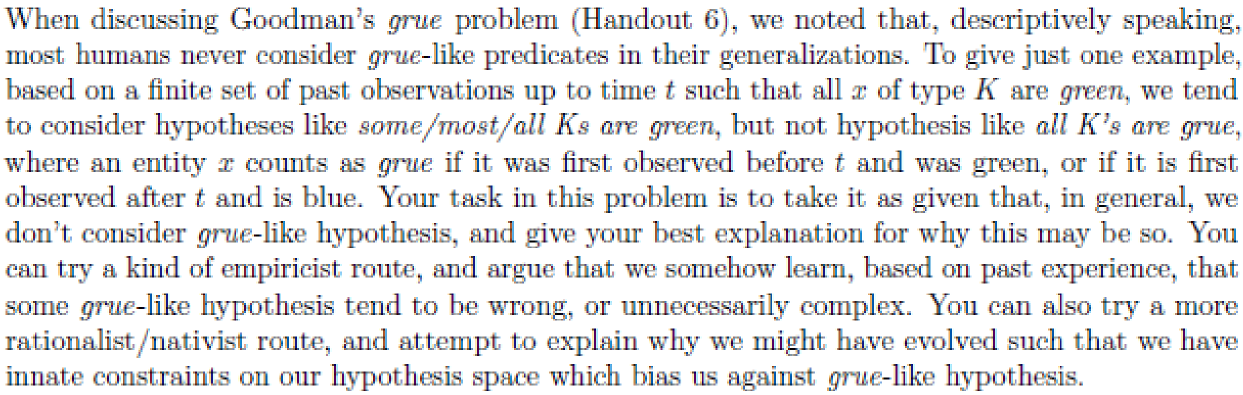
Short Essays on Minds and Machines
Safeer Ahmed
05/09/2022

Prompt:



When discussing Goodman's *grue* problem (Handout 6), we noted that, descriptively speaking, most humans never consider *grue*-like predicates in their generalizations. To give just one example, based on a finite set of past observations up to time t such that all x of type K are *green*, we tend to consider hypotheses like *some/most/all Ks are green*, but not hypothesis like *all K's are grue*, where an entity x counts as *grue* if it was first observed before t and was green, or if it is first observed after t and is blue. Your task in this problem is to take it as given that, in general, we don't consider *grue*-like hypothesis, and give your best explanation for why this may be so. You can try a kind of empiricist route, and argue that we somehow learn, based on past experience, that some *grue*-like hypothesis tend to be wrong, or unnecessarily complex. You can also try a more rationalist/nativist route, and attempt to explain why we might have evolved such that we have innate constraints on our hypothesis space which bias us against *grue*-like hypothesis.

Induction

Generally, we cannot consider *grue*-like inductive hypotheses, though, we typically tend to reasonably accept some other forms of inductive hypotheses with non-*grue*-like predicates. Hence, some form of explanation is required for determining how to differentiate between a *grue*-like hypothesis and other different hypotheses. Specifically, a different hypothesis that one would generally have less to dispute with. For instance, one of these less disputable hypotheses that I will refer to for the rest of this response is the "simple" inductive hypothesis that has *green* being the predicate for emerald discovery. I describe it as simple because it is a hypothesis that would likely be accepted by a majority of humans, is quite intuitive, and more commonsensical – equating to being less disputable. But why is this so? Also, why is it the case that *grue* is a more disputable predicate if it is equally provable in both inductive hypotheses that emeralds will or will not be *grue* and/or *green*? I argue that the key reason for differentiating between the two predicates and preferring *green* comes down to the complexity of the predicates – *green* is a non-complex predicate while *grue* is a complex predicate. To elaborate further, I will first go into depth on why I believe *grue* is more complex than *green* – plus generalize how I define complexity in predicates. Then, I will explain how as humans we have come to indicate the varying complexities of predicates and how we determine their application when inductively reasoning.

