How Smart Machines Think Debate

Consciousness in Self-Driving Cars

Debaters (in alphabetical order):

Emily Stark & Michael Teti

Resolution:

Consciousness is one of many ill-defined concepts which sparks controversy within the fields of artificial intelligence and psychology, and across them. Below are several definitions of consciousness:

> Sentience or awareness of internal or external existence
> ~ Merriam-Webster Dictionary
> The state of being awake and aware of one's surroundings
> ~ Oxford
> One capable of sensing and responding to its world
> ~ Armstrong (1981)
> An organism's awareness of something either internal or external to itself
> ~American Psychological Association

These definitions can roughly be distilled into the following, working definition of consciousness:

(1) Being aware of one's self and/or
(2) Being aware of one's surrounding

There is rich literature in the field of psychology surrounding conscious versus unconscious processing. It seems agreed-updo by psychologists that just because something is physically recorded by the nervous system, that does not mean it is consciously perceived (e.g. Ress & Heeger, 2003). This further reinforces the *awareness* piece of the definitions above which distinguishes sensing from conscious processing.

This brief overview of consciousness brings us to the present resolution:

> The architecture which drove Boss in the DARPA Urban
> Self-Driving Challenge contained modules which qualify Boss of
> having consciousness.

Debate Structure:

The following debate has developed *only* through written word. The two parties have agreed to not discuss the content of either sides in any format other than in the following pages in order to encapsulate as much of the thought-process as possible.

Each party has decided ahead of time on the debate resolution and have confirmed opposing viewpoints. Other than the preceding resolution, neither have discussed their thoughts on the subject.

Both parties have agreed to honor the highest standards of academic debate wherein they strive to produce reputable sources and/or support for all claims made. These references do not need to be produced before submitting their rebuttal, however the full citations must be supplied along with the rebuttal to allow the other party the chance to review the material.

The debate will be structured in the following manner:

- Opening Statements
  - These are done concurrently and with no input from the other party
- Rebuttals: Con-Pro
  - These are done sequentially
- Closing Statements
  - These are prepared concurrently and with no input from the other party

Opening and Closing Statements may be no longer than one full page, typed, single-spaced, 11pt font. Rebuttals may be no longer than one-half page, typed, single-spaced, 11pt font. These limitations exclude references.

In order to promote a timely debate, Opening Statements will be prepared within a 24 hour period of the start time. The first rebuttal must be submitted no later than 48 hours after the Opening Statements are posted. The following rebuttal must all be submitted no later than 48 hours following the previous rebuttal. Closing Statements must be posted no later than 24 hours after the final rebuttal. The Closing Statements will be posted at the same time.

Opening Statement 1: CON RESOLUTION

In this opening statement, I will argue that Boss's architecture, while impressive, does not contain consciousness in any module or combination of modules according to the agreed-upon definitions of consciousness. I will argue that these definitions of consciousness above depend entirely on the word "awareness", which I first show is regarded as distinct from perception, behavior, and even knowledge. Second, I will discuss certain cognitive characteristics of awareness in humans and show how Boss's architecture precludes it from displaying these characteristics.

It is well accepted in various fields of psychology, neuroscience, and philosophy that perception and behavior can happen entirely unconsciously - or outside a person's awareness - in most cases (Kanwisher 2001). An example of this is the phenomenon of binocular rivalry, in which separate images are displayed to each eye simultaneously, and the human subjects report seeing only one of the images at a time despite the unseen image activating the photoreceptors and upstream visual areas. Even emotions, which can serve as cues to certain behaviors in different situations and environments, are often processed unconsciously and arise automatically (Kandel et al. 2012). Indeed, any person who has ever driven a motor vehicle will recall that it is not uncommon to day-dream on long drives only to come to the frightening realization that they can't remember whole parts of the drive. This simple, yet effective, example helps illustrate that awareness is not only separate from perception and action, but it is also not required to perform complex behaviors involving many different cognitive abilities, such as driving a vehicle safely and effectively.

It would be extremely difficult to prove that Boss was - or was not - aware of the perception and planning it was performing during the task since it is not able to communicate its awareness like subjects in an experiment would. One possible way to determine whether Boss was aware of these operations would be to compare its architecture and abilities to observations and phenomena from human consciousness studies. For example, awareness of a thought or state allows the brain to recruit different cognitive abilities, which allows humans to better adapt and respond to various situations (Kandel et al. 2012). In humans, the thalamus, which is a relatively primitive structure in the mammalian forebrain, is largely responsible for routing information to various cortical areas depending on the nature of the information and top-down feedback from various cortical areas (Schmitt et al. 2017; Nakajima et al. 2019). When we look at the architecture of Boss, there is no evidence of top-down feedback or information routing, two characteristics which would likely be present in a system that is said to be aware. The entire network is feed-forward (with the exception of the planning module controlling the actuators), and the information is sent through it the same every time. The sensors will take in the same information and the object detectors will keep the same filters and output the same outputs to the planning module no matter what state the planning module is in.

