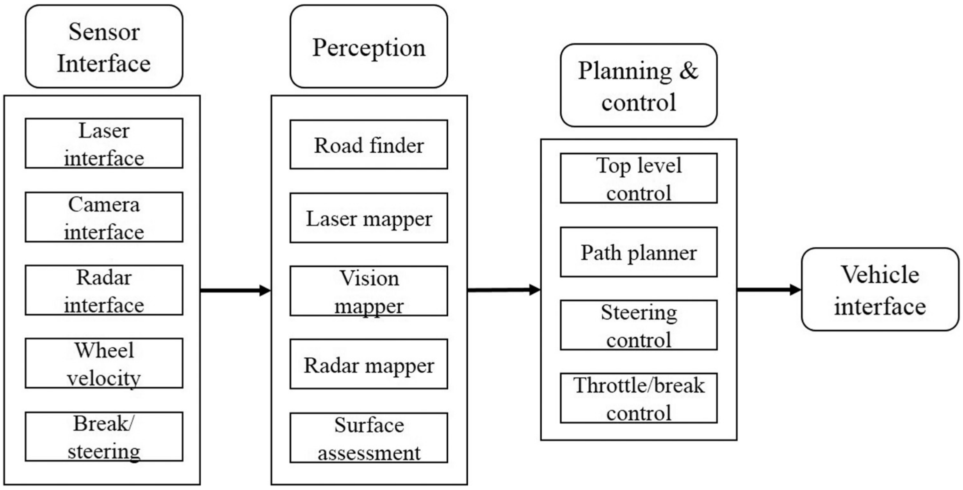

Fig. 2. A general description of the architecture of the first winner of the DARPA challenge 2005 named "Stanely" (Thrun et al., 2006). Reproduced with permission; John Wiley and Sons.)

may cause difficulties in detecting obstacles or the movement of an object in rain, snow or fog (Zang et al., 2019).

Second, after the perception system creates an internal representation of the environment and the corresponding car state, the decision making system determines the vehicle's next actions at different levels of abstraction. We summarize the main levels in Table 3.

It is worth bearing in mind that a self-driving car not only has to deal with clearly defined tasks such as the examples in Table 1. In reality, it faces a significant variety of challenging situations that eschew classification tasks into clear categories. As an example, consider a traffic jam at an intersection such that the reality of the traffic situation is in violation of traffic rules. In this situation the task is to clear

Table 2
Summary of the main sub-systems in the perception layer.

Perception sub-system	Description
Static obstacle mapping	Creating an internal representation of the static objects and their location in a created road map
Localization	Determining the location of the vehicle on an internal map. Localization can be done using LIDAR (Wolcott & Eustice, 2015), LIDAR with cameras (Xu et al., 2017) or just cameras (Wu, Tang, & Li, 2018).
Road mapping	Creating a map of the road lanes, specifying the number of lanes, lane merges and crossing lanes (Bresson, Alsayed, Yu, & Glaser, 2017).
Moving object tracking	Detecting and tracking moving objects such as pedestrians and other vehicles including current and future locations (Wang, Thorpe, Thrun, Hebert, & Durrant-Whyte, 2007).
Traffic signalization detection and recognition	Detecting and recognizing traffic lights and road signs (Wali, Hannan, Hussain, & Samad, 2015).

Table 3
Summary of the main decision-making subsystems.

Decision-making sub-system	Description
Route planning	Determines the route from the current position to the final destination. Invoked once per trip, or when the plan changes due to an external cause (Sanders & Schultes, 2007).
Behavioural layer	Takes the route plan from the route planning sub-system and then decides when to stop, change lane, negotiate an intersection or follow the current lane.
Motion planning	Takes the manoeuvre decision from the behavioural layer sub-system and produces the corresponding path trajectory to be executed.
Obstacle avoidance	Prevents collisions with objects that the perception system has detected. It can for example, reduce the velocity, carry out emergency braking, or change lane to avoid a predicted collision with a detected object (Jain & Malhotra, 2020).
Controller	Executes the planned motion and trajectory by the vehicle's actuators.

the traffic or the deadlock but there may not be direct instructions or rules of how to achieve this task. The difficulty of describing and then classifying and modelling all kinds of tasks that a self-driving car may face leads to a need for methods that go beyond the direct modelling from task description to component implementation. Historically, self-driving cars research and development has seen different paradigms for addressing these challenges. Current methods, for example, rely heavily on the latest advances in artificial intelligence (AI) and machine learning (ML), but previous approaches were more computationalist (see the early DARPA winners discussed above).

Next, we therefore discuss examples of a variety of AI approaches that implement the tasks summarized in this section.

3. Approaches to realizing self-driving car systems

This section focuses on the AI and ML techniques described in the literature on self-driving cars. We limit this discussion to papers that describe complete agents acting in real world settings and using AI-based approaches for the intelligent driving systems. In other words, we only consider work that fulfils all of the following criteria: (a) there is an explicit aim to contribute to the field of self-driving cars, (b) the self-driving car systems use AI techniques in their implementation, (c) the architecture involves tasks related to the intelligent driving systems, and (d) the systems are tested in a real world car or with real world datasets. Our aim here is not to present a comprehensive survey of self-driving car algorithms, but to demonstrate how AI and ML have contributed to the development of self-driving cars, and where their limits lie. In summary, we have identified thirteen papers, summarized in Table 4, according to the following criteria: (1) main symbolic AI and ML techniques used (see Sections 3.1 and 3.4), (2) type of task used (Table 1), (3) type of sensors, and (4) whether the behaviours are mainly derived from a given a priori knowledge base or discovered based on training data (see Table 4 for a complete summary).

Based on the analysis of these papers, we identify four main approaches: (1) symbolic AI, (2) neural network as subsystem, (3) neural network as end to end learning systems, and (4) reinforcement learning. In the following subsections, we describe each approach based on a rough chronological timeline.

3.1. Symbolic AI

Symbolic AI, creates an explicit symbolic representation of real-world objects. Reasoning then takes the form of symbolic operations carried out on these explicit and interpretable representations. Symbolic AI has been used in self-driving cars for different purposes, for example, to represent each object by its attributes and behaviours for knowledge inference (Oka et al., 1999). Notably, this approach produced the top three finishers of the DARPA challenge 2007, "Boss" (Urmson et al., 2008), "Junior" (Montemerlo et al., 2008), and "Odin" (Bacha et al., 2008). The challenge included sophisticated tasks such as navigation, lane change, parking and intersections. Despite differences in their architecture designs, all three agents modelled data in knowledge graphs with symbolic representation and used it to search the state space for maneuver decision making. The strength of this approach thus derives from its ability to explicitly infer knowledge from a given set of rules and a predefined model of object representation.

3.2. Neural networks as a sub-system

Another approach to the development of self-driving cars is to allow learning from large amounts of data (Grigorescu et al., 2020), in particular using modern deep learning methods. Presently, Convolutional Neural Networks (CNN) (Krizhevsky et al., 2012) are widely used, for example, to classify objects in the environment. The objects thus identified are then passed to the decision making subsystem, which determines actions to take (Geiger, Lenz, Stiller, & Urtasun, 2013). For example, Chen et al. (2015) used the AlexNet CNN for affordance detection and then a symbolic decision making system for driving action in an urban environment with low traffic. Similarly, Al-Qizwini et al. (2017) used the GoogLeNet CNN for affordance detection with higher accuracy and then a symbolic decision making system for driving action. They also improved training parameters and adapted more realistic assumptions. Another study (Kim & Canny, 2017) used neural networks for both the perception and the decision making components. Specifically, they used multi-step decoder as an attention system in order to learn to define which part of the sensory image contributes to driving learning. The attention system is followed by a long short-term memory (LSTM) to anticipate the driving control. The defining feature of neural networks used in this type of approach is that they are only used as a subsystem to solve one part of the driving task, in contrast with end-to-end architectures.

Table 4
Summary of different approaches and architectures for developing self-driving cars intelligent driving systems.

Approach	Project	Architecture main components	Lane keeping	Traffic	Pedestrians	Camera	Lidar	GPS	Distance sensors	Logical reasoning	Training data
Symbolic AI	Oka, Tashiro, and Takase (1999)	Vision- Decision-Motion	✓			✓				✓	
	Urmson et al. (2008)	Perception- Motion planer- mission planner- behavioural system.	✓	✓		✓	✓	✓	✓	✓	
	Montemerlo et al. (2008)	Perception-navigation-Behaviour.	✓	✓		✓	✓	✓	✓	✓	
	Bacha et al. (2008)	Perception-Planning-Driving behaviours.	✓	✓		✓	✓	✓	✓	✓	
NN as sub-systems	Chen, Seff, Kornhauser, and Xiao (2015)	CNN- logical control.	✓	✓		✓				✓	✓
	Al-Qizwini, Barjasteh, Al-Qassab, and Radha (2017)	CNN + logical decision making system.	✓	✓		✓			✓	✓	✓
	Kim and Canny (2017)	CNN- LSTM	✓	✓		✓					✓
NN as e2e	Pomerleau (1989)	NN	✓			✓					✓
	Bojarski et al. (2016)	CNN	✓			✓					✓
	Chen and Huang (2017)	CNN	✓			✓					✓
RL	Kendall et al. (2019)	CNN- Actor Critic DRL	✓			✓					✓
	Zhu et al. (2020)	Actor Critic DRL	✓						✓		✓
	Saleh, Hossny, and Nahavandi (2019)	Inverse RL- LSTM			✓	✓				✓	✓

3.3. Neural networks as an end-to-end system

Neural networks have also been used as end-to-end systems for driving agents. In such a system, the network architecture takes the input data (usually in its raw form) and outputs driving actions. One of the early examples of such an approach was “ALVINN” (short for Autonomous Land Vehicle In a Neural Network), Pomerleau (1989), which used a three layer neural network for real world road following based on inputs from camera images. The network was trained on simulated data as snapshots images generated from real world video data. The network was first tested in a simulation environment then it was deployed and tested in a real vehicle. More recently, NVIDIA (Bojarski et al., 2016; Chen & Huang, 2017) trained a Convolutional Neural Network (CNN) for real car driving. Real driving-data was collected using expert drivers. This data contained image inputs from cameras along with the corresponding driving actions of the steering wheel. A deep neural network was then trained using this data and deployed into a real car control environment. Although this approach shows high performance with manageable design and implementation efforts, it requires significant amounts of training data to be collected from driving in many diverse situations to be able to generalize to more than the most common traffic scenarios (Grigorescu et al., 2020).

