International Journal of Machine Consciousness

COALESCING MINDS: BRAIN UPLOADING-RELATED GROUP MIND SCENARIOS

KAJ SOTALA
Department of Computer Science, University of Helsinki
Helsinki, Finland
kaj.sotala@helsinki.fi

HARRI VALPOLA*
Department of Information and Computer Science, Aalto University
FI-00076 Aalto, Finland
harri.valpola@aalto.fi

Received Day Month Year
Revised Day Month Year

We present a hypothetical process of mind coalescence, where artificial connections are created between two or more brains. This might simply allow for an improved form of communication. At the other extreme, it might merge the minds into one in a process that can be thought of as a reverse split-brain operation. We propose that one way mind coalescence might happen is via an exocortex, a prosthetic extension of the biological brain which integrates with the brain as seamlessly as parts of the biological brain integrate with each other. An exocortex may also prove to be the easiest route for mind uploading, as a person's personality gradually moves away from the aging biological brain and onto the exocortex. Memories might also be copied and shared even without minds being permanently merged. Over time, the borders of personal identity may become loose or even unnecessary.

Keywords: Mind uploading; mind coalescence; whole brain emulation; exocortex.

1. Introduction

Mind uploads, or uploads for short (also known as brain uploads, whole brain emulations, emulations or ems) are hypothetical human minds that have been moved into a digital format and run as software programs on computers. One recent roadmap charting the technological requirements for creating uploads suggests that they may be feasible by mid-century [Sandberg & Bostrom, 2008].

There exists some previous work analyzing the societal consequences of uploading. Hanson [1994, 2008] has examined the economic consequences of being able to copy minds and Bostrom [2004] and Shulman [2010] consider some possible evolu-

*Permanent address: ZenRobotics Ltd., Mikonkatu 8 A, 00100 Helsinki, Finland

tionary scenarios. Sotala [this issue] discusses various advantages that any digital minds, including uploads and artificial intelligences, may benefit from. Sandberg & Bostrom [2008] briefly mention various ways by which neuroscience would benefit from uploading being possible.

Previous work has mostly assumed that while uploads may be copied, they will remain basically separate. Two uploads may share memories in the sense of being copies of each other, but they have no special means of sharing new information with each other.

A coalesced mind, or a coalescence for short, is a hypothetical mind created by merging two or more previously separate minds. Physical or software connections are created between the brains housing the minds, similar to the neuronal connections already existing within each brain. The brains begin communicating with each other directly, as if they were different parts of the same brain. Eventually, any stored information that one of the minds can consciously access becomes consciously accessible for the other minds as well.

In addition to information, brains house conscious thought processes. A normal human brain consists of two hemispheres that normally have only one conscious thought process between them. Coalesced minds could end up with either only one conscious thought process or several, depending on the implementation.

There could be varying degrees of coalescence, from full integration joining both information and thought processes, to a light integration where information is only occasionally exchanged. In any case, the minds involved could use each others' accumulated knowledge without needing to learn all of it themselves.

The purpose of this paper is to introduce the concept of mind coalescence as a plausible future development, and to study some of its possible consequences. We also discuss the exocortex brain prosthesis as a viable uploading path. While similar concepts have previously been presented in science fiction, there seems to have been little serious discussion about whether or not they are real possibilities. We seek to establish mind coalescence and exocortices as possible in principle, but we acknowledge that our argument glosses over many implementation details and empirical questions which need to be solved by experimental work. Attempting to address every possible problem and challenge would drown the reader in neuroscientific details, which we do not believe to be a productive approach at this stage. Regardless, we believe that compared to some uploading proposals, such as using nanoprobes for correlational mapping of neuronal activity [Strout, 2007], our proposal, while still speculative, seems like a much more feasible development to occur in the foreseeable future.

2. The Benefits of Coalescence

The reasons for wanting to coalesce with another mind may not be immediately obvious. Yet there are a number of reasons why somebody might want to do so, either at a light and superficial level, or more extensively.