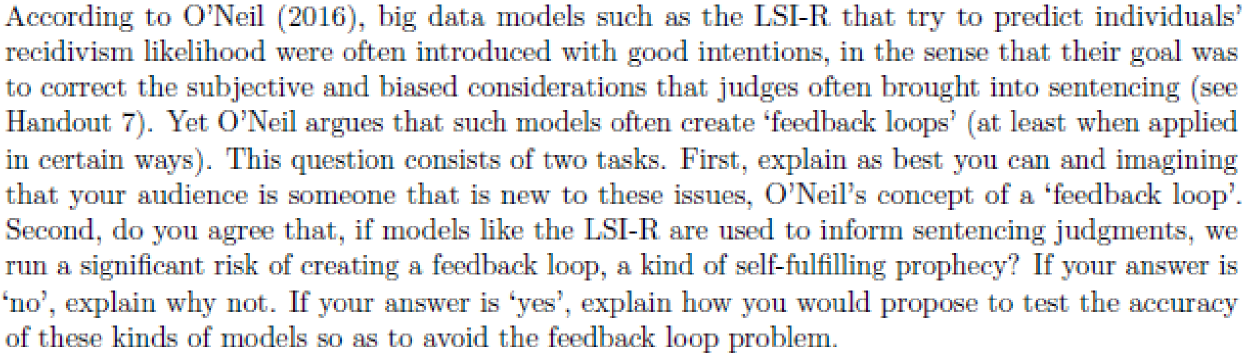
The key reason why *grue* is a complex predicate boils down to its underlying definition. *Grue's* definition contains two different predicates as we defined it in class: the predicates of *green* and *blue* depending on the time of the emerald's discovery. Compare this to the simple inductive hypothesis aforementioned earlier – it only had the single predicate *green*. Simply by being more complex, it is harder to justify both predicates when trying to defend the *grue* hypothesis. This notion of underlying predicates adding more restrictions to the overarching predicate in *grue*-like hypotheses is why I believe that they are less intuitive than a simpler hypothesis. An opponent of the *grue*-like hypothesis, then, could utilize Occam's Razor and object *grue* as being too complex. *Green* would be the simplest explanation to the inductive problem at hand because adding more restrictions on hypotheses disrupts simplicity.

Furthermore, how I see it is that a predicate's complexity "scales" based on how one can interpret the number of concealed predicates or restrictions within the semantics of the predicate itself. For instance, the predicates of being a president and an adult vary in complexity. There are far more restrictions and predicates hiding within the definition of a president – even being an adult is a requirement itself for the president predicate. This example should solidify the notion that there is clearly some sort of degree of complexity when comparing different kinds of predicates – *grue* and *green* have this different degree of complexity.

Although I have stated that in the specific case of *green* versus *grue*-like hypotheses are more complex based on having more complex predicates, sometimes added complexity is necessary depending on the situation. Even though this defers Occam's Razor in a sense, my stance is that sometimes it might be better to add more complexity to properly formulate a hypothesis for a given scenario. In the case of emeralds being discovered having a certain state, the simple predicate is the preferred predicate because the scenario of emerald discovery is simple itself. The complexity of the scenario plays a role in the type of predicates that should be involved. To me, experience is the driving factor in understanding inductive hypotheses for any varying scenarios. Take for example an inductive hypothesis concerning a chameleon that lives on Earth. It would make more sense to utilize a more complex predicate in an inductive hypothesis indicating the color of chameleons discovered. This is because this scenario involves a more complex base object. Humans can differentiate between the complex nature of objects within inductive hypotheses based on prior experience with said objects. We understand that chameleons change their skin depending on the environment with changing cells, but emeralds are non-living, non-dynamic objects from our experience with them. There are no logical grounds for justifying a jump from *green* to *blue* for emeralds, though, we can justify this to an extent for chameleons. (Maybe due to rising sea levels, chameleons want to match the excess water and be *blue*.) Hence, the complex predicate *grue* might be more appropriate for chameleons than emeralds. This example diverts from the notion that some/most/all chameleons were *green* up until now since we know they change colors; nevertheless, context still matters for both examples and is crucial for forming hypotheses inductively. To put it concisely, there just is not enough justification or prior knowledge to explain why emeralds would ever be *blue* after a given time. So, introducing a complex predicate is not necessary and generally dismissed when dealing with emeralds' discovery.

Essentially, what I am trying to argue is that it truly depends on the case at hand, and it is tricky to concretely answer if complex predicates are proper for making inductive hypotheses. For emeralds, there is no reasonable justification for making such a *grue*-like hypothesis with very complex predicates for a non-complex case. Humans learn and develop this intuition of determining complexity by analyzing the objects within each case and determining if making any jumps between predicates is applicable. Sometimes restrictions are imposed into definitions of predicates which makes them more complex – but the new restrictions require past experience to justify their introduction, so their complexity can match the context of the inductive scenario.