3.4. Reinforcement learning

Reinforcement learning is a paradigm in which agents learn from (inter)acting with an environment. In this approach, the agent senses the world and explores different actions that can be carried out. The consequences of these actions are evaluated (for example, based on a reward signal), leading to learning based on positive and negative outcomes. If such an agent is trained in simulation first, the learning can, at least in some cases, be transferred into real situations (Sutton & Barto, 2018). Reinforcement learning has gained recent attention after out-performing humans in many Atari games (Mnih et al., 2013, 2015). There are different methods for implementing reinforcement learning, including Q-learning (Watkins & Dayan, 1992) and Actor–Critic (Konda & Tsitsiklis, 2000) approaches.

Reinforcement learning typically performs well in simulation environments. However, due to the differences between the simulation and the real world, deploying agents trained in simulation in the real world remains challenging and the focus of ongoing research (Dulac-Arnold, Mankowitz, & Hester, 2019).

In terms of recent work on training a self-driving car for real world deployment, Kendall et al. (2019) demonstrated a lane following task where the driving agent was trained using a real world dataset and a simple reward function that returned positive feedback as long as the agent did not cross the side of the road. The trained agent was then tested in a modified Renault Twizy. Zhu et al. (2020) used reinforcement learning to train velocity control for car following. Car following events were extracted from the Next Generation Simulation (NGSIM) dataset, used for both training and testing. The reward function was optimized for safe and comfortable speed control. An actor critic network was used, taking distance and velocity of both the ego car and the following car as input and returning range-bounded acceleration as an output.

Reinforcement learning approaches were also demonstrated in interactions with pedestrians. For example, Saleh et al. (2019) modelled an Inverse Reinforcement Learning (IRL) and bidirectional recurrent neural network architecture (B-LSTM) for learning detecting pedestrian intention and trajectory. The model was then evaluated on real world dataset for pedestrian behaviour.

Computationally, reinforcement learning carries out its learning by value function optimization in the sense that the model attempts to optimize a control function based on a given reward function. However, reinforcement learning is not the only optimization method for modelling maximization functions and other approaches, such as, dynamic programming, are also used in autonomous vehicle research. Lu et al. (2019), for example, used policy iteration adaptive dynamic programming to model optimal control function for steering control. A car following example (Zhu, Dai, Huang, Sun, & Liu, 2017) used actor-critic reinforcement learning with dynamic programming to model acceleration decision policy. da Silva and de Sousa (2010, 2011) used dynamic programming for motion control and path-following by optimizing the utility function.

Optimization techniques mainly focus on the computational algorithm and equation optimization of a value or utility function regardless of the car’s relation to the environment. These techniques can be explicitly coupled with reinforcement learning (eg. Lu et al. (2019), Zhu et al. (2017)) or studied independently (eg. da Silva and de Sousa (2010, 2011)). In the remainder of this paper, we focus primarily on reinforcement learning approaches when discussing optimizations since it more explicitly bridges optimization techniques with biological learning (Sutton & Barto, 2018).

Reinforcement learning, if carried out in suitable learning environments, allows the agent to explore, and learn from, the consequences of both carrying out the same actions in different conditions (such as driving in different weather conditions when the road is icy and slippery or dry and stable as well as if the weather is foggy and vision is unclear) and different actions under the same conditions (such as different strategies for avoiding a pedestrian on a crosswalk).

On the other hand, reinforcement learning is based on trial and error which makes it dangerous to be trained in real roads shared with other vehicles and vulnerable road users. Allowing car accidents for the purpose of training an algorithm is clearly not a tenable position. Therefore, another controlled environment is needed, such as a simulation environment to carry out the training. This is challenging because any such simulation needs to be of high fidelity and capture, as realistically as possible, situations the vehicle would encounter in the real world. Otherwise, the car may not be trained properly and behave in undesirable ways in new situations it encounters.

3.5. Current challenges

Despite the large effort from both academia and industry to develop self-driving car systems, there still exist a number of challenges.

One way to see why these challenges exist is to consider the classes of problems that current approaches address. The simplest scenario, but one that is often used in current work, occurs when a well-defined environment is given that the self-driving car needs to learn and act in it. This is equivalent to an uncertain Markov Decision Processes (uMDP) problem in which the agent has a full observation of the state and what it needs to know about the environment to take the optimum action (Bellman, 1957). There are many known solutions and algorithms for this type of problems (Mundhenk, Goldsmith, Lusena, & Allender, 2000).

However, the actual class of problems self-driving cars face are not equivalent to an uMDP. In a real application, information can be missing and other factors that the vehicle has no access to may interact with the decision, such as weather conditions affecting vision or occluded objects. Under such conditions, the problem to be solved is equivalent to an uncertain Partially Observable Markov Decision Processes (uPOMDP) (Spaan, 2012). In this type of problems, the state of the car in the environment and the probability of the transition function are uncertain. Unlike a computational environment where the transition function and probabilities can be modelled and tested, a real world transition function is more difficult to compute and verify. Exploring the different possibilities for optimization, practical challenges aside, makes this problem ExpTime-hard (Chatterjee, Doyen, & Henzinger, 2010).

Additionally, a self-driving car is not the only agent in a real world environment. Rather, it acts and interacts with other agents. This increases the complexity of the problem further into an uncertain Partially Observable Stochastic Game (uPOSG), one of the hardest classes of computational problems to solve (Horák & Bošanský, 2019). The more unknown parameters and conditions, the more intractable the problem becomes.

Another aspect is that learning is a never ending process (Parisi, Kemker, Part, Kanan, & Wermter, 2019). Just like humans continuously learn and accumulate knowledge, artificial agents need to continuously learn (Smith & Slone, 2017), even in the case of self-driving cars, as it is very difficult to train for and anticipate every possible scenario in an ever-changing environment. Some have suggested addressing the problem with continuous learning (Rusu et al., 2016) or to build upon previous skills in the form of transfer learning (Parisotto, Ba, & Salakhutdinov, 2015). In this, the challenge of catastrophic forgetting (where, in continuous learning, later information modifies the network weights such that previously learned aspects are forgotten even though they are still relevant) remains a general problem for neural systems that need to cope with large changing environments (Kirkpatrick et al., 2017).

We discuss how theories from cognitive systems can help in addressing these challenges in Section 5. First, however, we discuss the relevant paradigms in that field.

4. Cognitive paradigms and theories

While it is essential to look at how current challenges in self-driving car development are presently addressed, it is also relevant to investigate the underlying theory of these approaches. This is important insofar as new solutions are not always found by improving current techniques but, sometimes, by re-evaluating the theories and paradigms in which the work is carried out.

There are different ways to conceptualize the relationship between implementation and theory. For example, Guest and Martin (2020) proposed six different layers at which computational models can be built in psychological science with data from psychological experiments at the bottom and frameworks (outlining general assumptions of the nature of all the studied phenomena) at the top. The point being that “scientific inquiry can be understood as a function from theory to data and back again, and this function must pass through several states to gain explanatory force” (Guest & Martin, 2020, p.5). This multi-layered approach is not limited to the psychological sciences and we adapt this layered model in our discussion of self-driving cars. We focus on four layers: (starting from the bottom) Data, Implementation, Theory, and finally Paradigm (as shown in Fig. 3).

Although the interplay between cognitive paradigm and computational implementation is commonly discussed in cognitive science and cognitive systems research (Vernon, 2014), it has received less attention in recent work on self-driving cars. We suggest that awareness of different paradigms/theories and their possible influence on the implementation/data layer provides opportunities for the development of self-driving cars. The first steps to realize these opportunities are to describe the relevant cognitive paradigms and theories to model artificial cognitive agents (Section 4.1), describe the relation between cognitive paradigms and implementation technique (Section 4.2), and to establish a mapping between these paradigms and the approaches used in self-driving car development discussed previously in Section 3 (Section 4.3).

4.1. Cognitive paradigms

Vernon, Metta, and Sandini (2007) identify two main classes of cognitive systems: *cognitivist systems* and *emergent systems*. *Hybrid* systems, which include aspects associated with both of these classes are also possible (and in fact the dominant category at present Kotseruba & Tsotsos, 2020); one can therefore think of this as a continuous space whose extremes are defined by certain properties that we elaborate on here.

Cognitivism views cognition as symbolic inference based on a knowledge base and a set of rules, and is based on symbolic information processing. Physical world objects are mapped onto internal symbols used by the agent to represent the information about the world, and are processed based on these representations (Newell, Shaw, & Simon, 1958). The associations between symbols define rules used by the agent to infer information and behaviour. Using this approach, the agent is thought to be able to infer more knowledge about the world through logical deduction, allowing the agent to adopt the actions that lead to the intended goal. The knowledge-based rules determine how the agent behaves in certain situations. Another central assumption of cognitivism is the separation of perception and action, the idea being a one directional flow of information according to a sense-model-plan-act sequence in which actions are the mere output of central planning mechanisms (Hurley, 2001).

The *emergent paradigm*, meanwhile, emphasizes (Maturana & Varela, 1987) self-organization through which the system is continually re-constituting itself in real-time to maintain its operational identity through moderation of mutual system-environment interaction and co-determination. In a wide sense of the term, it thus emphasizes the interactive nature of agents and includes approaches inspired by situated and embodied cognition (Ziemke, 2003), as well as different forms of enactivism. In brief, enactivism asserts that "cognition is a process whereby the issues that are important for the continued existence of a cognitive entity are brought out or enacted: co-determined by the entity as it interacts with the environment in which it is embedded" (Vernon et al., 2007, p.157) . It places a strong emphasis on the embodiment of the agent since the shape and structure of the agent's body affect its perceptions and actions (Pfeifer & Bongard, 2006). Perception is thus not merely an input, and action not merely an output; rather they are fundamental aspects of the cognitive mechanisms (Clark, 1997).

