Paper 1:

Explain the Chinese Room Argument and **exactly one** of the following responses:

1. The Robot + Systems reply*
2. The Connectionist Reply
3. The Functional Decomposition Reply

Evaluate this response. Is it successful? Why or why not? **Argue for your answer.**

*If you choose option 1, you are encouraged to consider the *combined* Robot and Systems reply, which is more plausible than either response in isolation.

Part 1: Argument Summary

Mental states have both syntactic properties and semantic properties (Irving, 2021a, slide 17). Because we have already established that computers can compute valid syntactic operations, it begs the question of whether computers can sufficiently generate semantics from these syntactic operations and thereby possess intentionality. According to Brentano "we can define mental phenomena by saying that they are those phenomena which contain an object intentionally within themselves" (Brentano, 1974). This means that intentionality is the *mark of the mental* or to put it another way, all and only minds have intentionality. That's how mental states have *meaning* (Irving, 2021a, 19).

A key part of computers and other computing machines is that they operate on formal logics, systems of rule-defined operations that operate on sets of parameter symbols. The rules are defined in such a way that the operations cannot produce a false or incorrect conclusion from true, valid inputs (Clark, 2001, p. 9). A key example of this is a game of chess. If a board with its orientation and layout of pieces and positions is considered an input and the set of valid moves are considered operations, then as long as both opponents only apply valid operations (moves) to the chess board, all states of the game (chess-board piece layout) are guaranteed to be legal.

This implies that a set of properly construed operations can be conducted syntactically on semantically valid inputs to produce conclusions that share the semantic validity as the original inputs. This implies that computers, which can follow syntactic rules, can operate on semantic properties such as rationality and truth (lecture CTM 2), preserving original semantics and arriving to conclusions that share the validity of the original inputs. This produces the idea that "If you take care of the syntax, *the semantics will take care of itself*" (Haugeland, 1981, p.23, original emphasis).

This leads into the idea of strong and weak AI. Weak AI is the idea that the use of a computer is only as a tool. A highly capable tool but a tool, nonetheless. Conversely, strong AI is the idea that a computer is far more powerful than what weak AI suggests. So much so that a properly programmed computer is not just a computational tool but a mind, a mind that can understand and possess cognitive states (Searle, 1980).

And if a mind has cognitive states as suggested, we can take Brentano's statement and combine it with the idea of Strong AI to conclude that Strong AI possesses a mind. Or to put it another way, a properly programmed computer *is* a mind. A machine that can understand its input, contextualize this input, and provide elaboration on questions addressed regarding this input.

Searle's response to the idea of Strong AI is the Chinese Room Argument. The Chinese Room Argument portrays a room that contains an English-fluent person with no knowledge of Chinese as well as an English rulebook that defines rules to "correlate one set of formal symbols

with another set of formal symbols" (Searle, 1980), specifically Chinese symbols. The person locked in the room is given 3 sets of Chinese symbols. The first set is a script, the second a story, and the third set is a series of questions, again all written in Chinese. Using the rulebook mentioned, the non-Chinese speaking person can "correlate elements of the third [set] with the first two [sets]" and "give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given in the third [set]" (Searle, 1980). The symbols generated through this correlation are offered back as answers to the questions (generated by following the rulebook). And these answers "are indistinguishable from those of native Chinese speakers". Further, if this person were to be offered the same set of script, story, and questions but in English, they would be able to answer just as well as any other English speaker and as well as the rulebook-generated Chinese output but in English.

The person in the room will never have any understanding of the Chinese characters he or she operates on because there is no way to get semantic meaning from only syntactic symbol operations. Because the characters that the person works with have no meaning to him or her, the person does not demonstrate intentionality. In this thought experiment, the human in the room is simply a computer, operating on inputs guided by a pre-defined set of rules just like Strong AI. Because the human in the room does not demonstrate understanding, then by extension, neither does Strong AI.

One of the primary responses against the Chinese Room Argument is the Connectionist Reply, created by the Churchlands. The argument agrees with Searle on the point that the Chinese Room does not understand Chinese, but they maintain that the Chinese Room thought experiment succeeds only because of our lack of understanding about the mechanisms of cognitive and semantic phenomena, appealing to our common-sense intuitions instead (Cole, 2020). For example, we know, empirically, that electromagnet oscillations produce light. But when we attempt to replicate this by waving a magnet around, no light is produced. This leads to the following fallacious argument:

4. Electricity and magnetism are forces
5. The essential property of light is luminance
6. Forces by themselves are neither constituted of nor sufficient for luminance

Conclusion: Electricity and magnetism are neither constituted of nor sufficient for light (Page View, n.d.)