In conclusion, I have argued that perception and awareness are independent entities. I have also illustrated one how Boss's architecture gives no evidence of displaying a property one would expect to see in an agent claimed to contain awareness of its actions. Therefore, it is possible to conclude that, although Boss is modeling and keeping track of certain things in the world, it is not necessarily aware of those things, and it cannot be said to have consciousness.

Opening Statement 2: PRO RESOLUTION

Psychologists are shy to define terms if it means the terms would be used to compare computational models with the human mind. However, they are much more talkative when defining terms to study in their labs. From the definitions above, it seems clear that an agent (biological or computational) must be able to sense the environment and/or themselves. If we do not expand on this definition further, one could conclude that any robot affixed with sensors might be considered “conscious”. However, this answer is unsatisfactory and such a naive approach is undoubtedly unhelpful in understanding how autonomous cars work and possibly draw inspiration from human consciousness. For this reason, my stance is that Boss, the autonomous car which competed in the DARPA Urban Challenge, did practice conscious processing.

It seems “unconsciousness” is more readily defined by experimental psychologists. In a popular study from 2004, Dijksterhuis defined unconscious thought as “cognitive and/or affective task-relevant processes that take place outside conscious awareness.” He also defined conscious thought similarly as taking place within awareness. The sticking point of awareness is why I believe the argument of sensing implies consciousness is insufficient.

There are a few key findings from psychology which support a cognitive processing pipeline of sensing => processing => conscious perception (awareness). From a simplistic viewpoint, there is the work in binocular rivalry, specifically using continuous flash suppression. It has been shown that when people are exposed to a stimulus they are unable to consciously perceive, there is evidence that they still processed the stimulus (Tsuchiya & Koch, 2005). From a more complex, viewpoint, it has been shown that when an agent, i.e. a human, is presented with information, but prevented from consciously thinking about it, they will use the information to make a superior decision than if they were to make the decision immediately (Dijksterhuis, 2004). Both of these findings indicate that unconscious thought takes in much more information than consciousness. In this way, unconsciousness acts as a “preprocessing” step which synthesizes the signals from the world and passes along relevant information which allows the agent to consciously make informed decisions.

Essentially, my argument can be laid out as follows:

- Humans sense and process information about external and internal states.
- Humans then preprocess that data and pass along information deemed relevant to the level of consciousness.
- Boss’s sensors process information about external and internal states.
- Early modules preprocess the information and post it to a “message board” where it may or may not be used.
- Some modules then use these messages to make decisions about what to do next. These are the consciously thinking modules.

First Rebuttal: CON

While my opponent and I agree on some points, such as the fact that sensation - and often perception - takes place outside of consciousness, we disagree on many other points. In his/her argument, my opponent attempts to draw a parallel between the cognitive processing pipeline observed in experimental psychology studies, which is sensation to processing to conscious awareness, and the sensing to perception to planning stages of Boss. I submit that this argument is superficial at best and dangerously inaccurate at worst. Upon viewing Boss's architecture, there does seem to be much similarity between Boss's processing pipeline and that observed in human cognition (which Gerrish doesn't help much by referring to Boss's architecture as its brain). However, if we look deeper at the actual architecture and the functions being performed and contrast this to functions and connectivity of the human brain from neurophysiological studies of consciousness, we will see that the similarities end at the superficial. My opponent does try to delve deeper into his/her examination of certain stages of the Boss architecture in an attempt to more thoroughly connect these modules to parts of human perception. However, I assert that these comparisons are inadequate for the question being asked. They spent much time comparing the perception stage of Boss to the preprocessing stage in the cognitive perception pipeline. They conclude from this that the planning stage of Boss must be the conscious module because the perception module - which is to Boss what the preprocessing stage is to human perception - feeds into the planning module after filtering the information. This makes sense under the three stages of human perception, but we still can't say at all whether Boss was aware or conscious of this information. Indeed, much of the goal-oriented planning and behavior humans perform has been observed to be outside of conscious awareness (Binsted et al. 2007), so it is entirely possible that Boss's planning modules are as well without any evidence to the contrary.

Second Rebuttal: PRO

My opponent has made several claims in his/her Opening Statement and the First Rebuttal which are not supported by Boss's architecture or scientific literature. The first assertion is that within Boss's architecture there is no evidence of top-down processing. However, this is inconsistent with the report on Boss in the AAAI magazine (Urmson et al., 2009). Specifically, Figure 3 in the report indicates that the Behavioral Executive module both receives information from and sends information to the Mission Planning and Motion Planning modules. More informative is Figure 4, which describes the object detection capabilities. This figure shows that the sensors send information after some amount of processing and feature extraction to a Fusion Layer which completes a number of tasks, but this Fusion Layer sends information back down to the sensor layer which validates the features.