Coalescing could help exchange information far more effectively than by simple speech. Instead of being limited to talking, thoughts and ideas could be integrated in a similar way that information is integrated between the two hemispheres of a human brain. Consider an economist who has coalesced with a physicist. To take full advantage of the physicist's knowledge for his own work, he needs to be able to draw on the parts of it that he does not himself know are useful. For example, there might be a connection between a phenomenon in economics and a phenomenon in physics that requires deep knowledge of both in order to be noticed. Neither the economist or physicist could notice the connection by themselves. They could notice it without coalescing if they first spent an extended amount of time explaining their knowledge of the phenomenon to the other, but then they would have to first know that such an explanation would be useful. With coalescing, such connections could be noticed quickly and with little effort.

Normal humans can only have a single conscious thought process at a time. The phenomenon of attentional blindness will serve as an example. A person who is focusing their attention on something will often fail to notice things that they are looking directly towards—someone looking for an empty spot in a movie theater may miss their friends waving right at them, or someone following a bouncing basketball may fail to notice a man in a gorilla suit walking through the game [Simons & Chabris, 1999]. A mind may wish to acquire the ability to have several simultaneous consciousnesses at the same time, letting it think and do many things at once. This might be possible by creating a copy of oneself and coalescing with it in the right way.

If two minds share a common desire or preference, merging together may help promote that preference. Groups of individuals working towards achieving a specific goal face the problem of free-riding and enforcing effective co-operation. If the individuals involved in the group have goals other than what the group is trying to achieve, each individual has an incentive to invest less than it could in the group, hoping that other members will accomplish their goal regardless. This is particularly the case in large groups, where the impact of a single individual is close to negligible [Olson, 1965]. Two or more minds coalescing could help ensure that their individual desires and their collective desires are the same.

More generally, coalescing may help ensure trust between two minds. Two minds that coalesce together could thereby assure each other that they both really are fully committed to their common goal. An added benefit would be the coordination of behavior. Humans acting in groups are often unaware of what the others are doing, and may duplicate work or fail to do needed work. Coalescing could allow for such information to be traded quickly and effectively.

Copying memories between two minds is a special case of coalescence, where one mind receives the memories and information another has, but retains its own distinctive thought processes. This could be used to learn things rapidly. A milder variant would be to lightly link together a teacher and a student, allowing for a more gradual transfer of information to the student.

Everyone might not have consciously held goals that they want to explicitly promote. As far as such people are interested in coalescence at all, they'll be more driven by urges such as curiosity and a desire to better understand another person's ways of thought. A person with a weak visualization ability might want to experience what a strong visualization ability feels like. One might also speculate that lovers or very close friends might choose to coalesce as an expression of their loyalty to each other. Parents may wish link to their children to better transmit their experience and values.

Even minds which are not interested in personally coalescing may be altruistically motivated to share their accumulated knowledge with others. They could allow parts of their knowledge to be copied and be made available for others to coalesce with.

3. Paths to Coalescence

Coalescence requires some technological means of connecting minds together. We consider three options: direct brain-to-brain connections, an exocortex-mediated connection, and an option based on doing a full upload first.

3.1. *Direct brain-to-brain connections*

The easiest approach seems to be to connect human brains directly in much the same way as the two brain hemispheres are connected. The corpus callosum, which connects the hemispheres, comprises 200–250 million axons crossing from one hemisphere to the other. It seems likely that to coalesce minds, the number of connections should be of a similar order of magnitude, probably at least millions.

The technology exists today for creating hundreds of connections: for instance, Hochberg *et al.* [2006] used a 96-microelectrode array which allowed a human to control devices and a robotic hand by thought alone. Cochlear implants generally stimulate the auditory nerve with 16–22 electrodes, and allow the many recipients to understand speech in every day environments without needing visual cues [Peterson *et al.*, 2010]. Various visual neuroprostheses are currently under development. Optic nerve stimulation has allowed subjects to recognize simple patterns and localize and discriminate objects. Retinal implants provide better results, but rely on existing residual cells in the retina [Ong & Crux, 2011]. Some cortical prostheses have also been recently implanted in subjects [Normann *et al.*, 2009]. We are still likely to be below the threshold of coalescing minds by several orders of magnitude. Nevertheless, the question is merely one of scaling up and improving current techniques.