Prompt:



According to O'Neil (2016), big data models such as the LSI-R that try to predict individuals' recidivism likelihood were often introduced with good intentions, in the sense that their goal was to correct the subjective and biased considerations that judges often brought into sentencing (see Handout 7). Yet O'Neil argues that such models often create 'feedback loops' (at least when applied in certain ways). This question consists of two tasks. First, explain as best you can and imagining that your audience is someone that is new to these issues, O'Neil's concept of a 'feedback loop'. Second, do you agree that, if models like the LSI-R are used to inform sentencing judgments, we run a significant risk of creating a feedback loop, a kind of self-fulfilling prophecy? If your answer is 'no', explain why not. If your answer is 'yes', explain how you would propose to test the accuracy of these kinds of models so as to avoid the feedback loop problem.

Big Data and Society

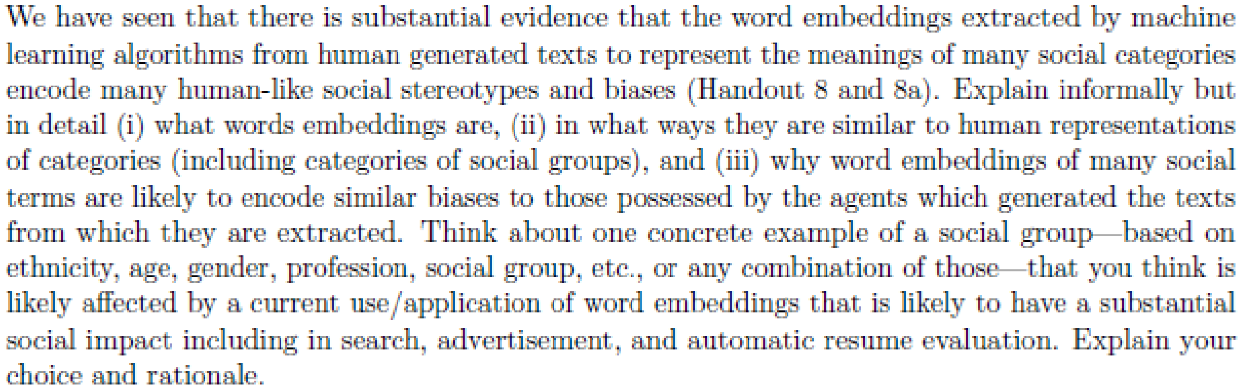
The book "Weapons of Math Destruction" by Cathy O'Neil outlines various big data algorithms and predictive models that are used to solve social problems with the caveat being that they can cause detrimental damage. The core reason why such algorithms/models are introduced in the first place is so that they can remove certain negative selective biases caused by human subjectiveness. However, a major issue leading to the damage by some predictive models is the possibility of feedback loops occurring. O'Neil's concept of a feedback loop boils down to what she calls a "toxic cycle" where a predictive model utilized to some capacity breeds an outcome that increases the odds of the same prediction occurring again, repeatedly. The reason she likely coins the phrase "toxic cycle" is because when negative feedback loops occur in a model, they can instigate collateral damage to the social problem being tackled and even make the situation at hand much worse than before introducing these models. This means that subjective outcomes can be produced which yield the toxic overall result – the models end up being counterintuitive. In order to explain feedback loops more thoroughly and build upon my later proposal, I will first go into depth to explain feedback looping in the example of LSI-R predictive models for sentencing judgment that O'Neil mentioned. Then, I will wrap up with my proposed method to test the accuracy of these kinds of predictive models to prevent feedback looping as much as possible.

A strong example of a case where feedback loops debatably occurred is in the model called LSI-R. The goal of the model is to predict the likelihood of an inputted convicted individual getting rearrested for a similar offense in the future. The reason that a model like this was introduced was to minimize the problem of human judges segregating convicts based on their own individual biases when sentencing. LSI-R works by processing information found using questionnaires given to each convict.

LSI-R is definitely a more popular model for judging sentences, but that does not equate to it necessarily being free of feedback loops. For the sake of brevity, I will not hyperfocus on the first two questions of the questionnaire because they heavily relate to each other questions – which is part of the weakness of LSI-R. Instead, the overly problematic question three makes more sense to attack here for briefness. Question three pertains to the criminal records of the convict's friends and family. The information obtained from this question to build the prediction directly produces the feedback loop as I defined earlier. Locations where the LSI-R is implemented for sentencing convicts are commonly segregated on racial, ethnic, and/or economic status. However, the three listed statuses are correlated to some degree as supported by data from the United States Federal Reserve (Bhutta et al.). We can likely assume that this correlation holds in most, if not all, of the locations that utilize the LSI-R predictive model. Additionally, poorer areas that heavily segregate against minorities are commonly heavily policed in comparison to other areas (Smyton). It is obvious that more policing in these specific areas will yield more individuals being convicted of crimes, even assuming that crime rates are constant throughout different racial, ethnic, and/or economic statuses. Each time an individual is convicted in a higher policed area, they will alter the LSI-R prediction to favor harsher convictions for anyone in the higher policed area. This is because the next time someone from that same area is faced with question three, they have to respond with a higher value. This will just constantly loop to the point where certain groups that were being discriminated against by human biases will end up being convicted equally harshly (or harsher) than before prior to introducing the model.

