Minds & Machines (2008) 18:169–178
DOI 10.1007/s11023-008-9098-2

Robosemantics: How Stanley the Volkswagen Represents the World

Christopher Parisien · Paul Thagard

Received: 1 June 2007 / Accepted: 21 April 2008 / Published online: 7 May 2008

Abstract One of the most impressive feats in robotics was the 2005 victory by a driverless Volkswagen Touareg in the DARPA Grand Challenge. This paper discusses what can be learned about the nature of representation from the car's successful attempt to navigate the world. We review the hardware and software that it uses to interact with its environment, and describe how these techniques enable it to represent the world. We discuss robosemantics, the meaning of computational structures in robots. We argue that the car constitutes a refutation of semantic arguments against the possibility of strong artificial intelligence.

Keywords Robotics · Representation · Semantics · Intentionality · Bayesian networks

Introduction

In 2005, a driverless Volkswagen Touareg sports utility vehicle won the DARPA Grand Challenge, a race for autonomous robots over a difficult 131-mile course in the Mojave Desert (DARPA 2005). The winner was developed by a team from the Stanford University Artificial Intelligence Laboratory, who called their vehicle *Stanley*. The magazine *Wired* (2006) declared Stanley the best robot of all time, and the self-navigated completion of the difficult course was certainly one of the major

C. Parisien
Department of Computer Science, University of Toronto, 10 King's College Road, Toronto, ON, Canada M5S 3G4
e-mail: chris@cs.toronto.edu

P. Thagard (✉)
Department of Philosophy, Faculty of Arts, University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada N2L 3G1
e-mail: pthagard@uwaterloo.ca

robotic feats to date. Stanley's success is the result of great sophistication in both hardware and software, including multiple instruments for sensing its environment and advanced programs for making inferences about its location and direction.

We aim to examine what can be learned about the nature of representation from Stanley's successful attempt to navigate the world. After a brief review of the hardware that Stanley used to interact with its environment, we discuss the software that enabled it to identify relevant features of the world and to plan an effective course using dynamic Bayesian networks and machine learning algorithms. We then describe how these techniques enabled Stanley to represent the world, and discuss what they tell us about *robosemantics*, the meaning of computational structures in robots. We also show that Stanley constitutes a refutation of semantic arguments against the possibility of strong artificial intelligence.

Stanley's Hardware and Software

Stanley was a diesel-powered Volkswagen Touareg R5 whose throttle, brakes, and steering were electronically controlled (Thrun et al. 2006). It perceived the world by means of sensors mounted on a roof rack which held five laser range finders, a color camera, and two antennas of a RADAR system, all pointed forward. Other antennas are used for GPS (global positioning system) and DARPA's system for stopping vehicles in an emergency. Stanley's trunk contained six Pentium M computers that were connected to each other, to the physical sensors, and to the actuators for throttle, brakes, and steering. The computers served to integrate all the information from the various sensors, determine Stanley's location, infer what obstacles lie ahead, and instruct the vehicle to drive at a manageable speed in the appropriate direction.

The director of the Stanford Artificial Intelligence Laboratory is Sebastian Thrun, co-author of an elegant recent textbook on probabilistic robotics that describes many of the computational techniques used in Stanley (Thrun et al. 2005; for a more elementary introduction, see Russell and Norvig (2003, Chap. 25), which was mostly written by Thrun). Robots need statistical techniques to deal with uncertainty because their environments are unpredictable, their sensors are limited, their actuators such as motors are not completely reliable, the internal models developed by their software are only approximate, and their computations are limited by time constraints.

In the Stanford approach to robotics, the primary technique for dealing with uncertainty is based on the Bayes filter, a method of representing beliefs over time as probability distributions over possible states. The state of a robot can be captured by a collection of variables that stand for its location, orientation, velocity, configuration of actuators, and features of objects in its environment. For a rigid mobile robot, location and orientation can be represented by six numerical variables: three for Cartesian coordinates relative to a global frame, and three for pitch, roll, and yaw. Other variables stand for measurements performed by sensors and for control actions carried out by the robot.

While the details of Bayes filtering (and its usual implementation, Kalman filtering) are beyond the scope of this paper, we can describe some of the

representational underpinnings using a simple model, the dynamic Bayes network. Essentially, a Bayes network is a graph in which each node is a variable that can take on a range of values (such as a robot's state). The relations between nodes are the conditional probabilities that connect the variables, and the function of a *dynamic* Bayes network is to use available evidence to update these probabilities over time. Figure 1 shows how a collection of measurement variables, u, influences a collection of state variables, x, which influences a collection of control variables, z. We use Bayes' theorem to estimate the state at time t, which in turn influences the state at time $t + 1$. In practice, Stanley was implemented using an *unscented Kalman filter* (UKF), an efficient method of probability estimation based on similar principles to those described here. In sum, a Stanford probabilistic robot is a machine that uses Bayes' theorem to repeatedly make inferences about its current state.