Vernon (2014) considers five key elements for designing a system at the emergent end of the scale; autonomy, emergence, experience, sense-making, and embodiment. Autonomy relates to the ability of the agent to act and interact in an environment without being controlled by another agent (usually, an engineer). Emergence suggest that cognition and behaviour is not merely the outcome of a central planner, but the result of the dynamics of the situation in which the agent acts. Sense making refers to the relationship between the knowledge that the agent possesses and the interaction with the environment. It suggests that the knowledge is shaped by the interaction and it is generated by the system itself. Embodiment refers to the coupling between the agent and the environment. Ziemke (2003) points out that robots may be embodied in different senses of the term, that is, system designs or designers may differ with respect to what kind of body is required for cognition. To mention a few alternatives, some approaches may focus on the interaction and adaptation of the system to its environment (the structural coupling with the environment in the terminology of Ziemke), while other emphasize having a physical body with sensors and actuators affected by sensor noise and friction (the physical embodiment) or having organism-like bodies exhibiting similar sensorimotor capabilities as natural cognitive agents (the organismoid embodiment), or even focusing on the type of processes that constitute a living body (the organismic embodiment) (Ziemke, 2003).

4.2. Multi-layer model: from cognitive paradigms to data

We now describe how the paradigm just discussed may influence actual implementations of self-driving cars. As seen in Fig. 3, there are four layers (cognitive paradigm, theory, implementation, and data) and different instantiations at each layer, with cognitivist influenced instantiation to the left (represented by the colour green) and emergentist instantiations to the right (represented by the colour brown). Here, we focus on describing the relations between the layers. In the next section we address what determines the horizontal location of an instantiation. At the top layer are the cognitive paradigms ranging from cognitivism to emergent systems (Vernon, 2014). Computationalism, one layer below, is a function of cognitivism, and can be implemented with symbolic AI (yellow solid arrows), dependent on embedded a priori knowledge. Connectionism and enactivism, however, are influenced, by emergent approaches in differing ways leading to differences at the implementation and data layers. Thus, the emergent paradigm, influencing connectionism at the theory layer, can lead to neural network implementations with larger control exerted by the designer in more restricted environments (red dashed arrows). The emergent paradigm might also lead to other types of implementation by the emergent paradigm's focus on self-organization and autonomy (blue dotted arrows). An example is coupling it with notions taken from enactivist theories to the degree that deep reinforcement learning using neural networks can accommodate these.

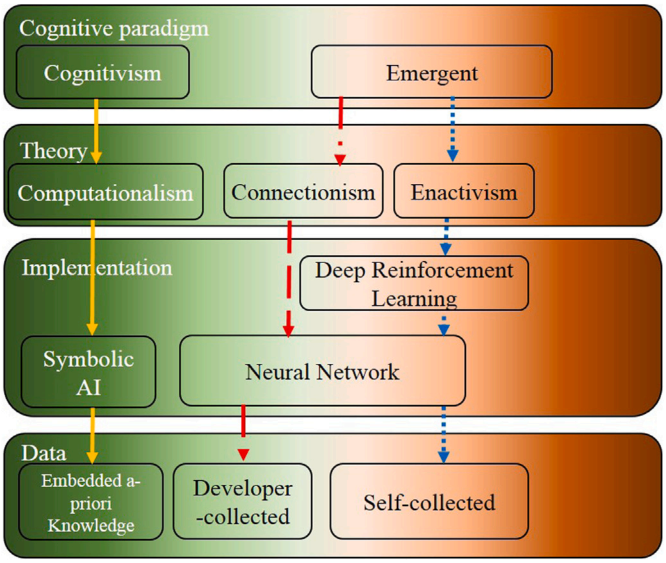

Fig. 3. From cognitive paradigms and theories to implementation and data, inspired by the work of Guest and Martin (2020). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Different implementation techniques also require data to be in compatible formats. For symbolic AI, knowledge and rules need to be provided to the system. These are usually chosen by the developer and they necessarily reflect the developer's mindset or may be limited to what data is available. Approaches such as reinforcement learning, meanwhile, interact with the environment and collect the data as experiences, and are in that sense, less dependent on explicit human input. Nevertheless, we do acknowledge that there are today always some selection bias from human designers.

4.3. Characterizing implementations based on their cognitive characteristics

The main point of this paper is to suggest that one side of Fig. 3, i.e., the (right) emergent paradigm side of the model, is underexplored in self-driving car systems and could be explored to approach some of the hard problems with self-driving cars. However, it may not be obvious what makes one implementation emergentist or not. Thus, to be able to systematically compare the different approaches (i.e., symbolic AI, neural network as subsystem, neural networks as end-to-end learning, and reinforcement learning, see Section 3 and Table 4), we make use of the ten cognitive characteristics proposed by Vernon et al. (2007) and Vernon (2014): *computational operation, representational framework, embodiment, perception, action, anticipation, adaptation, motivation, autonomous,* and *social cognition*. In the following, we first introduce each characteristic and define what a cognitivist and emergent implementation of each could consist of. In addition, we introduce a mid-point between the two ends of the spectrum. For visualization purposes, we map this on the same colour scheme used in Fig. 3, where dark green represents the cognitivist side to dark brown that represents the emergentist side (see Fig. 4). After analysing the self-driving car systems in each paper (Section 3), we determined the appropriate colour shade/paradigm for each characteristic. We then constructed a table consisting of the different papers and the cognitive characteristics, and finally applied the assigned colour to each cell. The

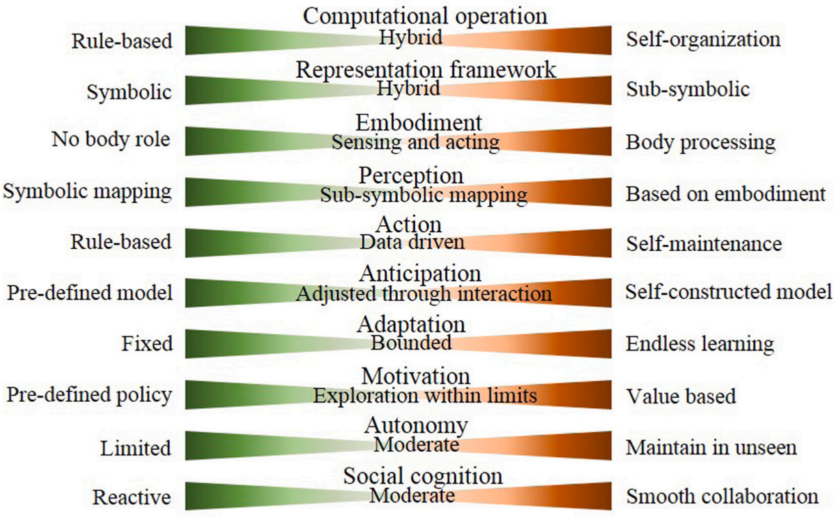

Fig. 4. The spectrum from cognitivism (green) to emergent systems (brown) for each of the ten cognitive characteristics. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Comparison of approaches for self-driving cars based on ten cognitive characteristics. Dark green cells indicates a cognitivist focus and dark brown cells indicates an emergent systems focus for each of the characteristics and approaches. Note that, for the present purposes, we do not distinguish between work that addresses the whole problem of self-driving and work that addresses specific aspects of the intelligent subsystem. The point of this table is merely to illustrate the currently dominant approaches. (For interpretation of the references to colour in this table legend, the reader is referred to the web version of this article.).

Comparison factors	Symbolic AI				Neural network						RL		
	Oka et al. (1999)	Urmson et al. (2008)	Montemerlo et al. (2008)	Bacha et al. (2008)	Chen et al. (2015)	Al-Qizwini et al. (2017)	Kim and Canny (2017)	Pomerleau (1989)	Bojarski et al. (2016)	Chen and Huang (2017)	Kendall et al. (2019)	Zhu et al. (2020)	Saleh et al. (2019)
Computational operation	█	█	█	█	▒	▒	🟧	⬜	🟧	🟧	🟫	🟫	🟫
Representation framework	█	█	█	█	▒	▒	⬜	⬜	⬜	⬜	▒	▒	▒
Embodiment	▓	▒	▒	▒	▒	▓	▒	▒	▒	▒	▒	▒	▓
Perception	▓	▓	▓	▓	▒	▒	▒	▒	▒	▒	▒	▒	▒
Action	█	▓	▓	▓	▓	▒	▒	▓	▓	▒	░	⬜	⬜
Anticipation	█	█	█	█	▒	▒	🟧	⬜	░	░	░	⬜	🟧
Adaptation	█	█	█	█	▓	▓	▒	▒	▒	▒	▒	▒	▒
Motivation	█	█	█	█	▓	▓	▒	▒	▒	▒	▒	▒	▒
Autonomy	█	█	█	█	▓	▓	▒	▒	▒	▒	▒	▒	▒
Social cognition	NA	▓	▓	▓	▒	▒	▒	NA	NA	NA	NA	▓	▓

resulting table thus visualizes to what extent cognitivism or emergent characteristics are present in the different approaches and specific implementations (see Table 5).

4.3.1. Cognitive characteristics

Here we briefly describe how cognitivism and emergentism differ on ten different characteristics as defined by Vernon et al. (2007) and Vernon (2014). The different characteristics are summarized in Fig. 4.

Computational operation refers to the nature of the computations performed by the system. On the cognitivist end, they are the rule-based syntactic transformation of symbol tokens, while on the emergent end, they are self-organized patterns of computations in a distributed network.