In order to understand the effects of a direct brain-to-brain connection, it is useful to consider what happens if the axons connecting the hemispheres are severed. This results in a condition known as split brain: two different conscious minds, one for each hemisphere. Each mind has its own set of knowledge, preferences, attention and motor control [Gazzaniga, 2005], implying two parallel conscious minds. The

effect of connecting brains directly would likely resemble the reverse of splitting the brain: the minds would coalesce such that there would be one conscious process which would have access to the knowledge of both brains. This would create a single, unified mind. If the connections were then separated, there would again be two separate brains instead of a single unified one.

Separating the connections might recreate two minds with their individual attentional processes, similar to the two original ones that existed before the connection was created. Either of the two minds would then have no access to the knowledge of the other mind. All knowledge transfer would thus have to happen during the uni-attention state. Although this approach seems technically feasible in principle, it does not seem like the optimal approach.

3.2. *Connecting minds via exocortices*

It appears that the biological brain cannot support multiple separate conscious attentional processes in the same brain medium. Merging brains would there therefore probably lead to a mind with only one focus of conscious attention. To implement a multi-attention mind, some sort of a mediating component that allows for the presence of multiple conscious attentional processes is required.

An optimal coalescence would fuse the memories of the participating brains to a shared knowledge base, available for each individual attentional process. The details will depend on the architecture of the upload but, as long as the architecture resembles the distributed synaptic memory of the biological brain, what needs to be done is combine the synaptic changes resulting from each individual attentional process.

To achieve this, we propose to connect to the human brain an exocortex, a prosthetic extension of the biological brain which would integrate with the mind as seamlessly as parts of the biological brain integrate with each other. Once the exocortex had become a part of a person's brain, it could be connected to the exocortices of other people, allowing for coalescence to occur.

We presume that in addition to directly connecting biological brains together, the brain interface technology surveyed in the previous section may be used connect the brain to an exocortex device carried by the user. Furthermore, we make three assumptions which will be further fleshed out in the following sections:

- **There seems to be a relatively unified cortical algorithm which is capable of processing different types of information.** Most, if not all, of the information processing in the brain of any given individual is carried out using variations of this basic algorithm. Therefore we do not need to study hundreds of different types of cortical algorithms before we can create the first version of an exocortex.
- **We already have a fairly good understanding on how the cerebral cortex processes information and gives rise to the attentional processes**

underlying consciousness.[a] We have a good reason to believe that an exocortex would be compatible with the existing cortex and would integrate with the mind.

- **The cortical algorithm has an inbuilt ability to transfer information between cortical areas.** Connecting the brain with an exocortex would therefore allow the exocortex to gradually take over or at least become an interface for other exocortices.

In addition to allowing for mind coalescence, the exocortex could also provide a route for uploading human minds. It has been suggested that an upload can be created by copying the brain layer-by-layer [Moravec, 1988] or by cutting a brain into small slices and scanning them [Sandberg & Bostrom, 2008]. However, given our current technological status and understanding of the brain, we suggest that the exocortex might be a likely intermediate step. As an exocortex-equipped brain aged, degenerated and eventually died, an exocortex could take over its functions, until finally the original person existed purely in the exocortex and could be copied or moved to a different substrate. This is similar to one of the scenarios discussed by Moravec [1988].

Strictly speaking, an exocortex could act as merely an intermediate component that allows for mind coalescence, without necessarily leading to mind uploading. An exocortex alone would not be enough to replicate all the necessary functions of a brain: the various non-cortical regions would also need to be replaced. On the other hand, if a large part of a person's brain functions moved to the exocortex, he could be considered a partial upload even while many brain functions persisted in the biological brain.

3.2.1. *A general cortical algorithm*

An adult human neocortex[b] consists of several areas which are to varying degrees specialized to process different types of information. The functional specialization is correlated with the anatomical differences of different cortical areas. Although there are obvious differences between areas, most cortical areas share many functional and anatomical traits. There has been considerable debate on whether cortical microcircuits are diverse or canonical [Buxhoeveden & Casanova, 2002; Nelson, 2002] but we argue that the differences are variations of the same underlying cortical algorithm, rather than entirely different algorithms. This is because most cortical areas seem to have the capability of processing any type of information. The differences seem to be a matter of optimization to a specific type of information, rather than a different underlying principle.