Stanley had three different sensing modalities: laser, vision and RADAR. The laser system had a range of approximately 25 m, which is only adequate for low speeds. In contrast, the vision and RADAR systems were good for a range of up to 200 m, but provided much coarser information than the laser measurements. Measurements from these sources were used to detect obstacles by functions determined by a machine learning algorithm that used human driving for training. The vision processing module was a Bayes network that used online machine learning to adapt continually to different terrain types. Data from all sensors were integrated into a *drivability map*, which is a single model of the environment that marks each cell in a two-dimensional map as either unknown, drivable, or undrivable. This information, along with other variables for the general condition of the environment such as terrain slope, are used to set the driving direction and velocity of the vehicle, which in turn control the steering, throttle, and brake. With six fast computers, Stanley was able to update its localization up to 100 times/s, update its visual discrimination of road from obstacles 8–75 times/s, and generate steering and velocity controls 20 times/s. Much more detail about how Stanley moved from measurements to actions is available elsewhere (Thrun et al. 2005, 2006). At the conclusion of this paper, we will briefly compare Stanley with the winner of the 2007 DARPA Urban Challenge, a Chevy Tahoe from Carnegie Mellon University.

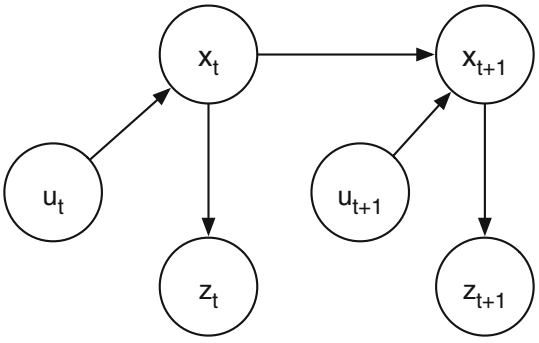

Fig. 1 Dynamic Bayes network that characterizes the evolution of measurements u, states x, and controls z. Based on Thrun et al. (2005, p. 25)

Meaning

We take Stanley's success at autonomously navigating a complex environment to be prima facie evidence that its representations are meaningful, but we will not attempt to review the many competing philosophical theories of meaning (see e.g. Cummins 1989). Instead, we will adapt the "neurosemantics" theory of Eliasmith (2005), who proposes a representational framework based on the abilities of neurons. This semantic theory is connected to a rich neurocomputational account of how brains encode, decode, and transform information (Eliasmith and Anderson 2003; Eliasmith 2003). Eliasmith (2005, p. 1044) presents a four-place schema for defining a representation:

> A {vehicle} represents a {content} regarding a {referent} with respect to a {system}.

He proceeds to explain how neural systems fit each of these components in order to give us mental meaning, and we will perform a similar analysis for Stanley.

Systems

For human mental representation, the system is typically considered to be the person, that is, the natural biological system of the human being. For robots, we are concerned with how representations arise, how they are stored, and how they are used, so we consider the entire information processing system of the robot. For Stanley, these included sensors, processing units, data storage, and the actuators that control the SUV's movements.

Vehicles

In Eliasmith's account, vehicles are internal states of physical objects that carry representational contents (in what follows, we shall use "vehicle" in this sense rather than the automotive one). It is important to distinguish vehicles from contents: in the human brain, the vehicles are neurons and groups of neurons, and contents are the properties that they ascribe to the world. Because the job of the probabilistic robot is to ascribe properties to the world by computing values for variables, its vehicles are the states of computer chips that store conditional dependence relationships, values for variables, and perform Bayesian updating.

Referents

Referents are the entities in the world that representations are about. In human mental states, referents are actual dogs, buildings, and so on—the targets of thinking. Since robots operate in the same world as people do, it is desirable that they share the same possible referents. For a robot like Stanley, some important referents are the landscape, obstacles, other vehicles, and the path through the landscape. According to Eliasmith (2005, p. 1046), the referent of a vehicle is the set of causes that has the highest statistical dependence under all stimulus

conditions. Stanley was causally connected to the world by its three kinds of visual sensors (laser, camera, RADAR) and by the GPS and inertial system used for localization. Thus the referents for its localization variables are Stanley itself and its place in the world, and the referents for variables representing features of the environment are the objects in the world that cause laser, light, and RADAR beams to be reflected back to the sensors.