Representation framework refers to the relationship between the agent's knowledge of the external world and the agent's internal world. On the cognitivist end, the designer feeds the system with the required knowledge about the world using symbolic annotation (e.g. Newell (1990) and see perception below). On the other end, a system can self-organize based on its interaction with the environment in a way that may not require explicit representations; for example if a sub-symbolic end-to-end neural network is used.

Embodiment refers to how the physical body – e.g. the properties of the actual platform used – affects cognitive mechanisms, including how the agent interacts with and senses the world. The cognitivist paradigm puts little focus on embodiment, which would mean that any architecture could simply be transferred onto a different platform with no consequences, while the emergent paradigm suggests behaviour will be highly co-determined by the body that the architecture is implemented in. However, as previously discussed there are different levels of embodiment, which highlight different aspects of the body of the system in question.

Perception from a cognitivist perspective is a direct mapping between what the agent senses from the outer world and the symbolic representation defined for the agent. Dorffner (1999) describes this view as based on an objectivist epistemology:

> The underlying view is an objectively existent outside world which must be mapped onto a faithful image in the cognitive agent in order for the latter to act intelligently. (…) For instance, to say that a symbol 'CHAIR' represents the category of chairs, one must not only specify the symbol, but must also assume that a category chair exists in the world, independently from whether the observer or the agent to be modelled interacts with the world (Dorffner, 1999, p. 24).

This implies that the perceptions of the world are formed independently of the individual subject, and only establish a reference to something in the external world (cf. also Newell, 1990 Newell's law of representation). A consequence of this is that perceptions are thought to be independent of action, forming what has been called a sandwich model of cognition (Hurley, 2001). Emergent paradigm breaks the separation of action and perception and focuses on how the agent's actions and interaction with the world create the inner world of the agent (Clark, 1997). As pointed out above, the particular embodiment of the agent may thus also influence the perceptions.

Action refers to how the agent changes the outer environment. In the cognitivist case, actions are only the end product of cognition computed according to some sequential procedures that take the agent from the current state to the goal state. The agent computes the changes between states based on the rules of symbols it holds that takes it to the defined goal. In emergent paradigm, often, the starting point of cognitive development is action where action may also serve cognitive purposes (cf. e.g. Clark (1997)). Emergent paradigm also emphasizes the continuous interaction with the environment based on representations tied to action. Thus, an important distinction, from the emergent paradigm perspective, is between the reactive nature of the controller/control system and the (externally observable) reactive behaviour of the agent (e.g. Nolfi and Floreano (2000)). The distinction can be seen in Chapman and Agre's (1989) two meanings of planning. First, "to plan" can have the general meaning of reasoning about action without the mechanism. For example, a slime mold's behaviour might be described as goal-directed even though the slime mold has no encoding or representation of its goal (cf. Anderson and Rosenberg (2008), Von Uexküll (1992)). Second, planning can be used in a more restricted sense referring to the process of constructing plans (or programs) to be executed in a step-wise manner (Chapman & Agre, 1989). Planning in the second more restricted sense is independent of the general meaning of planning because there could be other means than the construction of step-wise plans to achieve planning in the first general sense. Emergent paradigm emphasizes that planning in the first sense may often be the result of simpler mechanisms but as seen in the next section have also envisaged new types of mechanisms for planning in the restricted sense.

Anticipation refers to how the agent predicts the next state. A cognitivist view of anticipation sees the system as making a plan to reach a goal, by having a predefined state space where the possible states are logically connected and predictable. The agent can link the sequence of next states by predicting the change in symbol representations. Thus, in this sense anticipation is planning (in the restricted sense, as previously defined). An emergent view, often sees anticipation as an intrinsic aspect of brains and cognition (Bar, 2009). The focus of emergent paradigm has often been on continuous online interaction with the environment, but it is clear that cognition is also geared for mental time travel and thinking about future states not given by the current environmental situation, so called off-line cognition (cf. Clark and Grush (1999)). From an emergent perspective off-line cognition may be seen as based on the off-line reactivation of sensorimotor processes rather than a separate system (Grush, 2004; Hesslow, 2002, 2012; Möller, 1999).

Adaptation refers to how the agent develops or learns. One the one extreme, cognitivist approaches view learning as loading a priori knowledge for the agent and then build upon the knowledge structure. On the other, emergent approaches treat learning as changing the internal state during the process. In emergent systems, the agent reorganizes its topology or structure as a response to the change in the environment every time the agent gets a feedback by interaction to adapt to the change.

Motivation is what drives the agent to take an action, including influencing attention and adaptation. A cognitivist view of motivation is based on a given criteria that associates preferred future states with specific actions. In this case, the agent takes an action based on what the current state is and what the desired state should be. An emergent interpretation would be to fulfil certain values. The agent does not have a direct mapping between the current state and the desired state, but have a value that it needs to maintain where there could be several ways of satisfying this value.

Autonomy refers to the ability to freely interact in an unknown environment, which varies from, in the cognitivist case, being less relevant and restricted to a particular environment to, in the case of emergent systems, being free to interact and adapt in unencountered environments.

Social cognition is the extent to which the agent is perceived as part of the environment, where it interacts with the surrounding and other agents. In order to interact or collaborate with other agents, the agent needs to act in support of the goals of the other agents or the shared goal. This involves among other things reading faces, recognizing emotional experiences and detecting eye gaze. The agent's ability to have an effective social interaction with other agents depends on its ability to interpret information about other agents' activities and intentions. Reading intentions can be categorized into low-level intention associated to movement and high level intention associated with actions and motives (Vernon, 2014). From a cognitivist perspective other agents are mere inputs and have lesser impact on the individual behaviour, whereas on the emergent end of the spectrum social agents together create a shared understanding of the situation and the other social agents forms a new situation with distinct dynamics that could not be observed with a single agent.

It should be noted here, that the purpose of contrasting the paradigms is to establish differences between paradigms within cognitive systems research and how they might influence the design of self-driving cars. Note, in this paper we are not arguing for a particular approach, instead we aim to highlight how emergent paradigms may be underexplored in the development of self-driving car systems. We also provide some suggestion regarding which the emergent system paradigm might be most useful.

5. Opportunities for self-driving cars and cognitive systems research collaboration

To summarize, so far we have identified four common approaches the design of self-driving cars (symbolic AI, neural network as subsystem, neural network as end-to-end learning, and reinforcement learning), described how the cognitivist paradigm differs from the emergent systems paradigm, and described a model for how the paradigms might influence the type of techniques and data used in the design of the self-driving car system. Table 5 reflects and visualizes to what extent the 13 papers (discussed earlier in Section 3) contain aspects of the systems akin to cognitivist thinking or emergentist thinking. For the social cognition category, if the agent proposed in the paper runs in an environment where no other agents are included we mark the characteristic with Not Applicable (NA) in the table. We have also collapsed the neural network as subsystem and neural network as end-to-end learning approaches to the heading Neural Network to match the three implementation classes in Fig. 3.

It is evident from our analysis (and indicative of a visual inspection of the table) that, overall, cognitivist (dark green) and connectionist (lighter green) perspectives dominate current work as only 8 out of 130 cells can be said to be right of centre of the spectrum, i.e., including characteristics that are representative for the emergent system paradigm. As also illustrated in the layered model (Fig. 3), the papers representing a symbolic approach are cognitivist on most characteristics, the deep learning approaches could be seen as connectionist including both cognitivist and emergent system aspects, and finally the reinforcement learning paradigm includes more aspects of the emergent system paradigm than the two others. The reinforcement learning approach still includes many characteristics that are more toward the cognitivist side of the spectrum.

In more detail, the computational operation characteristic shows more diversity and seven of the papers exhibit some aspects representative of the emergent system paradigm. Five of the papers do not include other social agents. The absence of other social agents could be seen as indicative of a cognitivist approach per se, as the emergent systems paradigm emphasizes the interplay with the environment including other social agents for cognition. However, it is difficult to determine if the absence of social agents was mainly a methodological and pragmatic choice, rather than being a consequence of the cognitivist paradigm. The lack of emergent system characteristics in the approaches and particular system implementations suggests that there are still opportunities for exploring different paradigms, not the least the right side of our model visualized in Fig. 3 where the emergent paradigm and enactivist theory resides. We explore some of these opportunities in this section.

5.1. Vision-based perception system and rare scenarios

It is quite evident in the field that vision has been of central importance for the development of self-driving cars. Perception in current work tends to be largely cognitivist with hybrid approaches (similar green shades in Perception in Table 5), mainly using neural networks for object detection and recognition. Visual sensory data is considered one of the key elements of the perception system for self-driving cars and the main channel of observing the outer world is through 2D cameras, often supported with LIDAR. With the availability of different annotated datasets such as road object detection (Geiger et al., 2013), bird eye views of road[1] and pedestrians[2] as well as the rapid development of ML and CNN for image processing (Al-Qizwini et al., 2017; Chen et al., 2015), self-driving cars have shown high performance in detecting and categorizing the surrounding objects. Despite this massive development in computer vision for object detection, difficulties and failures in rare situations are still common (Grigorescu et al., 2020).

One of the challenges for perception is the need for training on large amounts of data to learn how to handle the input data, which has significant consequences for self-driving cars. The nature of the sensory data to be collected is more than just dispersed images of objects but a composition of a real world scenes and scenarios. For a level five self-driving car, the number of different possible scenarios is huge and many situations are unique or so rare they are difficult to identify (Grigorescu et al., 2020). These rare cases are not only difficult to collect but also difficult to train on. When rare cases are not well represented in the data-set, the learning mechanisms may not be able to identify them as important, for example, sorting them as noise. In the actual traffic situations, rare cases, such as fatal accidents, are usually the most important ones and require careful handling because they may cost lives (Da Lio, Doná, Papini, Biral, & Svensson, 2020).

Several approaches have been proposed to solve the less-present-data problem in general (Feldman, 2020) or for specific applications (e.g. recommender systems Park & Tuzhilin, 2008), nevertheless the field is still understudied (Johnson & Khoshgoftaar, 2019). Current approaches for solving this challenge are either to work on the sample data level and introducing the model to more of the rare samples, or to work on the decision making level by giving additional weight or value to less represented samples.