[a]This relates to the mechanisms of consciousness, or what is called the "easy problem"; we make no claims about the so-called "hard problem of consciousness" [Chalmers, 1995].

[b]The neocortex is the part of the cortex that developed in mammals and has expanded dramatically in humans and other mammals with large brains.

The cortical areas do lose much of their plasticity during maturation[c]. For instance, it is possible to lose one's ability to see colors if a specific visual cortical area responsible for color vision is damaged. The adult brain is not plastic enough to compensate for this damage, as the relevant regions have already specialized to their tasks. If the same brain regions were to be damaged during early childhood, color blindness would most likely not result.

However, this lack of plasticity reflects learning and specialization during the lifespan of the brain rather than innate algorithmic differences between different cortical areas. Plenty of evidence supports the idea that the different cortical areas can process any spatiotemporal patterns. For instance, the cortical area which normally receives auditory information and develops into the auditory cortex will develop visual representations if the axons carrying auditory information are surgically replaced by axons carrying visual information from the eyes [Newton & Sur, 2004]. The experiments were carried out with young kittens, but a somewhat similar sensory substitution is seen even in adult humans: relaying visual information through a tactile display mounted on the tongue will result in visual perception [Vuillerme & Cuisiner, 2009]. What first feels like tickling in the tongue will start feeling like seeing. In other words, the experience of seeing is not in the visual cortex but in the structure of the incoming information.

Another example of the mammalian brain's ability to process any type of information is the development of trichromatic vision in mice that, like mammalian ancestors, normally have a dichromatic vision [Jacobs *et al.*, 2007]. All it takes for a mouse to develop primate-like color vision is the addition of a gene encoding the photopigment which evolved in primates. When mice are born with this extra gene, their cortex is able to adapt to the new source information from the retina and to make sense of it. Even the adult cortical areas of humans can be surprisingly adaptive as long as the changes happen slowly enough [Feuillet *et al.*, 2007]. Finally, Marzullo *et al.* [2010] demonstrated that rats implanted with electrodes both in their motor and visual cortices can learn to modulate the output from their motor cortex based on feedback given to visual cortex. This type of input-output device is interesting because it can be considered as a first step towards an exocortex which communicates with the cerebral cortex.

3.2.2. *The processes underlying attention and consciousness*

We have good reason to believe that an exocortex would not only be able to provide information to the cortex. If it were designed properly, it would also participate in creating a unified mind as seamlessly as the two hemispheres of our cortex do. It is already known from split-brain research that the lateral connections between the

[c]Our brain does lose much of its plasticity, preventing us, for instance, from becoming as fluent in new languages as native speakers. However, it might be possible to continuously augment the exocortex with new areas ready to specialize in new tasks, or to make sure that enough of the exocortex would remain plastic.

cortical regions play a critical role in creating a unified attention and consciousness. Severe those connections and you have two independent minds [Gazzaniga, 2005]. The key feature of our cortical algorithm behind attention and consciousness appears to be so called biased competition [Desimone & Duncan, 1995]. Each cortical area processes and represents information in its own primary inputs (e.g., from the sensory organs in primary sensory cortices; from primary sensory cortices in the case of secondary sensory cortices). The observed neural activation is primarily caused by activation of the primary input. However, in case there are several different interpretations or things to represent, lateral and top-down input will bias the competition between alternative representations. The representation which gains more support laterally wins. As cortical areas provide lateral input to each other, this local competition and global lateral transmission of information leads to an emergent, attentional process: at any time, different cortical areas tend to represent information about the same object or event. The biased competition model has been implemented in simulations and has successfully replicated many aspects of attention, including bottom-up phenomena and, for instance top-down search for objects [Deco & Rolls, 2004].