Contents

In Eliasmith's theory of neurosemantics, the content of a representation is the set of properties of the referent encoded by the vehicle. It is obviously not possible for robots or humans to represent every possible aspect of something in the world. Drawing information from the world requires filtering and encoding performed by the vehicle. For a robot, filtering begins with the sensors. A radar unit, for example, will extract the distance to a solid obstacle at various angles around the robot. A camera, on the other hand, will sense the visible light reflected off of the surroundings. Both sensors observe the same referent, the landscape, but extract different properties of it.

In a Kalman filter, content is captured by the values of different variables. A robot will construct maps of its terrain and action plans about where to go and how to get there. Each of these contents has a probabilistic component, and is merely a subset of the possible properties of the outside world.

In sum, the semantic capability of a Stanford probabilistic robot fits comfortably into Eliasmith's four-place schema:

> A {Kalman filter running on computer chips} represents {statistical properties} regarding an {environment feature} with respect to a {robot's information-processing system}.

The neurosemantic view defines how each component of representation can be satisfied by a neural system. For each of these components necessary for representation, an analogue exists in Stanley.

Misrepresentation

Sophisticated representational systems are capable of making mistakes (Dretske 1995). My own visual faculties are generally good enough that when I think I am looking at a dog, it really is a dog. As proud as I am of this fact, this miracle of perception occasionally breaks down. Late at night in a fog-shrouded park, raccoons look like dogs, shadows look like people, and I am frequently confused.

Stanley had similar problems. Perception is a notoriously difficult problem to solve in robotics, and Stanley's laser and vision systems were by no means perfect. One prominent example comes from the way Stanley uses its laser rangefinders to judge the terrain directly in front of the car. A rotating laser sweeps the ground in an arc several meters ahead, and the rangefinder computes depth information along that arc. As the car moves forward, it pushes the arc like a broom, combining the information from multiple sweeps to create a three-dimensional map. However, this

process depends on the car's stability, because when the car pitches forward over a bump, the laser rescans a previous area, and then skips far ahead. This puts the scan lines out of sequence, making the rangefinder perceive a large, impassable obstacle (Thrun et al. 2006). Consequently, Stanley would carry out often dangerous avoidance maneuvers for an obstacle that never existed. The problem was eventually solved using better machine learning techniques, but the important lesson remains: since the perceptual system can ascribe the wrong properties to whatever lies in front, Stanley is capable of misrepresentation.

Beyond the Chinese Room

Stanley's ability to represent and misrepresent the world provides a decisive counterexample to John Searle's notorious argument that digital computers are inherently incapable of intelligence. Here is his most recent version (Searle 2004, p. 90; see also Searle 1980, 1992):

> Imagine that I am locked in a room with boxes full of Chinese symbols, and I have a rule book, in effect, a computer program, that enables me to answer questions put to me in Chinese. I receive symbols that, unknown to me, are questions; I look up in the rule book what I am supposed to do; I pick up symbols from the boxes, manipulate them according to the rules of the program, and hand out the required symbols, which are interpreted as answers. We can suppose that I pass the Turing test for understanding Chinese, but, all the same, I do not understand a word of Chinese. And if I do not understand a word of Chinese on the basis of implementing the right computer program, then neither does any other computer just on the basis of implementing the program, because no computer has anything that I do not have.

Searle thinks that a computer operates purely syntactically with uninterpreted symbols, whereas the human mind attaches meaning to the symbols.

As Holyoak and Thagard (1995) noted, Searle's thought experiment is an argument from analogy: just as Searle in the Chinese room does not understand Chinese, so computers are incapable of understanding anything. Analogical arguments have force only if they point out similarities between the source and target analogs that are relevant to the conclusion. Figure 2 makes clear the structure of the analogs in Searle's thought experiment, which works fairly well for the type of computer that people today have sitting on their desks. People type into their computers and get output back from their screen, and even if the computer is running a sophisticated artificial intelligence program it is legitimate to say that the computers, like Searle in the Chinese room, do not on their own attach meaning to their symbols.