Taking an emergent perspective may suggest new ways to conceptualize and address perception in autonomous system, for example, by emphasizing the interactive perspective and that perception is for action (Hurley, 2001). In embodied AI experiments with simple robots, this has been demonstrated to be a useful perspective and has been described as a form of sensori-motor coordination, that is, behaviours where an agent structures its sensory input through interaction with its environment (Hallam, Dario, Jean-Arcady, & Gillian, 2002; Scheier, Pfeifer, & Kunyioshi, 1998). For example, Scheier et al. (1998) showed that a robot that learned to circle objects could differentiate between large and small objects, despite the fact that its sensory system could not do such a categorization by itself. In addition to the structuring of sensory input, the active structuring of its own environments is a pervasive characteristic of humans and other animals emphasized by embodied and enactive theories of cognition. Ziemke, Bergfeldt, Buason, Susi, and Svensson (2004) demonstrated with a simple simulated robot experiment that a task requiring memory could be solved by a reactive (memory-lacking) agent evolving a strategy of placing road-signs in its environment that perturbed its behaviour so that it could reach the goal position. While not immediately transferable to the design of self-driving car systems, they indicate different ways in which taking an emergent perspective have given rise to new design solutions.

Another solution inspired by embodied theories (Hesslow, 2002; Svensson, Thill, & Ziemke, 2013) directly applied to research on self-driving cars suggests that it is possible to take the on-line driving experiences off-line (while the car is not under operation) and enable the car to “dream” about unseen and uncommon situations derived from its online experiences and data that has been collected from the real driving (Da Lio et al., 2020; Da Lio et al., 2017; Plebe, Papini, Donà, & Da Lio, 2019).

5.2. Understanding human behaviours and intentions

One of the takeaway lessons from the DARPA 2007 challenge winner (Urmson et al., 2008) is that driving is a social activity that involves understanding other agents’ behaviours and intentions. Although not trivial, several driving tasks such as lane keeping, overtaking, and adapting to other cars on highways have been shown to be handled by self-driving cars at least to some degree. However, once the traffic situation requires a meaningful interaction with other vehicles or pedestrians, which at least in the case of a human driver would require a need to identify others and communicate one’s own intentions, the abilities of handling the situation drops significantly. In these more complex situations, where the traffic rules and guidance of road markings and the like is not enough, factors such as culture, eye contact, body language, feelings and empathy play a large role.

A self-driving car would therefore be more like a social agent that interacts with other social agents. Although there are several approaches that try to solve the problem of interacting with other vehicles and vulnerable road users as an intention prediction problem, the dominant approach in Table 5 is toward the cognitivist perspective: social cognition is reduced to predicting the trajectory of the other agent. The interactive aspect, namely that the participating agents both act to mutually create an understanding of the situation is still less explored in these approaches. For example, a pedestrian who approaches or is waiting at a crossing area may be perceived as intending to cross. However, if the pedestrian does not actually have the intention to cross this may lead to the autonomous car (designed with single dynamical model) to unnecessarily slow down or stop, potentially creating a traffic disturbance. To avoid this, an autonomous driving system needs to be able to appropriately interpret the pedestrian intention and behaviour, not just the kinematics of the movement. Indeed, research on pedestrian crossing behaviour suggests that it is determined by a mix of pedestrian factors (e.g. social norms, age, walking speed) and environmental factors (e.g. vehicle speed, lighting, traffic flow) (Rasouli & Tsotsos, 2020). While efforts on building more complex sensor systems (cf. e.g. Konrad, Shan, Masson, Worrall, and Nebot (2018)) might advance the perception system, pedestrians sometimes rely on eye contact with the driver in which case autonomous vehicles also need to have some way of communicating its intentions (Rasouli & Tsotsos, 2020).

Social interaction is thus better characterized as a dance than as a transfer of information (Lindblom, 2015). There already are enactivist models that demonstrate the core mechanisms in this regard, for example when it comes to coordination between agents (Di Paolo, Rohde, &

[1] http://cvgl.stanford.edu/projects/uav_data/.

[2] http://www.gavrila.net/Datasets/Daimler_Pedestrian_Benchmark_D/daimler_pedestrian_benchmark_d.html.

De Jaegher, 2010; Froese & Di Paolo, 2008). In autonomous vehicle research, efforts to address this have, for example, used Bayesian models to account for different possible intentions. Hashimoto, Gu, Hsu, and Kamijo (2015), for example, proposed a Dynamic Bayesian Network with Bayesian filtering framework for predicting the pedestrian intention at the crosswalk area. Quintero, Parra, Lorenzo, Fernández-Llorca, and Sotelo (2017), on the other hand, used a Hidden Markov Model for intention recognition along with the body language. In their study, intention recognition is not restricted to finding the maximum similarity of the current observation with the pedestrian motion sequences; rather, it also evaluates how the intention of the pedestrian has evolved. Understanding the pedestrian's intention includes predicting the next behaviour of the pedestrian. Li et al. (2019) conducted a comprehensive study of the approaches for predicting pedestrian trajectories especially for long sequences.

Modelling self-driving cars to smoothly interact with other human agents therefore requires both technological development and insights from cognitive systems research, simply because this is fundamentally an interaction between two cognitive agents. Similar to findings in social cognition research (De Jaegher, Di Paolo, & Gallagher, 2010), and in line with the aforementioned dance metaphor, all aspects of this interaction (including the interaction dynamics themselves) are relevant and need to complement the technological development such as sequence recognition techniques as presented by Hashimoto et al. (2015).

5.3. Driving styles, motivations, and embodiment

Another social aspect is understanding the motivation behind the decision-making of other drivers. Motivation in driving can be reflected in the driving style. Different drivers have different driving styles (Sagberg, Selpi, Bianchi Piccinini, & Engström, 2015). The driver's personality trait and emotions (Poó & Ledesma, 2013), as well as age and experience (Miller & Taubman-Ben-Ari, 2010) plays a big role in defining the driving style.

For self-driving cars, understanding motivation affects both the design of the driving and understanding the motivation of the other road participants. For the development of self-driving cars, as shown in Motivation in Table 5, most of the literature undervalues the factor of motivation in the architectural design of the agent (represented as the green shade which is either predefined or within limits). Some studies of self-driving cars do point to the idea of having autonomous agents with different driving styles and adapt to the preferences of the current users of the self-driving car (e.g. Kraus, Althoff, Heißing, and Buss (2009), Kuderer, Gulati, and Burgard (2015)).

However, motivation, as viewed from an emergent perspective, does not involve a direct mapping that associates preferred future states with specific actions, but is rather something that comes from within. Borrowing an example from the psychologist von Hofsten (2004) highlights this point:

> For example, before infants master reaching, they spend hours and hours trying to get the hand to an object in spite of the fact that they will fail, at least to begin with. For the same reason, children abandon established patterns of behaviour in favour of new ones. For instance, infants often try to walk at an age when they can locomote much more efficiently by crawling. In these examples there is no external reward. It is as if the infants knew that sometime in the future they would be much better off if they could master the new activities.

Thus, the notion of motivation is not necessarily tied to external rewards or a mapping to an explicit future state, but it is inherent in the design of the agent, and this has implications for the designers of such systems because it clarifies where the decision to prefer one driving style over another in a specific context takes place.

It is worth noting here that the design of the agent, in principle, extends to the specific nature of the body that it has. As previously mentioned, an emergent view on embodiment entails that the embodiment of an agent cannot be separated from the control system of that agent. Some researchers in the field of situated and adaptive/evolutionary robotics have taken this to mean that it is intrinsically difficult or impossible for a human designer to design the control system from the robots point of view, but that the robot must itself generate its own representations of the world and let the control system adapt to the embodiment of the agent (Pfeifer & Bongard, 2006). Thus, to some extent what is good data and training scenarios for a self-driving car may be different from that of a human learning to drive, and thereby affect what the agent considers to be reasonable driving styles. Reinforcement learning approaches do, to some extent, address this aspect as the data is self-collected by the agent from the environment along with the corresponding reward, as opposed to supervised learning approaches (Barto & Dietterich, 2004) with the data collection mainly conducted and labelled based on human involvement. However, taking inspiration from emergent paradigms in adaptive/evolutionary robotics would suggest finding design methods that take the human even more out of the loop.

5.4. User experience in self-driving cars

Although our main focus is on the development of the intelligent system of self-driving cars (and the relevance of cognitive systems research in this context), it is worth pointing out that research in the cognitive domain is also relevant for another aspects of autonomous vehicles: the human pilot, and thus the user experience aspects of self-driving cars. This includes, for example, the interaction between the system and the user in partial autonomy driving (e.g. SAE levels three and four), in which a human driver remains involved in the driving task, as well as the user experience of a fully autonomous driving system (SAE level five), in which no human driver is involved in the driving process.

Trust and safety are key elements for critical systems as autonomous vehicles. That is, the users trust that the driving agent is intelligent enough to safely drive in complex and unpredictable traffic environments. A system that is perceived as not trustworthiness may not be used in the proper way or not at all (Raats, Fors, & Pink, 2020). While great effort is invested in developing autonomous vehicles with high capabilities, the field of autonomous vehicles needs also to address the aspects of user experience, in particular the user perception of a trustworthy system.

Psychological studies on autonomous driving user experience show that the appearance of the driving system can, for example, influence the user's perception of the trustworthiness of the system. A study by Lee, Kim, Lee, and Shin (2015) showed that a human-like appearance of the driving system increases the trustworthiness for how the user perceives the driving agent's intelligence.

Further, the importance of including end users (and an understanding of how they perceive AI systems) into the algorithmic development has been highlighted recently (Shin, 2021a). Without such efforts, the autonomous vehicle can be in danger of being perceived as a "black box" even though the importance of aspects such as transparency and explainability in this context are well understood (Shin, 2021b).