The Churchlands go on to say that because we are unsure about how exactly mind-based computations occur, we cannot propose that such computations operate as proposed in the Chinese Room Argument. Instead, they propose a Connectionist architecture over the symbolic program architecture that Searle suggests. Cetic goes further to suggest the removal of the rulebook altogether and provide only sample inputs and their corresponding outputs.

Part 2: Evaluation

The connectionist response to Searle's argument is highly successful at dismantling the Chinese Room Argument. Primarily due to the connectionist argument's focus on one of the key assumptions of the Chinese Room Argument; the idea that a computer must inherently operate via a symbolic architecture. The problem with the Chinese Room thought experiment is that any computer that relies only on syntactic manipulation, the basis of a symbolic architecture, cannot demonstrate sufficient complexity to create a mind. To put it another way, the Chinese room thought experiment is doomed to fail from the get-go due to the assumption that the computer must exclusively perform syntax manipulation. By expanding beyond the symbolic architecture, the Connectionist Reply introduces possibilities for the Chinese room thought experiment to succeed.

A notable mode for the implementation of connectionist architectures are DCNNs, modeled after the human brain. DCNNs are artificial neural network that feature layers of hidden nodes. Each node performs operations on collections of inputs, inputs generated from either combinations of other nodes or from the original input itself. These operations are trained and re-trained against vast sets of data. Through thorough training on this data, DCNNs can restructure themselves such that innately, within their structure, DCNNs store the generalizations that enable them to correctly map inputs to outputs.

And this is the crux of the argument that back-propagates to the initial exposition, detailed in the argument summary. The idea that Strong AI goes beyond the tool-based properties of Weak AI is true when the artificial intelligence is constructed on a connectionist architecture. A connectionist architecture possesses intentionality due to its innate ability to represent the inputs that it receives. The DCNN will represent relevant aspects of its input via the activation of specific sets of neurons. Just like the human brain, which we assume possess intentionality, each unique input will produce a novel firing of neurons, an aspect that demonstrates that the neural network has the capability for understanding, and by extension intentionality.

A common objection to this idea is presenting the fact that DCNNs are highly vulnerable to adversarial images. According to Goodfellow, DCNNs "consistently misclassify adversarial examples", adversarial examples are inputs that are created by applying nearly unnoticeable "worst-case perturbations" to dataset examples. The problem is that the DCNN will go on to offer an "incorrect answer with high confidence" to the input image (Goodfellow, 2014). But, to a human the image would still appear the same. The objection continues to state that because DCNNs are so easily tripped up by simple perturbations, this must indicate that DCNNs possess insufficient machinery to properly classify images. This argument is furthered by experiments conducted on rubbish images, or images that are so perturbed that even a human would be unable to identify the image. DCNNs continue to incorrectly classify rubbish images with high confidence. From the example of the rubbish images, the argument proceeds to suggest that

DCNNs fail to represent relevant aspects of images. Instead, the argument suggests that DCNN's innate representation of relevant aspects are dissimilar to what humans would consider relevant aspects and are therefore not demonstrating intentionality.

This argument is ultimately fallacious. Just because a DCNN implemented on a computer does not represent objects like a human does not imply that a DCNN does not possess intentionality, that is only a product of Human Chauvinism. More importantly, the fact that DCNNs represent objects differently from humans (evidenced by the mistakes they make with perturbed images) improves the case for DCNN intentionality and understanding. It suggests that DCNNs represent images in a way that is fundamentally different from humans, implying that DCNNs *understand* their input in a completely different way from the way that humans understand their input.

Another objection is the idea that DCNN architecture is fundamentally insufficient for demonstrating intentionality and understanding. After all, a DCNN is simply a series of mathematical operations conducted at various nodes. It should be impossible for any individual node to understand its input. If each node does not understand its input, how could it be that a system of these nodes can understand its input and demonstrate intentionality? Further, at the end of the day couldn't the computations conducted by DCNN's be replicated by syntactical operations? Wouldn't DCNNs be subject to the same fallacies as symbolic architectures?

Humans share the same architecture as DCNNs and one way or another, we demonstrate intentionality. No individual human neuron can demonstrate the intentionality and understanding that is possessed by the collection of billions of neurons makes up the human brain. And the human brain possesses intentionality. This must imply that sufficiently advanced neural networks do indeed possess intentionality. Sufficiently large neural networks can not only be trained to generate input/output mappings, but they can demonstrate intentionality as well. This is due to three of the aspects of DCNNs and other connectionist-based architectures.