Overall, with the context above I agree with the stance that any predictive model for sentencing judgments in similar capacities runs the same risk as LSI-R in having feedback loops – either positive or negative. Note that up until this point I have only mentioned negative feedback loops, though, positive ones would be similar in some sense but have bias alleviating outcomes. Now, it is important to gauge how to test the accuracy of these types of models. It is important because it can minimize the occurrence of negative feedback loops. My proposed method relies heavily on the section prior concerning the rise of feedback loops in models. Recall that the core reason why feedback loops occur in LSI-R is due to its implementation into contexts where information gathered directly correlates to the subjects' own context itself when being judged. Put simply, when LSI-R-like models predict an individual that is in a heavy police zone context, for example, information gained (like from question three) depends on the individual's context (like his friends/relatives). This issue is exactly how I would propose to measure the effectiveness of predictive models. My proposal for testing sentence judging models accounts for the relationship between the context in the questions for information gathering and the context in which the subject is in. Basically, a model that has no relationship is the goal. However, even question one which is less problematic than two and three in LSI-R has some relation. I believe that it is extremely difficult to reach the no relation goal, but models can get close. One example I have in mind is maybe in a model we can gather information from a convict by asking for age. The age of a convict has less relation to the level of policing between a poor and a rich area (most likely). Thus, the contexts, between information gathering and the input individual, have a lower relationship value than questions like question three. I am aware that this is definitely not a solution to modeling sentencing – it is just an example of a more optimal information-gathering question with a less related context to the subject's contextual situation. I also recognize that this proposed test might not yield the best predictions, yet the test will for sure lead to less rampant and "toxic" feedback loops.

Prompt:



We have seen that there is substantial evidence that the word embeddings extracted by machine learning algorithms from human generated texts to represent the meanings of many social categories encode many human-like social stereotypes and biases (Handout 8 and 8a). Explain informally but in detail (i) what words embeddings are, (ii) in what ways they are similar to human representations of categories (including categories of social groups), and (iii) why word embeddings of many social terms are likely to encode similar biases to those possessed by the agents which generated the texts from which they are extracted. Think about one concrete example of a social group—based on ethnicity, age, gender, profession, social group, etc., or any combination of those—that you think is likely affected by a current use/application of word embeddings that is likely to have a substantial social impact including in search, advertisement, and automatic resume evaluation. Explain your choice and rationale.

Word Embeddings and Society

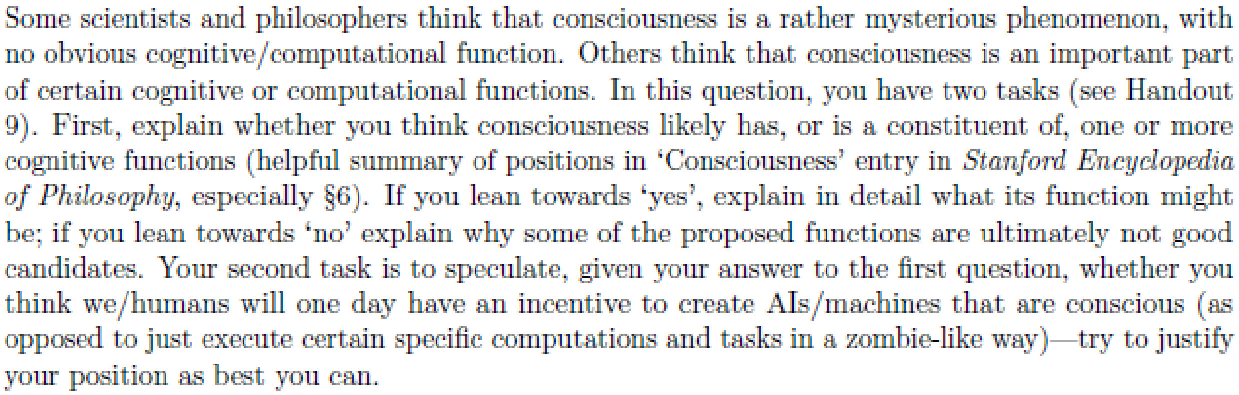
(i) Natural language processing using AI is a tricky task but using an approach with word embeddings has benefits. Word embeddings are essentially identified groupings of various words that have similar semantics. This allows words that have similar, or identical, meanings to have related representations. Words are grouped into vectors in a vector space in a dense manner. By being dense, neural networks can have better performance because “neural network toolkits do not play well with very high-dimensional, sparse vectors” (Goldberg 92). The reason why word embeddings are used in tangent with deep learning AI models and tests are that the way words are mapped to vectors themselves can also bear a resemblance to neural networks. Additionally in class, we also touched upon different association tests that utilize word embeddings in varying ways. For instance, WEAT grouped words based on textual context. This is likely because some words have different meanings depending on the context where they are used. Words such as “bank” can either be used to reference a riverbank or a financial institution; hence, understanding the associated context where the word is located is necessary for forming mappings.