However, Stanley was a very different machine from personal computers. Figure 3 shows how far Stanley goes beyond the analogs in Searle's thought experiment. Stanley as a system is different from the Chinese room and the personal computer because its sensors and control actions give it ongoing causal interactions with the world, many times/s. Stanley's computer chips are vehicles for representing

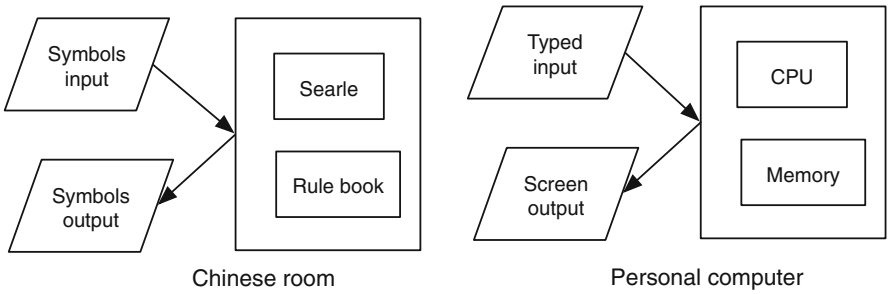

Fig. 2 Searle's analogy between the Chinese room and computers. CPU is the central processing unit of a digital computer

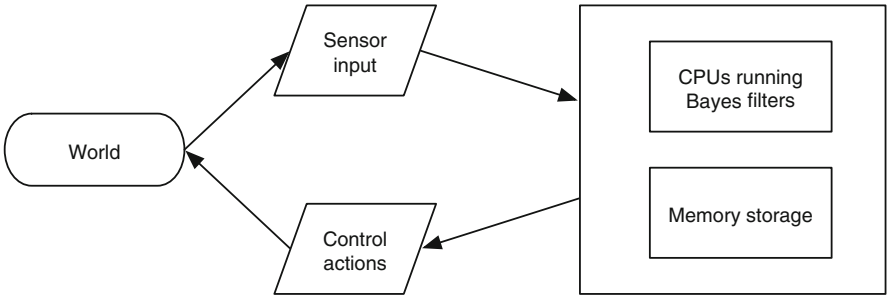

Fig. 3 Stanley's systematic relations to the world

the world in the same way that human neurons are, because of the causal, statistical dependencies between their operations and what goes on in the world.

Searle's reply to the claim that robots show the possibility of a computer having meaningful symbols is a modified version of his thought experiment. Suppose that the Chinese room is placed inside a robot so that the input comes from a television camera and the output contains motor instructions. There may even be statistical causal dependencies between the person's manipulation of symbols and the television input and motor output that provide connections to the world. Searle maintains that nothing has changed, in that the person in the room is still processing meaningless symbols, so by analogy the robot's CPU is also.

However, Searle's analogy is defective. Stanley's probabilistic variables may lack meaning when considered only in relation to its six CPU's, but are meaningful when considered with respect to the robot's whole information processing system, including its sensors that generate statistical properties regarding features of the environment. To see how this works, consider the Chinese room to be an analogy for something we already know to have meaningful content: the human neurobiological system. If we use the robot version of the Chinese room, then the correspondences are quite straightforward. The robot's input corresponds to the human sensory-perceptual system, including visual and auditory areas of the cortex, and so on. The robot's output matches up with cortical motor areas, the cerebellum,

and the musculoskeletal nerves. The rule book is simply memory, instantiated in the hippocampus and neocortex. Now where is the man in the room? In the thought experiment, the man is a control center, carrying out rules for manipulating symbols. To correspond, we might choose the dorsolateral prefrontal cortex, a major site for executive functions including working memory. This area takes neural spikes as input, performs some transformations, and passes neural spikes as output. All that the region sees are spikes, which are meaningless for it alone. Thus by Searle's argument, it might seem that brains cannot have meaningful symbols, contrary to his own assumptions. But brains, of course, rely not just on a single region for central processing, but many regions including ones dedicated to processing sensory information. Similarly, Stanley uses multiple CPUs interacting with each other and multiple sensors.

Thus Stanley's abilities undermine Searle's argument. The robot's Bayesian networks give it representational power. Sensory inputs give Stanley's representations statistical causal dependencies with the world, assigning the representations content with respect to the system. Furthermore, Stanley's performance in the real world is evidence that the content works. Because Searle's analogy with the Chinese room is defective, and because Stanley's performance in the world is so successful, we have reason to attribute meaning to Stanley's symbols, the variables and links in its Bayes networks. Their meaning does not derive simply from the programmers who wrote the C code for Stanley's computers, but also from ongoing interactions with the world and with ongoing machine learning that make possible Stanley's effective operations.

Shani (2005) has attempted to supplement Searle's thought-experimental argument against robot intentionality with another argument derived from the work of Mark Bickhard (e.g. Bickhard and Terveen 1995). Shani contends (2005, p. 220) that "mental structures cannot function as representations, cannot be intrinsically informative, *in virtue* of the fact that they *encode* whatever it is they encode". But we argued above that Stanley's representations had referents in just the same way that human brains do, by having high statistical dependence under all stimulus conditions. If such causal relations are sufficient for neurosemantics, they should also be sufficient for robosemantics.