Firstly, connectionist architectures perform a series of mappings. This may seem obvious, but connectionist architectures are interconnected in such a way that enables them to have series of mappings wholly unlike symbolic architectures. Connectionist architectures do not perform a single mapping like in the case of symbolic architectures but a *series* of mappings. This enables exponentially more complexity and nuance within the mapping functions which in turn produces intentionality. The mapping functions expressed by symbolic architectures are merely a single node, a component of the total mapping that is performed by connectionist architectures. But by combining billions of transformations, each tailored and trained on millions of input/output samples, highly complex information can be incorporated into the very structure of computational function, to be elaborated on in the next section.

The next notable property of connectionist architectures is the way in which they are created. Connectionist architectures are trained and forced to create their own generalizations, and in turn mappings, independent of the knowledge and intentionality of its creators. Unlike a

symbolic architecture which has mappings that are only derived from the intentionality of its creators, the connectionist architecture can create intentionality, distinct from that of its creators. Not only is this information stored in a unique manner, but it is stored in a way that is fundamentally different from the storage of symbolic architectures. Connectionist architectures store information *alongside* their computational mechanisms. Or to put it another way, the computations themselves are the storage mechanism for connectionist architectures. Whereas in symbolic architectures, computations are simply a means to an end, in connectionists architectures, the computations are what *define* the system.

The third and final aspect that is unique to connectionist architectures is their ability to be highly adaptable. Even when exposed to inputs that are outside the domain of the data set that they were trained on. The fact that connectionist architectures can restructure themselves at a fundamental level, independent of any outside intervention, is a paradigm shift in what computers can accomplish. Unlike syntactical architectures that need specific mappings for all inputs, breaking when exposed to unexpected inputs, connectionist architectures can adapt to incorporate these novel inputs. In the context of the Chinese Room Argument, this describes a machine that would be able to perform the same operations that it performs in Chinese in Portuguese, French, Latin and more.

For these reasons, a Connectionist reply to the Chinese Room Argument is successful in rebuttal. By attacking one of the key assumptions of the Chinese Room Argument, the Connectionist reply can crack open the Chinese Room Argument. Further, by providing an alternative to the architectures assumed by the Chinese Room Argument, the Connectionist reply can expand on the pre-conceived capabilities of computers, solidifying and ensuring the capabilities of Strong AI.

Works Cited

Brentano, F. (1874). *Psychology From an Empirical Standpoint* (Issue 2, p. 241). Routledge.

Clark, A. (2001). *Mindware: An Introduction to the Philosophy of Cognitive Science*. New York: Oxford University Press.

Cole, D. (2020). The Chinese Room Argument. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/chinese-room/

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and Harnessing Adversarial Examples*. https://arxiv.org/abs/1412.6572v3

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea* (Vol. 38, Issue 151, p. 249). MIT Press.

Irving, Z. (2021). *Computer Theory of Mind II* [PowerPoint presentation]. Retrieved from Collab.

Page View. (n.d.). Retrieved October 22, 2021, from https://debategraph.org/Details.aspx?nid=825

Searle, J. R. (1980). Minds, brains, and programs. *THE BEHAVIORAL AND BRAIN SCIENCES*, 8

Paper 2:

Suppose Ankit becomes so intoxicated that he cannot remember anything he thought, felt, or did from midnight until 2am. As we saw in lecture, Locke's theory entails that sober, post-2am Ankit is not the same person as drunk, midnight-to-2am Ankit, and this consequence, we saw, can seem quite implausible.

Locke tries to deal with this sort of worry about his view in §22 of his "Of Identity and Diversity." **Explain** and **evaluate** what Locke is saying in this passage. Is his response convincing? Why or why not? **Argue** for your answer.

Part 1: Argument Summary

A key motivation of Locke's memory theory is the idea that humans, and other sentient beings, are fundamentally different from objects such as masses, substances, and organisms (Locke, Ch. 2-6). Locke states that people are conscious, thinking beings (Locke, Ch.9). Therefore, the identity of a person cannot be anchored in the same things that anchor the identities of substances, vegetables, and animals which are components, functionality, and life-requisite functions. So, whatever grounds a person's identity must be the same thing as what makes a person that specific, distinct person. Locke concludes that this grounding-object is the mind. This leads us to our next question of what enables that mind to persist over time.