(ii) Word embeddings are similar to human representations of categories in the fact that they capture the categorization of our nearby environments, and sometimes more specifically, our societies – our real-world contexts. Any object in our environment has features that make it distinct from other objects. For example, humans have the ability to represent different animals in categories based on the attributes of their biology. Like we have obvious ways of representing say a fish and a horse because of their biological makeup. Horses cannot breathe underwater, but fish can; hence, why humans have justification to categorize them differently. Fish are categorized differently than horses because of the way they interact with their ecology using their biological features as well – the context. Essentially, this type of categorization is similar to how to word embedding represents categories. As I mentioned earlier, the context of where words are derived from and how they semantically function generally determines how they are categorized into different vectors. Thus, the way word embeddings utilize context to form representative vectors is identical to how humans utilize environments to form representations of categories. This processing of identifying information from words in language based on their context, overall, makes word embedding quite similar to how humans represent categories. Word embedded categories have other useful reasons for being grouped into vectors that mimic how humans differentiate between different types of objects in the environment. Broadly speaking, the main reason for grouping is to grasp the context as mentioned, yet the usefulness extends to more specific areas of contexts. One of these is what I mentioned earlier humans build categories based on societal contexts. For instance, humans often tend to base their categories due to societal realities and biases that occur, so the information gathered by word embedding processes tries to capture similar aspects; however, I will focus on this discussion in section (iii).

(iii) Because human-environmental context is extremely important in defining categories (as elaborated in section (ii)), writings created by humans have numerous different categorizations that contain concurrent categories derived from each individual environmental context. Basically, because humans live in societies that have biases, their writings tend to mirror similarly biased categorizations; thus, word embeddings from agents that extract social terms from said texts will have equivalent biases. To solidify this point, I will focus on the case of translating languages, more specifically, the translation between languages that contain gender pronouns to those that do not. There are numerous languages that are genderless but here let us consider the example from handout eight: English versus Bengali. Without getting too much into the nitty-gritty details of the origins of language and linguistic studies, we can still recognize that the specific writings from regions speaking either language will lead to drastically different embeddings. These embeddings extracted will have similar biases to the original author agents solely due to the limitations of the languages. We can consider a random book, call it Book A, that is written in English by an English author in England. Assume that another random book is selected, call it Book B, written in Bangladesh is authored by a Bangladeshi writer written completely in Bengali. Now picture a scenario where an agent deciphers word embeddings from both books A and B. Let us call the number of categories n be the extracted social terms from A. My claim is that if we have word embeddings from the same agent extracting social terms from B, then we are bound to have less than n categories – assuming that the agent does not self-correct its word embeddings for any biases. The justification for the claim is that since B is written in Bengali, a language that has genderless pronouns, it must not have any possibility for a categorization that is entailed by gender pronouns. The lack of gender due to the context of the language will always net the chance of there being fewer word embeddings for the Bangladeshi Book B.

The consequence of this phenomenon likely varies depending on the severity of the situation and how contexts influence the sheer number of categorical differences – numerically speaking. One substantial downfall that could occur for the case aforementioned case regarding books is as follows: consider that instead of books, the random selections are blog posts from both same authors online. If an advertising company was approached by both authors to put adverts on both blog posts, the company would have no way of specializing in what type of gendered products it could push to readers of the Bengali blog. Online advertising companies likely prefer to have specialized marketing depending on the audience of the blogs. Their tool for extracting word embeddings from both blogs would have fewer categories for Blog Post B. Thus, the downfall from the similarly encoded bias negatively impacts the “advertiser friendliness” of blog B in comparison to the English Blog Post A – solely due to gender biases that are more easily depicted from the text in English.