Until full automation is reached, the human driver will need to interact with the driving system. For example, when the autonomous system is unable to handle a driving situation, the system needs to shift control from the intelligent system to the human driver. This opens the opportunity for the field of self-driving cars user experience to study the process of handling the driving responsibility exchange from and to the intelligent driving system from the interactive perspective (Walch, Lange, Baumann, & Weber, 2015) as well as the legal and ethical considerations (McCall, McGee, Meschtscherjakov, Louveton, & Engel, 2016).

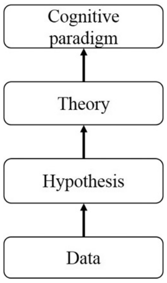

Fig. 5. From data and hypothesis to cognitive paradigms and theories.

In addition, the driver assistance system needs to predict the driver's intention for smooth interaction. An example of level two SAE is a system that predicts the driver intention to change lane and warns the driver for another car in the blind spot. If the assistance system falsely predicted that the driver intends to change lane while not intended, the system would give a false alarm that may be irritating for the driver if frequently repeated and the driver may then turn off the assistance system (Wen, Zhang, Wang, & Han, 2015).

5.5. *The relevance of self-driving cars in cognitive systems research*

So far, we have discussed how the development of self-driving car systems can benefit from inspiration from work in cognitive systems, in particular within emergent paradigms. However, it is also important to highlight the converse: cognitive systems research can benefit from work on self-driving cars.

For example, in cognitive systems research, agents are often studied in simplified environments, either in simulation, or, if a robotic agent is used, in a constrained experimental setting. Self-driving cars, by contrast, necessarily have to deal with a rich environment with various intentional interactions. From a cognitive systems perspective, a self-driving car is thus an agent that deals with a rich environment, but retains simplicity in terms of control since they are characterized by two degrees of freedom (longitudinal and lateral control). Most, if not all, work in cognitive systems research that has made use of wheeled robots (Nolfi & Tani, 1999; Scheier et al., 1998; Tani & Nolfi, 1999; Ziemke et al., 2004) to model cognitive mechanisms could be further studied in autonomous vehicles. As is evident throughout this paper, these mechanisms are all relevant for autonomous vehicles, and the degree to which they can be successfully implemented can lead to new theories in cognitive systems via a bottom-up path in Fig. 5 (see also Guest and Martin (2020)).

6. Discussion and future work

Self-driving cars are interesting for both robotics and AI as well as cognitive systems research because they are autonomous systems that act and interact with other autonomous agents in real life situations. They need to solve the full range of tasks that other cognitive systems need to solve in the real world and should therefore also be understood as a cognitive system in their own right, not the least because it opens the potential for new collaborations between the traditional AI domain and cognitive systems research.

While significant work has been done in the field of self-driving cars, the largest focus has been on improving the hardware components for sensors, such as cameras and LIDARs, or software algorithms for localization, object detection or trajectory planning (Pendleton et al., 2017). Methods such as dynamical modelling are suitable for deterministic problems but they show shortcomings in more versatile application like self-driving cars (Hashimoto et al., 2015). We argue that the development of self-driving cars requires more than the technological development of algorithms. This includes several aspects.

The intelligent driving system, responsible for decision making, needs to handle various of situations. While the designers and developers may be able to model in advance some of these situations or provide data for them, the system needs to handle rare situations that are less encountered (Da Lio et al., 2020; Da Lio et al., 2017). Those are usually the most dangerous and important ones. Current techniques of machine learning such as deep learning lack the ability to cover this need. Such situations require the system to act with cognitive abilities such as autonomy and adaptability. In cognitive systems this is largely studied as an emergent system in which the system focuses on the relation between the agent and the environment. Exploring emergent systems for self-driving cars opens opportunities for various research directions regarding how to build an intelligent driving system that learns and adapts in real world situations. Learning from dreaming is an example of the innovations for cognitive inspired techniques for self-driving cars. This research direction studies how humans learn by dreaming of un-encountered experiences to improve the performance in real world situations. We intentionally left free space in Fig. 3 to indicate that further theories may still emerge, for example by formalizing enactivist theories further. Accordingly, additional implementation approaches may come forth from these theories that demonstrate enactive systems. The space in the data layer suggests the ability to have new opportunities in the knowledge/data used for learning and interaction with the environment. Although the current self-collected data represent what enactivists claim about self-constructed models, it is still bounded by the developers' modelling. We suggest that additional theories and implementation techniques may require a shift in how the data is perceived and represented. Again, this highlights the symbiotic potential between autonomous vehicle development on one hand and cognitive systems research on the other.

Another large research area that requires cognitive systems studies in self-driving cars is the social interaction. Self-driving cars are social agents that interact with other road users that could be vehicles or pedestrians. The interaction with other road users requires a highly developed understanding of their intentions and needs. This involves body language categorization and understanding to predict the intention of the different road users in the scene and how to act accordingly. This is one of the difficult aspects of self-driving cars because the same observable may originate from different intentions. For example, a pedestrian standing in a crosswalk area may have the intention to cross or just to wait. Assuming that every pedestrian presented in the crossing area has the intention to cross may lead into traffic disturbance, while misunderstanding the intention of crossing may cost a human life. Therefore, wide study of human interaction is required for developing social cognition for self-driving cars.

While many techniques have shown successful outcomes for the different aspects of self-driving cars, the narrow focus on techniques may not lead to the desired progression. For example, prediction is a rising topic in modelling other agent's behaviours, such as predicting trajectories or intentions. Predictions that are only based on data analysis and categorization may be misleading without taking human cognition into account (Quintero et al., 2017). Prediction needs to be considered from a cognitive system point of view along with considering the technical aspects. For example, taking into account the emergent system characteristic of perception being for action, suggests prediction may not be seen as only a trajectory but also for how it is integrated and related to the control of the car. Another example is using synthetic data generation for training self-driving cars on less encountered experiences. There are techniques for generating synthetic data such as Generative Adversarial Networks (Goodfellow et al., 2014) that learn to generate new data by combining features from the original dataset. Initial work has been done to train driving agents with this type of synthetically generated data (Ha & Schmidhuber, 2018; Santana & Hotz, 2016). However, real world applications like self-driving cars need cognitively inspired techniques for generating the synthetic data to avoid generating dangerous unrealistic scenarios. This opens the opportunities for additional research in how to use current techniques for cognitive systems. An example is how to generate meaningful synthetic experiences for training purposes with the least human crafting of these situations.

We claim that the collaboration of self-driving cars development and cognitive systems research does not only benefit the former but also the latter. Cognitive system research lacks real world applications for cognitive studies beyond lab bounded robotic systems. The field of self-driving cars automatically and naturally extends the typical embodied AI experiment to a realistic setting with high complexity where the role of autonomy, emergence, experience, sense-making, and embodiment cf. Vernon (2014) can be investigated in new ways. For example, it introduces the cognitive system to situations where aspects of human trust and acceptance are part of the context, to interactions with other social agents in the form of self-driving car, human driver, and vulnerable road user interactions, and to interactions with infrastructure (e.g. vehicle-to-vehicle and vehicle-to-X communication Dey, Rayamajhi, Chowdhury, Bhavsar, & Martin, 2016; Harding et al., 2014).

7. Summary and conclusion

In this paper, we first presented the field of self-driving cars from an AI and machine learning point of view. We then introduced ten cognitive characteristics to describe to what degree self-driving car systems have properties that are either compatible with or influenced by the cognitivist paradigm or the emergentist paradigm. By applying the ten characteristics on thirteen papers describing real world implementations of self-driving car systems, we showed that on most characteristics there is a prevalence of cognitivism. We then tried to show that there is an unknown but possible opportunity to exploit the emergentist side of the characteristics in self-driving car system development by pointing to some specific examples of how embodied AI experiments previously have exploited principles from the emergent paradigm to design new solutions. Although we have primarily highlighted emergent paradigms as a likely candidate, many of the ideas in that field, particularly around enactivism, remain at a theoretical level and, as such, how to transform the theory to the implementation layers below is an open question. Self-driving car systems do offer an opportunity to demonstrate these ideas in practice by providing an ideal agent, situated in the real world, that can demonstrate different implementations. As just pointed out in the previous section there is also a bi-directional flow of information between building self-driving cars and building a better understanding of natural cognitive systems. To conclude, we suggest that advances in both domains can be made if this potential is acted upon.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Al-Qizwini, M., Barjasteh, I., Al-Qassab, H., & Radha, H. (2017). Deep learning algorithm for autonomous driving using googlenet. In *2017 IEEE intelligent vehicles symposium (IV)* (pp. 89–96). IEEE.

Anderson, M. L., & Rosenberg, G. (2008). Content and action: The guidance theory of representation. *The Journal of Mind and Behavior*, 55–86, Publisher: JSTOR.

Bacha, A., Bauman, C., Faruque, R., Fleming, M., Terwelp, C., Reinholtz, C., et al. (2008). Odin: Team victortango's entry in the darpa urban challenge. *Journal of Field Robotics*, *25*, 467–492, Publisher: Wiley Online Library.

Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., et al. (2020). Self-driving cars: A survey. *Expert Systems with Applications*, Article 113816, Publisher: Elsevier.

Bar, M. (2009). The proactive brain: memory for predictions. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *364*, 1235–1243.

Barto, A. G., & Dietterich, T. G. (2004). Reinforcement learning and its relationship to supervised learning. In *Handbook of learning and approximate dynamic programming, vol. 10*. Publisher: New York: Wiley-IEEE Press.

Bellman, R. (1957). A Markovian decision process. *Journal of Mathematics and Mechanics*, *6*, 679–684, Publisher: JSTOR.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., & Goyal, P. others (2016). End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316.

Bresson, G., Alsayed, Z., Yu, L., & Glaser, S. (2017). Simultaneous localization and mapping: A survey of current trends in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, *2*, 194–220, Publisher: IEEE.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, *2*, 14–23, Publisher: IEEE.

Campbell, M., Egerstedt, M., How, J. P., & Murray, R. M. (2010). Autonomous driving in urban environments: approaches, lessons and challenges. *Philosophical Transactions of the Royal Society of London A (Mathematical and Physical Sciences)*, *368*, 4649–4672, Publisher: The Royal Society Publishing.