The second claim, which attempts to undermine the comparison between Boss's architecture and human consciousness, asserts that because sensors will report information in the same manner (i.e. using the same filters) regardless of context, Boss is behaving without consciousness. However, this is simply refuted by the notion that the processing of sensor information is remarkably similar to the human visual system. The retina takes in the same information regardless of context and that information is processed in V1 in the same way regardless of situation. It is true consciousness is not found in V1, but it is an important step.

The last, possibly most egregious, claim made is that goal-oriented planning and movement can be achieved without conscious processing. However in the study my opponent cited, the task was an incredibly short-term perception task (Binsted et al. 2007). The results showed that participants could point to a visual stimulus they were not consciously aware of, however to assert that they could drive a car using only unconscious processing or awareness remains unjustified. My opponent touched on this notion of unconsciousness planning in the application of driving claiming that during an average commute, drivers often experience the phenomenon of driving without conscious awareness. That experience is well-documented and uncontested. However, “highway hypnosis” does not occur in busy city traffic, much less when the agent driving does not have familiarity with the path or is trying to win a race which involves extensive and longer-term planning. Neither of which are addressed in the cited articles or my opponent's rebuttal.

Closing Statement: CON

Here, I will make my closing statement in an effort to solidify my assertion that the Boss architecture does not, and cannot contain consciousness. I will begin by discussing the arguments made in my opponent's rebuttal of this stance and illustrate why they are misguided. Specifically, I will discuss the fact that, while the depiction of Boss's architecture in the report by Urmson et al. (2009) indicates that Boss's architecture does contain top-down processing, the functionality of these connections bears little resemblance to that observed in neurophysiological studies. I will then attempt to disprove my opponent's statement that the information processing pipeline used by the Boss architecture is “remarkably similar” to that of the human visual system by asserting that this is largely due to the high level of abstraction used to illustrate it. I will then conclude this statement by reiterating my central argument in light of these points.

My opponent has argued against my assertion that Boss's architecture does not contain top-down processing -- something that you would expect to find in a conscious agent -- by introducing a figure in the report by Urmson et al. (2009). Specifically, my opponent referred to Figure 4 in the report, where there is an arrow pointing from the Fusion Layer to the Sensor Layer that is labeled Validated Features. By merely observing this figure, it does seem likely that there is some top-down processing incorporated into the Boss architecture. However, in their description of this connection's function, Urmson et al. state that it is used to validate sensor measurements by comparing them with the instantaneous obstacle map, and these measurements are then discarded if there is a discrepancy between the two. First, assuming the sensor measurements are analogous to the outputs of retinal ganglion cells (since there seems to be no issue with mapping Boss's architecture onto that of the human brain), my opponent's argument is incorrect because top-down visual processing does not happen in the retina; it interacts with the outputs coming from the retina (Gilbert and Li, 2013). Furthermore, it has been observed that, rather than discarding inputs, top-down connections to a given neuron causes the neuron's entire function -- or “algorithm” -- to change. This alone highlights why discarding measurements at the sensor layer is not something observed in the human visual system, and, thus, why my opponent's argument is not entirely accurate.

Second, my opponent asserts that Boss's visual processing pipeline is similar to that of humans, and, therefore, there is evidence for consciousness in the former. In support of this argument, they state that in the human primary visual cortex, neurons process information the same way “regardless of context”, which is simply not true (Gilbert and Li, 2013). In this assertion we can see the dangers of anthropomorphizing algorithms and machines (not that it can't be useful sometimes), as Gerrish and many others do when they talk about Boss's brain. Typically, when these claims are made, the algorithm or architecture is being examined at a very high level of abstraction which can make it very easy to map these things onto the human brain. However, when you zoom in on the mechanisms and operations of the architecture, it is often true that these things look almost nothing like the brain at all.

Here, I conclude this closing statement by reiterating my key arguments as to why Boss does not contain consciousness: 1) the top-down feedback in the Boss architecture does not look like anything in the human brain functionally, and 2) anything can look like anything else if you zoom out far enough. Thank you for reading.

Closing Statement: PRO

The common thread between my opponent’s and my reasoning is that definitions of consciousness are often lacking and insufficient. That said, we agreed on the point that sensing does not necessitate conscious awareness. We ultimately disagreed on what beyond sensing would necessitate consciousness and what that looks like in humans, as well as algorithms. My opponent also pointed out that the common language used in discussing Boss and other AI-agents performing human-like tasks (speech, driving, etc.) are typically referred to as “them” or make reference to their “brain”. We anthropomorphize the algorithms to the point of giving them names, but refuse to believe that they might have what would be considered consciousness if it occurred in a biological agent.