Locke builds on the Consciousness Theory, the idea that two people at different times are identical if they share the same consciousness. Locke states that consciousness is the *only* thing that defines personal identity and that a being is the same "as far as the consciousness can be extended backwards to any past action or through, so far reaches the identity of that person" (Locke, Ch. 9). This unifies a person across all time and all ages, suggesting that the person you are right now is the same as the person a second ago, a year ago, and even 10 years ago, all the way back to the inception of your consciousness. Consciousness is a seemingly abstract term so to make this statement more concrete, we assume Locke to imply that consciousness is memory. To put it more specifically, shared episodic memories imply the persistence of a person across time.

Episodic memories are memories of past events "(roughly) as [one] experienced them" (Irving, 2021a). These memories are an essential component of one's identity due to the ability for these memories to be recalled and then replayed and relived such that the net-effect of these memories as they originally occurred can be reinstated. This connects the person in the future time with the person in the previous time who made these memories. This future person is numerically identical to the past person because the future person shares the episodic memories of the past person.

Locke's memory theory gives the proper verdict on three-cases that the body theory struggles with. Namely the non-human persons, body/brain swapping, and life after death cases.

In the non-human persons case, the body theory fails because it assumes that a sentient being must have a human body/brain and is not generalizable to other intelligent creatures such as parrots, elephants, and even sufficiently advanced robots. Locke's memory theory addresses this flaw through one of its core premises. The premise that the only pre-requisite for identity is episodic memories. So, any being, if it has the ability to create episodic memories, can be a person and persist across time (Locke, Ch. 8).

In a scenario where one was to swap bodies, Locke states that the identity follows the location of the consciousness. So, in the movie Freaky Friday, Lindsey Lohan's character persists in her mother's body over the duration of the swap. Locke addresses this scenario

specifically with his Prince and the Cobbler example (Locke, Ch. 15). In this example, Locke suggests the swapping of the Cobbler's consciousness with that of the Prince, or to put it in simpler terms, the exchange of episodic memories between bodies. Locke goes on to make the argument in the consecutive chapter that legal punishments must be determined based not on "substance, material or immaterial, or no" refuting the body theory's idea that the self is dependent on being "made up of the same or other substances" but instead on self-consciousness and the persistence of memory throughout time.

Locke's final address to the life after death case is the most distinct from that of the viewpoint presented by the body/mind theory. An implication of the body/mind theory is that there is no life after death. After one's body is destroyed it is impossible for that same person to live again, unless of course the body is resurrected. Locke's Memory Theory suggests that post-resurrection one will retain their identity as long as they have "episodic memories of [their] old life" (Irving, 2021). The memory theory shoves the question of life after death off of the mechanism of identity and onto the mechanism of resurrection/death, specifically whether or not memories are preserved beyond death.

There is one other scenario where Locke's memory theory is placed in jeopardy. The case of a drunk actor. In the case of someone who commits a crime when drunk, Locke notes that the individual is held legally liable despite lacking moral responsibility. He proceeds to argue that this is only due to a shortcoming of mortal judges. Mortal judges have no knowledge of the defendant's memories and can only judge assuming that a person was conscious throughout the act, disregarding the defendant's memory of events. But Locke notes that in an idealized scenario, only memory of the crime can be used as a measure of liability. This solidifies Locke's position that memory and only memory is the locus of identity (Locke, Ch. 22).

Part 2: Evaluation

Locke's response to the drunk actor scenario is solid and in-line with his previously stated opinions. Particularly artful in Locke's response is that he acknowledges the clear discrepancy between his stated theory and the real-world but is still able to use this scenario to drive home one of his key points. The fact that human judges use a body theory framework in assigning legal culpability is begotten of our own limitations. The inability to independently verify that the defendant does or does not have memories of a crime that their body committed. Locke continues to note that if this independent verification were possible, then only memory of crime would be necessary to assign culpability. This nuance leads to the conclusion that moral responsibility and legal responsibility are two distinct phenomena.

From Locke's argument, we find that moral responsibility is contingent upon memory of the action that one is responsible for. This is in-line with Locke's Prince and Cobbler scenario where the exchange of identities causes the responsibility of a crime to follow consciousness. To put it another way, the lack of common, shared memories between a drunk actor and the sober individual indicates that the two are different people. This leads to the conclusion that the sober individual should not be punished for the actions of the drunk actor. Locke essentially states that in an ideal world, memory leads to identity which leads to moral responsibility which in turn leads to legal responsibility. This is because according to Locke's memory theory, identity resides in memory.