Chapman, D., & Agre, P. (1989). What are plans for? *Designing Autonomous Agents*, 17–34.

Chatterjee, K., Doyen, L., & Henzinger, T. A. (2010). Qualitative analysis of partially-observable Markov decision processes. (pp. 258–269). arXiv:0909.1645 [cs] ArXiv: 0909.1645 6281.

Chen, Z., & Huang, X. (2017). End-to-end learning for lane keeping of self-driving cars. In *2017 IEEE intelligent vehicles symposium (IV)* (pp. 1856–1860). IEEE.

Chen, C., Seff, A., Kornhauser, A., & Xiao, J. (2015). Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE international conference on computer vision* (pp. 2722–2730).

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. Cambridge, MA: MIT Press.

Clark, A., & Grush, R. (1999). Towards a cognitive robotics. *Adaptive Behavior*, *7*, 5–16, Publisher: Sage Publications Sage CA: Thousand Oaks, CA.

S. O.-R. A. V. S. Committee (2014). Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. *SAE Standard Journal*, *3016*, 1–16.

Da Lio, M., Doná, R., Papini, G. P. R., Biral, F., & Svensson, H. (2020). A mental simulation approach for learning neural-network predictive control (in self-driving cars). *IEEE Access*, *8*, 192041–192064, Publisher: IEEE.

Da Lio, M., Mazzalai, A., Windridge, D., Thill, S., Svensson, H., Yüksel, M., et al. (2017). Exploiting dream-like simulation mechanisms to develop safer agents for automated driving: The Dreams4Cars EU research and innovation action. In *2017 IEEE 20th international conference on intelligent transportation systems (ITSC)* (pp. 1–6). IEEE.

da Silva, J. E., & de Sousa, J. B. (2010). A dynamic programming approach for the motion control of autonomous vehicles. In *49th IEEE conference on decision and control (CDC)* (pp. 6660–6665). IEEE.

da Silva, J. E., & de Sousa, J. B. (2011). A dynamic programming based path-following controller for autonomous vehicles. *Control and Intelligent Systems*, *39*, 245.

De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences*, *14*, 441–447.

Dey, K. C., Rayamajhi, A., Chowdhury, M., Bhavsar, P., & Martin, J. (2016). Vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication in a heterogeneous wireless network–performance evaluation. *Transportation Research Part C (Emerging Technologies)*, *68*, 168–184, Publisher: Elsevier.

Di Paolo, E., Rohde, M., & De Jaegher, H. (2010). Horizons for the enactive mind: Values, social interaction, and play. In *Enaction: Towards a new paradigm for cognitive science* (pp. 32–88). Cambridge MA: The MIT Press.

Dorffner, G. (1999). The connectionist route to embodiment and dynamicism. In *Understanding representation in the cognitive sciences* (pp. 23–32). Springer.

Dulac-Arnold, G., Mankowitz, D., & Hester, T. (2019). Challenges of real-world reinforcement learning. arXiv preprint arXiv:1904.12901.

Feldman, V. (2020). Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd annual ACM SIGACT symposium on theory of computing* (pp. 954–959).

Froese, T., & Di Paolo, E. A. (2008). Stability of coordination requires mutuality of interaction in a model of embodied agents. In *International conference on simulation of adaptive behavior* (pp. 52–61). Springer.

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, *32*, 1231–1237.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, *37*, 362–386.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, *27*, 377–396, Publisher: Cambridge University Press.

Guest, O., & Martin, A. E. (2020). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*, 789–802, Publisher: PsyArXiv.

Ha, D., & Schmidhuber, J. (2018). World models. arXiv preprint arXiv:1803.10122.

Hallam, B., Dario, F., Jean-Arcady, M., & Gillian, H. (2002). The road sign problem revisited: Handling delayed response tasks with neural robot controllers. In *From animals to animats 7: Proceedings of the seventh international conference on simulation of adaptive behavior* (pp. 228–229). MIT Press, 978-0-262-29117-0.

Harding, J., Powell, G., Yoon, R., Fikentscher, J., Doyle, C., Sade, D., et al. (2014). *Vehicle-to-vehicle communications: Readiness of V2V technology for application: Technical Report*, United States. National Highway Traffic Safety Administration.

Hars, A. (2016). Top misconceptions of autonomous cars and self-driving vehicles. *Inventivio Innovation Briefs*, 12.

Hashimoto, Y., Gu, Y., Hsu, L.-T., & Kamijo, S. (2015). Probability estimation for pedestrian crossing intention at signalized crosswalks. In *2015 IEEE international conference on vehicular electronics and safety (ICVES)* (pp. 114–119). IEEE.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences*, *6*, 242–247.

Hesslow, G. (2012). The current status of the simulation theory of cognition. *Brain Research*, *1428*, 71–79.

von Hofsten, C. (2004). An action perspective on motor development. *Trends in Cognitive Sciences*, *8*, 266–272.

Horák, K., & Bošanský, B. (2019). Solving partially observable stochastic games with public observations. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 33* (pp. 2029–2036). Number: 01.

Hurley, S. (2001). Perception and action: Alternative views. *Synthese*, *129*, 3–40, Publisher: Springer.

Hussain, R., Lee, J., & Zeadally, S. (2018). Autonomous cars: Social and economic implications. *IT Professional*, *20*, 70–77, Publisher: IEEE.

Jain, S., & Malhotra, D. I. (2020). A review on obstacle avoidance techniques for self-driving vehicle. *International Journal of Advanced Science and Technology*, *29*(06), 5159–5167.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*, 27.

Kendall, A., Hawke, J., Janz, D., Mazur, P., Reda, D., Allen, J.-M., et al. (2019). Learning to drive in a day. In *2019 international conference on robotics and automation (ICRA)* (pp. 8248–8254). IEEE.

Kim, J., & Canny, J. (2017). Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision* (pp. 2942–2950).

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*, 3521–3526, Publisher: National Acad Sciences.

Konda, V. R., & Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in neural information processing systems* (pp. 1008–1014).

Konrad, S. G., Shan, M., Masson, F. R., Worrall, S., & Nebot, E. (2018). Pedestrian dynamic and kinematic information obtained from vision sensors. In *2018 IEEE intelligent vehicles symposium (IV)* (pp. 1299–1305). IEEE.

Kotseruba, I., & Tsotsos, J. K. (2020). 40 Years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*, 17–94.

Kraus, S., Althoff, M., Heißing, B., & Buss, M. (2009). Cognition and emotion in autonomous cars. In *2009 IEEE intelligent vehicles symposium* (pp. 635–640). IEEE.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kuderer, M., Gulati, S., & Burgard, W. (2015). Learning driving styles for autonomous vehicles from demonstration. In *2015 IEEE international conference on robotics and automation (ICRA)* (pp. 2641–2646). [ISSN: 1050-4729].

Lee, J.-G., Kim, K. J., Lee, S., & Shin, D.-H. (2015). Can autonomous vehicles be safe and trustworthy? Effects of appearance and autonomy of unmanned driving systems. *International Journal of Human-Computer Interaction*, *31*, 682–691, Publisher: Taylor & Francis.

Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., et al. (2011). Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE intelligent vehicles symposium (IV)* (pp. 163–168). IEEE.

Li, Y., Xin, L., Yu, D., Dai, P., Wang, J., & Li, S. E. (2019). Pedestrian trajectory prediction with learning-based approaches: A comparative study. In *2019 IEEE intelligent vehicles symposium (IV)* (pp. 919–926). IEEE.

Lindblom, J. (2015). Embodiment and social interaction. In *Embodied social cognition* (pp. 115–159). Springer.

Lu, X., Tang, S., Zhang, L., Li, P., Li, C., & Wang, Y. (2019). A novel steering control for real autonomous vehicles via PI adaptive dynamic programming. In *2019 chinese control and decision conference (CCDC)* (pp. 926–930). [ISSN: 1948-9447].

Marina, L., & Sandu, A. (2017). Deep reinforcement learning for autonomous vehicles-state of the art. *Bulletin of the Transilvania University of Brasov. Engineering Sciences. Series I*, *10*, 195–202.

Mathibela, B., Newman, P., & Posner, I. (2015). Reading the road: road marking classification and interpretation. *IEEE Transactions on Intelligent Transportation Systems*, *16*, 2072–2081, Publisher: IEEE.

Maturana, H. R., & Varela, F. J. (1987). *The tree of knowledge: the biological roots of human understanding*. New Science Library/Shambhala Publications.

McCall, R., McGee, F., Meschtscherjakov, A., Louveton, N., & Engel, T. (2016). Towards a taxonomy of autonomous vehicle handover situations. In *Proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications* (pp. 193–200). ACM.

Miller, G., & Taubman-Ben-Ari, O. (2010). Driving styles among young novice drivers–The contribution of parental driving styles and personal characteristics. *Accident Analysis and Prevention*, *42*, 558–570, Publisher: Elsevier.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., et al. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529.

Möller, R. (1999). Perception through anticipation. a behaviour-based approach to visual perception. In *Understanding representation in the cognitive sciences* (pp. 169–176). Springer.

Montemerlo, M., Becker, J., Bhat, S., Dahlkamp, H., Dolgov, D., Ettinger, S., et al. (2008). Junior: The stanford entry in the urban challenge. *Journal of Field Robotics*, *25*, 569–597, Publisher: Wiley Online Library.

Mundhenk, M., Goldsmith, J., Lusena, C., & Allender, E. (2000). Complexity of finite-horizon Markov decision process problems. *Journal of the ACM*, *47*, 681–720, Publisher: ACM New York, NY, USA.

Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving.. *Psychological Review*, *65*, 151.

Nolfi, S., & Floreano, D. (2000). *Evolutionary robotics: The biology, intelligence, and technology of self-organizing machines*. MIT Press.

Nolfi, S., & Tani, J. (1999). Extracting regularities in space and time through a cascade of prediction networks: The case of a mobile robot navigating in a structured environment. *Connection Science*, *11*, 125–148.