Conclusion

Despite its impressive navigational accomplishments, Stanley fell far short of human-level intelligence. It has no capability of processing natural language, and no one would claim it has consciousness. Its problem solving ability is less than a cockroach, which can not only navigate a complex environment but also find food, mate, and avoid being stomped. Nevertheless, there was dramatic progress between the 2004 DARPA Grand Challenge, when none of the robotic vehicles managed to complete the course, and the 2005 challenge, when four other vehicles completed the course after Stanley, using a variety of kinds of hardware and software (see the technical reports available at DARPA 2005). These impressive achievements were repeated 2 years later in the 2007 DARPA Grand Challenge, which required

autonomous vehicles to complete a 60-mile *urban* course safely, obeying traffic laws, in less than 6 h. Six vehicles completed the difficult course, led by Carnegie Mellon University's Tartan Racing Team. Stanford placed second.

Carnegie Mellon's 2007 winner was a Chevy Tahoe called Boss. Like most of the 2007 competitors, it used a powerful new laser sensing technology produced by Velodyne, consisting of a spinning unit with 64 lasers firing thousands of times/s. To interpret sensory information, it used similar kinds of probabilistic filtering techniques employed by Stanley and many earlier robots. Boss and most of its competitors translated sensory information into discrete rules for guiding action. Such translation was necessary because urban traffic is governed by precise rules of conduct, unlike the less constrained desert navigation in the 2005 competition. Boss had more computing power than Stanley, with 10 computers containing 20 CPUs. It took months of testing to get Boss ready for the Urban Challenge, including considerable tuning of software by Boss's programmers but also use of machine learning algorithms to improve its interpretation of sensory inputs (Paul Rybsky, personal communication, Feb. 22, 2008).

Thus Stanley's Bayes network software is not the only way of building robots, and it is debatable whether the human brain is similarly Bayesian. Obviously, the brain is a sophisticated processor of statistical information, but that does not imply that it uses the particular machinery of Bayes networks: directed acyclic graphs obeying crucial conditions about probabilistic dependence. Some psychologists think that the human mind incorporates Bayes networks (Gopnik et al. 2004), but others are skeptical (Rehder and Burnett 2005). One of the advantages of looking at robots is that we know how Stanley was built to learn about and to interact with the world, and probabilistic reasoning is a central part of its design. We have described how they enabled Stanley to represent the world in ways that derive from its own sensing, acting, inference and learning capabilities, not just those of its initial programmers.

Philosophers use the term *intentionality* for the human capability of having internal representations that are about the world. If you are wondering whether robots can have intentionality, the reasonable answer is: they already do.

Acknowledgments We are grateful to Chris Eliasmith and Abninder Litt for comments on an earlier version, and to the Natural Sciences and Engineering Research Council of Canada for funding. Thanks to Sebastian Thrun for information about Stanley, and to Paul Rybsky and Drew Bagnell for information about the 2007 Urban Challenge winner from Carnegie Mellon.

References

Bickhard, M. H., & Terveen, L. (1995). *Foundational issues in artificial intelligence and cognitive science: Impasse and solution*. Amsterdam: Elsevier.

Cummins, R. (1989). *Meaning and mental representation*. Cambridge, MA: MIT Press.

DARPA. (2005). *Web archive of the October 2005 Grand Challenge*. Retrieved June 16, 2006, from http://www.darpa.mil/grandchallenge05/index.html.

Dretske, F. I. (1995). *Naturalizing the mind*. Cambridge, MA: MIT Press.

Eliasmith, C. (2003). Moving beyond metaphors: Understanding the mind for what it is. *Journal of Philosophy, 100*, 493–520.

Eliasmith, C. (2005). Neurosemantics and categories. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 1035–1054). Amsterdam: Elsevier.

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation and dynamics in neruobiological systems*. Cambridge, MA: MIT Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schultz, L. E., Kushur, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 2004*, 3–32.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press/Bradford Books.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50*, 264–314.

Russell, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*, 417–424.

Searle, J. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.

Searle, J. (2004). *Mind: A brief introduction*. Oxford: Oxford University Press.

Shani, I. (2005). Computation and intentionality: A recipe for epistemic impasse. *Minds and Machines, 15*, 207–228.

Thrun, S., Burgard, W., & Fox, D. (2005). *Probabilistic robotics*. Cambridge, MA: MIT Press.

Thrun, S. et al. (2006). Stanley: The robot that won the DARPA Grand Challenge. *Journal of Field Robotics, 23*, 661–692.

Wired. (2006). *The 50 best robots ever*. Retrieved June 16, 2006, from http://www.wired.com/wired/archive/14.01/robots.html.