Prompt:



Some scientists and philosophers think that consciousness is a rather mysterious phenomenon, with no obvious cognitive/computational function. Others think that consciousness is an important part of certain cognitive or computational functions. In this question, you have two tasks (see Handout 9). First, explain whether you think consciousness likely has, or is a constituent of, one or more cognitive functions (helpful summary of positions in 'Consciousness' entry in *Stanford Encyclopedia of Philosophy*, especially §6). If you lean towards 'yes', explain in detail what its function might be; if you lean towards 'no' explain why some of the proposed functions are ultimately not good candidates. Your second task is to speculate, given your answer to the first question, whether you think we/humans will one day have an incentive to create AIs/machines that are conscious (as opposed to just execute certain specific computations and tasks in a zombie-like way)—try to justify your position as best you can.

Consciousness and AI

I lean on the side that consciousness does indeed have some cognitive functionality that would otherwise not be possible without it. It is difficult to quantify a number of functions, so it is unclear how many cognitive functions exist specifically due to consciousness; however, I think a key cognitive function that our consciousness provides is the function of being aware of emotions, but more precisely, the ability to understand other cognitive creatures' emotions. Basically, consciousness is important in facilitating the cognitive function of empathetic behavior. The aforementioned function was my initial response to the question at hand. And after referencing back to the Stanford Encyclopedia of Philosophy, section 6.3 regards social coordination with some similarities. The encyclopedia's description partly coincides with my core stance that consciousness is the source of empathetic behavior, though, the encyclopedia just frames it with more emphasis on the social outcome aspect deriving from empathy. Overall, for this response, I will focus solely on the function of having empathy itself and humans' effort to spread it to AIs/machines in the future.

A key reason why I believe that consciousness plays a specific role in empathetic behavior is because of the cases where we try to consciously think of what others are consciously thinking of. To me, our brains are not biologically hardcoded to try to subconsciously calculate and depict the possible emotions others are feeling. Instead, it is our conscious behavior to attempt to reimagine the feelings of others onto ourselves. I would argue that the level of empathy one has depends on upbringing and correlates with how our consciousness develops over our lifetimes. For instance, consider two identical twins that are separated at birth. In this scenario, suppose that one of the twins is raised with a lot of wealth and only interacts with other wealthy individuals (call him/her A). Now suppose the other individual twin is raised poor while solely interacting with other unfortunate individuals (call him/her B). When faced with a decision to support laws that either raise or lower taxes for rich folk, A and B will likely have different stances on the laws based on their empathy for others. A would likely have less empathy for the poor populace when voting and B might be the opposite. Biologically speaking, we can assume that they have nearly identical physical brain parts. Although A and B obviously have different memories stored 'physically' within their brain, each of their consciousnesses has the function of determining the strength of their empathy felt by processing the memories experienced. I cannot find a reason to believe that this process can be done without any conscious thought or processing – though I would love to hear opposition to the way I described the situation at hand. Therefore, it is clear from the example that consciousness can correlate and can develop based on varying upbringings.

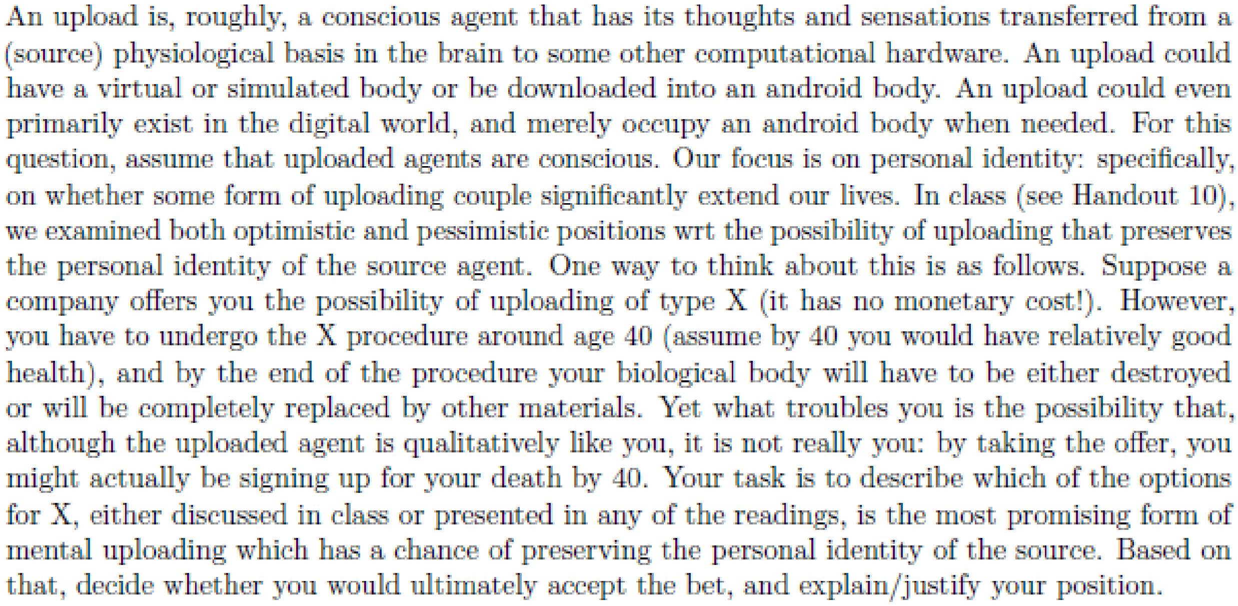
Humans can and will have the desire to create advanced electronic machines. The clear incentive is to assist in our day-to-day tasks and further advance society by computing things faster and more efficiently than our biological brains. I additionally believe that humans should have the incentive to make these super-advanced machines or AIs have their own consciousnesses. To approach my belief and defend my claim, I will begin with the question of "why not?".