But memory is a poor vessel for identity. Particularly for humans, memories are particularly fragile. Human memories can be affected by a variety of diseases and ailments including Alzheimer's, dementia, and more. Even disregarding catastrophes such as diseases even seemingly innocuous events can cause memory loss. This includes amnesia, depression, anxiety, and trauma. Therefore, Locke's theory that memory is the keystone of identity is built on shaky ground.

I propose a different locus of identity. I propose that identity is composed of how a person is shaped by their experiences and memories. This person's moral identity is most succinctly defined by a core set of beliefs and ideals. This formulation of identity is removed from an individual's capacity to remember and recall memories and is instead dependent only on what an individual learns from his or her experiences. This definition of identity is only a qualification of memory theory but is essential due its ability to define identity independently of memory's volatility.

For example, a renowned mathematician suffers a freak accident that causes this mathematician to lose the much of his memories prior to the accident. He recognizes no one and nothing around him. But when taught basic computations he recognizes them instinctively and quickly regains his ability to perform computations. Further, when exposed to his own mathematical research, he recognizes it as his own despite having no knowledge of conducting said research. This phenomenon extends to his memories of people. When exposed to a variety

of his acquaintances, the mathematician distinguishes friend from foe despite being unable to say why. Locke's memory theory suggests that this mathematician is not identical to the mathematician prior to the accident. But my experience-based theory produces a different result. The mathematician retains his identity despite lacking his memories, the mathematician's experiences such as practicing calculations, conducting research, and interacting with people fundamentally changed who he was as a person, shaping his identity permanently.

In the context of the drunk actor scenario, this experience-based theory produces a different conclusion than Locke's memory theory. Recall that Locke stated that in mortal legal procedures, a defendant would be determined to be legally responsible despite being unable to prove moral responsibility or common identity/memory. While in an ideal world, moral responsibility would lead to legal culpability. The experience-based theory suggests that in a drunk actor scenario, the drunk actor would always be responsible for any actions taken regardless of whether the actor can recall his or her actions.

This conclusion is drawn from the fact that through the experience-based identity framework, the actor is identical before drinking, while drinking, and after drinking making him or her legally responsible regardless of whether he or she remembers what occurred while inebriated. This is because the same person that is willing to rob or murder while drunk possessed those intentions while sober. It is only that the actor's inhibitions were lowered while drunk that enabled him or her to follow through with the crime. The same built-in beliefs and ideals that led this person to commit a crime while inebriated were still present when this person was sober its just that these beliefs were inhibited. Therefore, the assignment of moral responsibility is essentially the punishment for retaining beliefs and ideals that are incongruent with society's laws.

Intentions, wants, and needs are derived from a person's personality, which in turn originates from what he or she learned from their experiences. The same morality that would possess someone to rob a bank while inebriated is the same morality that conceived of such a want while sober but said want simply went unexecuted.

A potential objection to this view is that personality changes can occur as well, which by this definition would indicate a change in identity. For example, a conniving businessperson can become a pious saint later in life, discarding all connections to his previous life. The objection would argue that because this person is so different in personality, he must be a completely different person. But I would disagree with the conclusion of such an example, this change of outward personality is not a reflection of the change of inward identity, it is only a different expression of this individual's identity, he only selects to present a different personality than the one before. To pursue the moral responsibility framework, this newly made saint would still need to answer for the actions taken as a conniving business-person indicating that the saint and the conniving businessman are one and the same people despite displaying different personalities.

Another objection would be the immortal being objection. Humans are malleable creatures with varying beliefs and ideals over time, therefore a person who grows and ages over time would not be the same person throughout their lives. This is even more true of an immortal being who would experience only more drastic changes over the course of their lifetime. To this I respond exactly. It is the persistence of beliefs and ideals that is linked to a constant identity and nothing else. The manipulation of these beliefs and ideals is what transforms someone into someone new. The beauty of the experience-based identity theory is that it accommodates change, particularly experience-induced change that is characteristic of a long life.

Due to the problems inherent in Locke's memory theory of identity, particularly in retaining memory over extended durations of time, as well as in the assignment of moral responsibility, Locke's memory theory fails when dealing with memory loss. Although the proposed experience-based identity theory is far-from perfect, it demonstrates the failures of the memory theory of identity and supports the viability of alternatives to the memory theory of identity.

Works Cited

Locke, J. (1690). Of Identity and Diversity. In *An Essay Concerning Human Understanding*.

Irving, Z. (2021). *Memory Theory I* [PowerPoint presentation]. Retrieved from Collab.