Operationalizing the notion of consciousness gives us a common ground to build arguments for or against Boss containing consciousness. However by assigning such a strict definition to such an abstract concept, we lose a lot of nuance, as is always the risk of dimensionality reduction. In this way, I believe it is clear that with the limited, agreed-upon definition of consciousness that Boss’s architecture contains what could be considered consciousness.

I have previously cited studies from psychology and perception to support my conclusion. Consciousness is by and large the next step in the perceptual pipeline which follows unconscious processing. Our conscious awareness does not have the bandwidth to acknowledge every stimulus we perceive. Our unconscious cognition preprocesses the data and makes assumptions, fills in gaps, and triage’s the information brought in to most effectively use our limited conscious resources. This is a direct parallel to the architecture that drives Boss. All of these preprocessing filters can be thought of as the different sensing systems Boss contains that post message for the planning module (wherein consciousness would reside) to use as it needs.

The parallels I have drawn between human consciousness and Boss’s architecture have been refuted by my opponent, but not effectively. It was claimed that there does not exist “top-down” processing, but we do see how the architecture enables for self-checking. Granted, this processing feedback may not necessarily map as neatly onto human top-down processing, but I contest that top-down processing is necessary for consciousness. Indeed, research has shown that infants have consciousness as early as 5-months old (Kouider et al., 2013), however the earliest investigation into infant top-down processing is at 9-months (Xiao & Emberson, 2019). If infants can be conscious prior to top-down processing, the opinion that Boss does not contain top-down processing is irrelevant.

In fact, even the article my opponent cited to support his/her claim that motion-planning and action can take place without conscious processing was not relevant to the task of novel, urban driving. Nor was the claim that since humans can “zone out”, a phenomenon called highway hypnosis, while driving, consciousness is not required for the task. When the details are brought to light about where these two types of unconscious processing is sufficient to complete tasks, in the former during a very short-term experimental task and in the latter during a very familiar and/or monotonous drive, it is clear they are not relevant to the conditions in which Boss successfully drove autonomously. This render’s my opponent's rebuttal irrelevant.

Debate Conclusion

Now the debate has concluded, we ask you, Dr. Marques, to decide who you believe constructed the most convincing arguments in the context of this debate. To keep this blind, we have refrained from listing our names attached to the stances (pro/con) up until this point. Once you have made your decision as to the winner (which we would love to hear your opinion as we both think we are the more brilliant debater of the two), you can see who took which stance below. We have written the name of the debater in white text on the white background so you will have to highlight it to reveal the identities.

Con Stance:

Pro Stance:

References

Opening Statement: Con

Kanwisher, N. (2001). Neural events and perceptual awareness. Cognition, 79(1-2), 89-113.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., & Hudspeth, A. J. (2013). Principles of Neural Science, Fifth Editon.

Schmitt, L. I., Wimmer, R. D., Nakajima, M., Happ, M., Mofakham, S., & Halassa, M. M. (2017). Thalamic amplification of cortical connectivity sustains attentional control. Nature, 545(7653), 219-223.

Nakajima, M., Schmitt, L. I., & Halassa, M. M. (2019). Prefrontal cortex regulates sensory filtering through a basal ganglia-to-thalamus pathway. Neuron, 103(3), 445-458.

Opening Statement: Pro

Dijksterhuis, A. (2004). Think different: the merits of unconscious thought in preference development and decision making. Journal of personality and social psychology, 87(5), 586.

Tsuchiya, N., & Koch, C. (2005). Continuous flash suppression reduces negative afterimages. Nature neuroscience, 8(8), 1096-1101.

First Rebuttal: Con

Binsted, G., Brownell, K., Vorontsova, Z., Heath, M., & Saucier, D. (2007). Visuomotor system uses target features unavailable to conscious awareness. Proceedings of the National Academy of Sciences, 104(31), 12669-12672.

Second Rebuttal: Pro

Urmson, C., Baker, C., Dolan, J., Rybski, P., Salesky, B., Whittaker, W., ... & Darms, M. (2009). Autonomous driving in traffic: Boss and the urban challenge. AI magazine, 30(2), 17-17.

Binsted, G., Brownell, K., Vorontsova, Z., Heath, M., & Saucier, D. (2007). Visuomotor system uses target features unavailable to conscious awareness. Proceedings of the National Academy of Sciences, 104(31), 12669-12672.

Closing Statement: Con

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. Nature Reviews Neuroscience, 14(5), 350-363.

Closing Statement: Pro

Kouider, S., Stahlhut, C., Gelskov, S. V., Barbosa, L. S., Dutat, M., de Gardelle, V., ... & Dehaene-Lambertz, G. (2013). A neural marker of perceptual consciousness in infants. Science, 340(6130), 376-380.

Xiao, N. G., & Emberson, L. L. (2019). Infants use knowledge of emotions to augment face perception: Evidence of top-down modulation of perception early in life. Cognition, 193, 104019.