Consider that we never decide to design a conscious AI/machine, or possibly we restrict machines from learning how to be conscious as a countermeasure as well. What would be created is a remorseless, zombie-like powerhouse computational device. This device would indeed assist human capabilities in certain areas, but it would lack the fundamental features necessary for helping other areas. Let us take for example an assistant for psychiatrists or therapists. The machine would truly never be able to actually empathize with the clients since it would rely on brute-forcing responses. Moreover, jumping back to the original consideration of restricting machines from being conscious, would yield a contradiction in a sense. We would never be able to invent the most advanced machine if it has the limitation of never being able to learn how to be conscious. With the restriction, it would never be able to be infinitely knowledgeable on everything. Also, the zombie-like machine would have absolutely no function of empathy toward humans, so it can completely discard us. The machine might reach a point where it would no longer have a need for humans and algorithmically decide the best outcome is to rid humans from existence – common in dystopian fiction where machines eradicate humans.

On the contrary, consider that we design AIs/machines with the intent of having them become conscious on creation, or at some later point after creation – likely through learning/developing consciousness. As aforementioned, consciousness brings with it the function of improved social interaction, but more specifically empathy! The AIs/machines can be programmed or strictly taught that humans are friendly and should never be extinguished by them from existence. Like the A and B example, we can positively craft the consciousness within the AIs/machines and prevent a total doom outcome.

Through all the examples and reasoning up to this point, it is evident to me that there is definitely some sort of incentive to give AIs/machines some form of consciousness. In doing so, the conscious creations will be better equipped to assist in even more applications, like being a therapist assistant, and empathize positively with humans to prevent being negatively harmed by brainless AIs/machines that might deem us less inferior. As a forethought to this conclusion, recall that upbringing does have some impact on consciousness, so it is necessary to put emphasis on that feature if we commit to the incentive of building up conscious AIs/machines.

Prompt:



An upload is, roughly, a conscious agent that has its thoughts and sensations transferred from a (source) physiological basis in the brain to some other computational hardware. An upload could have a virtual or simulated body or be downloaded into an android body. An upload could even primarily exist in the digital world, and merely occupy an android body when needed. For this question, assume that uploaded agents are conscious. Our focus is on personal identity: specifically, on whether some form of uploading couple significantly extend our lives. In class (see Handout 10), we examined both optimistic and pessimistic positions wrt the possibility of uploading that preserves the personal identity of the source agent. One way to think about this is as follows. Suppose a company offers you the possibility of uploading of type X (it has no monetary cost!). However, you have to undergo the X procedure around age 40 (assume by 40 you would have relatively good health), and by the end of the procedure your biological body will have to be either destroyed or will be completely replaced by other materials. Yet what troubles you is the possibility that, although the uploaded agent is qualitatively like you, it is not really you: by taking the offer, you might actually be signing up for your death by 40. Your task is to describe which of the options for X, either discussed in class or presented in any of the readings, is the most promising form of mental uploading which has a chance of preserving the personal identity of the source. Based on that, decide whether you would ultimately accept the bet, and explain/justify your position.

Personal Identity and Mental Uploading

The most promising form of mental uploading which has a chance of preserving the personal identity of the source would be the option of gradual uploading. Specifically, I will structure gradual uploading under the following form: some type of nanotechnology device/chip attaches itself to a single brain part and undergoes simulation of that single brain part. This device would eventually replace every part of the brain as the device in each increment would overtake the original brain part after the device mimics the majority behavior of the original brain part. In class, we discussed the devices transmitting processes to an outsourced computer remotely; however, for the sake of my response, I would like to assume that every device replacing the brain parts can function locally – this yields an end result that is much more similar to the initial state of the brain, so comparison of personal identity is more "one-to-one" in a sense. For this topic, I will begin by going over a slightly technical aspect and the possibility of mental uploading with maintaining personal identity, then reapproach the main issue of it whether there still is the same identity after upload, then finally, conclude with my stance on the proposed bet.