Oka, T., Tashiro, J., & Takase, K. (1999). Object oriented benet programming for data-focused bottom-up design of autonomous agents, robotics and autonomous systems. *Elsevier*, *28*, 127–139.

Paden, B., Čáp, M., Yong, S. Z., Yershov, D., & Frazzoli, E. (2016). A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on Intelligent Vehicles*, *1*, 33–55.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, *113*, 54–71, Publisher: Elsevier.

Parisotto, E., Ba, J. L., & Salakhutdinov, R. (2015). Actor-mimic: Deep multitask and transfer reinforcement learning. arXiv preprint arXiv:1511.06342.

Park, Y.-J., & Tuzhilin, A. (2008). The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM conference on Recommender systems* (pp. 11–18).

Pendleton, S. D., Andersen, H., Du, X., Shen, X., Meghjani, M., Eng, Y. H., et al. (2017). Perception. *Planning, Control, and Coordination for Autonomous Vehicles, Machines*, *5*, 6, Publisher: Multidisciplinary Digital Publishing Institute.

Pfeifer, R., & Bongard, J. (2006). *How the body shapes the way we think: A new view of intelligence*. MIT Press.

Plebe, A., Papini, G. P. R., Donà, R., & Da Lio, M. (2019). Dreaming mechanism for training bio-inspired driving agents. In *International conference on intelligent human systems integration* (pp. 429–434). Springer.

Pomerleau, D. A. (1989). ALVINN: An autonomous land vehicle in a neural network. In *Advances in neural information processing systems* (pp. 305–313).

Poó, F. M., & Ledesma, R. D. (2013). A study on the relationship between personality and driving styles. *Traffic Injury Prevention*, *14*, 346–352, Publisher: Taylor & Francis.

Quintero, R., Parra, I., Lorenzo, J., Fernández-Llorca, D., & Sotelo, M. A. (2017). Pedestrian intention recognition by means of a hidden markov model and body language. In *2017 IEEE 20th international conference on intelligent transportation systems (ITSC)* (pp. 1–7). IEEE.

Raats, K., Fors, V., & Pink, S. (2020). Trusting autonomous vehicles: An interdisciplinary approach. *Transportation Research Interdisciplinary Perspectives*, *7*, Article 100201, Publisher: Elsevier.

Rao, Q., & Frtunikj, J. (2018). Deep learning for self-driving cars: chances and challenges. In *Proceedings of the 1st international workshop on software engineering for AI in autonomous systems* (pp. 35–38).

Rao, K., Harris, C., Irpan, A., Levine, S., Ibarz, J., & Khansari, M. (2020). RL-CycleGAN: Reinforcement learning aware simulation-to-real. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11157–11166).

Rasouli, A., & Tsotsos, J. K. (2020). Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE Transactions on Intelligent Transportation Systems*, *21*, 900–918, Publisher: IEEE.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., et al. (2016). Progressive neural networks. arXiv preprint arXiv:1606.04671.

Sagberg, F., Selpi, Bianchi Piccinini, G. F., & Engström, J. (2015). A review of research on driving styles and road safety. *Human Factors*, *57*, 1248–1275, Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Saleh, K., Hossny, M., & Nahavandi, S. (2019). Contextual recurrent predictive model for long-term intent prediction of vulnerable road users. *IEEE Transactions on Intelligent Transportation Systems*.

Sanders, P., & Schultes, D. (2007). Engineering fast route planning algorithms. In *International workshop on experimental and efficient algorithms* (pp. 23–36). Springer.

Santana, E., & Hotz, G. (2016). Learning a driving simulator. arXiv preprint arXiv: 1608.01230.

Scheier, C., Pfeifer, R., & Kunyioshi, Y. (1998). Embedded neural networks: exploiting constraints. *Neural Networks*, *11*, 1551–1569.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Shin, D. (2021a). Embodying algorithms, enactive artificial intelligence and the extended cognition: You can see as much as you know about algorithm. *Journal of Information Science*, Publisher: SAGE Publications Sage UK: London, England.

Shin, D. (2021b). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, Article 102551, Publisher: Elsevier.

Smith, L. B., & Slone, L. K. (2017). A developmental approach to machine learning? *Frontiers in Psychology*, *8*, 2124.

Spaan, M. T. (2012). Partially observable Markov decision processes. In *Reinforcement learning* (pp. 387–414). Springer.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Svensson, H., Thill, S., & Ziemke, T. (2013). Dreaming of electric sheep? Exploring the functions of dream-like mechanisms in the development of mental imagery simulations. *Adaptive Behavior*, *21*, 222–238.

Tani, J., & Nolfi, S. (1999). Learning to perceive the world as articulated: an approach for hierarchical learning in sensory-motor systems. *Neural Networks*, *12*, 1131–1141.

Thill, S., & Vernon, D. (2017). How to design emergent models of cognition for application-driven artificial agents. In *Neurocomputational models of cognitive development and processing: Proceedings of the 14th neural computation and psychology workshop* (pp. 115–129). World Scientific.

Thórisson, K. R. (2009). *From Constructionist to Constructivist AI Keynote, AAAI Fall Symposium Series–Biologically Inspired Cognitive Architectures, Washington DC: Technical Report, AAAI Tech Report FS-09-01*, (pp. 175–183). Menlo Park, CA: AAAI Press.

Thórisson, K., & Helgasson, H. (2012). Cognitive architectures and autonomy: A comparative review. *Journal of Artificial General Intelligence*, *3*, 1–30.

Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., et al. (2006). Stanley: The robot that won the DARPA grand challenge. *Journal of Field Robotics*, *23*, 661–692, Publisher: Wiley Online Library.

Urmson, C., Anhalt, J., Bagnell, D., Baker, C., Bittner, R., Clark, M. N., et al. (2008). Autonomous driving in urban environments: Boss and the urban challenge. *Journal of Field Robotics*, *25*, 425–466, Publisher: Wiley Online Library.

Vernon, D. (2014). *Artificial cognitive systems: A primer*. MIT Press.

Vernon, D., Metta, G., & Sandini, G. (2007). A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *IEEE Transactions on Evolutionary Computation*, *11*, 151–180, Publisher: IEEE.

Von Uexküll, J. (1992). A stroll through the worlds of animals and men: A picture book of invisible worlds. *Walter de Gruyter*, Publisher: Walter de Gruyter, Berlin/New York Berlin, New York.

Walch, M., Lange, K., Baumann, M., & Weber, M. (2015). Autonomous driving: investigating the feasibility of car-driver handover assistance. In *Proceedings of the 7th international conference on automotive user interfaces and interactive vehicular applications* (pp. 11–18). ACM.

Wali, S. B., Hannan, M. A., Hussain, A., & Samad, S. A. (2015). An automatic traffic sign detection and recognition system based on colour segmentation, shape matching, and svm. *Mathematical Problems in Engineering*, *2015*, Publisher: Hindawi.

Wang, C.-C., Thorpe, C., Thrun, S., Hebert, M., & Durrant-Whyte, H. (2007). Simultaneous localization, mapping and moving object tracking. *International Journal of Robotics Research*, *26*, 889–916, Publisher: Sage Publications Sage UK: London, England.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*, 279–292, Publisher: Springer.

Wen, Y., Zhang, X., Wang, F., & Han, J. (2015). Predicting driver lane change intent using HCRF. In *2015 IEEE international conference on vehicular electronics and safety (ICVES)* (pp. 64–68). IEEE.

Wolcott, R. W., & Eustice, R. M. (2015). Fast LIDAR localization using multiresolution Gaussian mixture maps. In *2015 IEEE international conference on robotics and automation (ICRA)* (pp. 2814–2821). IEEE.

Wu, Y., Tang, F., & Li, H. (2018). Image-based camera localization: an overview. *Visual Computing for Industry, Biomedicine, and Art*, *1*, 1–13, Publisher: SpringerOpen.

Xu, Y., John, V., Mita, S., Tehrani, H., Ishimaru, K., & Nishino, S. (2017). 3D point cloud map based vehicle localization using stereo camera. In *2017 IEEE intelligent vehicles symposium (IV)* (pp. 487–492). IEEE.

Yaqoob, I., Khan, L. U., Kazmi, S. A., Imran, M., Guizani, N., & Hong, C. S. (2019). Autonomous driving cars in smart cities: Recent advances, requirements, and challenges. *IEEE Network*, *34*, 174–181, Publisher: IEEE.

Yurtsever, E., Lambert, J., Carballo, A., & Takeda, K. (2019). A survey of autonomous driving: common practices and emerging technologies. arXiv preprint arXiv:1906.05113.

Zang, S., Ding, M., Smith, D., Tyler, P., Rakotoarivelo, T., & Kaafar, M. A. (2019). The impact of adverse weather conditions on autonomous vehicles: how rain, snow, fog, and hail affect the performance of a self-driving car. *IEEE Vehicular Technology Magazine*, *14*, 103–111, Publisher: IEEE.

Zhao, W., Queralta, J. P., & Westerlund, T. (2020). Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 737–744). IEEE.

Zhu, Q., Dai, B., Huang, Z., Sun, Z., & Liu, D. (2017). An adaptive longitudinal control method for autonomous follow driving based on neural dynamic programming and internal model structure. *International Journal of Advanced Robotic Systems*, *14*, Article 1729881417740711, Publisher: SAGE Publications.

Zhu, M., Wang, Y., Pu, Z., Hu, J., Wang, X., & Ke, R. (2020). Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving. *Transportation Research Part C (Emerging Technologies)*, *117*, Article 102662.

Ziemke, T. (2003). What's that thing called embodiment?, in: Proceedings of the annual meeting of the cognitive science society, volume 25. In *Proceedings of the annual meeting of the cognitive science society* (pp. 1305–1310). Issue: 25.

Ziemke, T., Bergfeldt, N., Buason, G., Susi, T., & Svensson, H. (2004). Evolving cognitive scaffolding and environment adaptation: a new research direction for evolutionary robotics. *Connection Science*, *16*, 339–350.