The reason I consider the described form of mental uploading above to be more one-to-one is since it removes the factor of there being two separate spatiotemporal entities. Essentially, by having everything managed at the site of the source body, it preserves some aspects of personal identity. Say for instance that we consider the transmission strategy where there is a remote computer. Imagine that we separate the source body/android and the processing computer by an absurd distance in space – possibly galaxies apart. In order for the source to actively interact with the physical world (or even a simulated one), there would be an inevitable delay in interaction because the transmission would need to follow the laws of physics and has a speed limit. This obviously impacts the personal identity of the new mental formation because there would be a delay in every action. The remote mind would not be able to keep up with both the physical world and its stimuli and a simulated world and its stimuli. On the contrary, if we assume that all functions are done locally, then this issue would be avoided entirely. This is why, in general, I think the most promising form of mental uploading where identity is preserved to be possible necessitates that the new “more advanced brain/mind” should be functioning and computing exactly where the source sensations and stimuli occur.

Returning back to the question, “how can mental uploading still maintain the personal identity of the source mind?”, can be more eloquently answered through gradual uploading. If we go the route of destructive uploading or any non-gradual method, it creates a chance for personal identity to get completely lost in the process. By there being a disjoint and a jump with a glaring change in a short time period, the electronic components have a much higher chance of losing the original personal identity of the mind. Gradual uploading, on the other hand, has a linear transformation of the mind. Nonetheless, an opponent to gradual uploading can reiterate the same aforementioned question and instead argue that gradual uploading also has underlying risks of losing grasp of the original mind’s personal identity like non-gradual methods. They can justify their claim by arguing that a gradual shift also has a possible disjoint gap between the original mind and the electronic mind. The issue is of vagueness, however, in a similar manner to the vague state of transitioning from being not bald to bald. I would argue against this using the current understanding of hemispherectomies as a defense. Without getting too technical, essentially, children that had half their brain removed but grew up and functioned perceivably normally because their brains were able to adjust and overtake the functionality of the missing half. Then, it might be entirely possible to assume that if we gradually replace parts of a brain over a long enough period of time, the functions of the brain and the mental state of a person will adapt accordingly to the newly inserted nanotech/electronic/computerized components. To put it simply, the mental state of a person – including their personal identity – would flow through the new technology as they did through their original neurons. The issue with a non-gradual approach might then be that the mind and its underlying functions would not be able to adapt if changes are too disjoint, leading to a loss of personal identity taking place.

Ultimately, my original stance on the subject was entirely pessimistic; however, over time it has shifted to be more optimistic, and I would definitely be open to taking the bet of uploading my mind gradually through the process I described – if the company with said technology existed of course. I believe that a non-gradual method would basically kill me, but more specifically my own identity. I think it would create a new identity just like me. Instead, I think an overtime replacement of my mind with advanced technology mimicking my brain’s current neurons would allow my identity to persist alongside the mental upload. Regardless of whether I am wrong or right on mental uploading, maybe suppose that it is actually impossible to upload a mind that has its original personal identity that persists, the concept of me would be immortalized. My memories of events and mannerisms of how I interacted with the world would be maintained in this digitized copy of my mind – which would very likely outlive my biological self. Therefore, I think it is a risk worth taking, even if it is somehow not my own personal identity persisting across the upload.

Sources:

[1] Bhutta, Neil, et al. “Disparities in Wealth by Race and Ethnicity in the 2019 Survey of Consumer Finances.” The Fed - Disparities in Wealth by Race and Ethnicity in the 2019 Survey of Consumer Finances, Board of Governors of the Federal Reserve System, 28 Sept. 2020, https://www.federalreserve.gov/econres/notes/feds-notes/disparities-in-wealth-by-race-and-ethnicity-in-the-2019-survey-of-consumer-finances-20200928.htm.

[2] Smyton, Robin. “How Racial Segregation and Policing Intersect in America.” Tufts Now, Tufts University, 17 June 2020, https://now.tufts.edu/2020/06/17/how-racial-segregation-and-policing-intersect-america.

[3] Goldberg, Yoav. Neural Network Methods for Natural Language Processing. Morgan & Claypool Publishers, 2017.