The information filtering implemented by biased competition model seems to be compatible with findings about conscious, subconscious and subliminal processes in the brain [Dehaene *et al.*, 2006]. If a sensory stimulus is too weak to produce significant activation, it will be subliminal. Thus, the subject will deny perceiving a stimulus, even though the activation has been recorded in the brain and the effects of the activation can be probed with clever tests. Such an activation has not been able to bias the representations of neighboring cortical areas. If the sensory stimulus is able to activate neighboring regions but not the whole cortical network—in particular, not the regions which are connected to the hippocampus and related areas responsible for long-term memory—the stimulus remains subconscious. If probed immediately afterwards, the subject is able to recall and report the stimulus, but otherwise the subject will forget ever observing the stimulus. Only sensory stimuli which trigger cortex-wide activation will be stored in long-term memory and can be recalled even after longer delays. Taken together, these findings indicate that an exocortex will integrate with the mind as long as it follows the same kind of rules of biased competition and has enough connections with the cortex so that either the cortex or exocortex can tip the balance of the other to represent the same information.

In the cortex, the meaning of the lateral input is learned [Martin, 2002]. This means that just connecting the brain with an exocortex will not immediately result in an integrated mind. Rather, both parties have to learn the associations through experience: each cortical area receiving lateral inputs needs to learn what kind of bottom-up inputs the lateral inputs predicts. It therefore takes time for the cortex and exocortex to grow together.

3.2.3. *Knowledge transfer in the cerebral cortex*

As far as we know, having an exocortex doing part of your thinking would not feel like anything in particular. People are not aware where their cognitive processes take place. They are not aware that a familiar memory has moved from one brain area to another. For example, it is known that long-term memories are first formed in the hippocampus but that they are gradually consolidated in the cerebral cortex [Lassalle *et al.*, 2000]. Also, several mammal species are capable of unihemispheric sleep where one of the cortical hemispheres is asleep while the other is awake [Rattenborg *et al.*, 2000]. The sleeping hemisphere will not remember the events that took place during their sleep before they have been awake together with the hemisphere which was awake during the event [Bloom & Hynd, 2005]. The cortical algorithm therefore seems to be prepared for moving memories around. This also relates to the fact that a gradual loss of parts of cortex can be compensated by the remaining cortex [Feuillet *et al.*, 2007]. From the perspective of an exocortical implant, this means that our minds could gradually be transferred to an exocortex as the original cortex ages and degenerates.

3.2.4. *Connecting exocortices*

It is likely that the initial integration of an exocortex and the brain will be a relatively slow learning process whose timescale is comparable to learning new skills or recovering from brain injuries, that is, at least days but more likely months or even years. However, once an exocortex has coalesced with the rest of the cortex, several new opportunities open up.

An exocortex, especially one that has taken over completely, could be used to allow several simultaneous conscious thought processes. Our brains are not able to do this because the neural activations corresponding to individual thought processes would interfere with each other, for instance, through local inhibition. With an exocortex, it would be possible to keep track of two or more sets of non-interfering neural activations and related processes with short timescales. As discussed earlier, the memory of each thought process would become available for all the processes through synaptic changes as the synaptic weights would be shared by all the processes.

Even before an exocortex had taken over completely, an interesting possibility would be to create a standard interface for exocortices. This would be analogous to natural languages which need to be learned first during encounters between people but once learned, can be used immediately by strangers meeting the first time. Two people, each with a standard interface on their exocortex, might link their exocortices together to be able to rapidly communicate with each other on a neural level. Such a high-bandwidth neural link could allow the communication of sensations such as tastes, sounds, images and complex episodes and even high-level abstractions much faster than symbolic communication using words because the neural patterns would contain so much information.

3.3. *Mind coalescence via full uploading*

The third possible way to achieve mind coalescence might be to first fully upload a human brain to a digital substrate somehow. Once this had been accomplished, the task of connecting two or more brains to each other would no longer be a problem of biological feasibility, but rather a more straightforward computer science problem. If the brains of Albert and Bob were both emulated in the same computer, then adding a connection between a neuron in Albert's brain and a neuron in Bob's brain might not have any essential difference from adding a connection between two neurons in Albert's brain. Full uploading could then be used to either implement a direct brain-to-brain connection, or to create a software exocortex to mediate the connection. However, we suspect that the technology for a physical exocortex will become available before the technology for full uploading will. As surveyed above, various brain prostheses are actively being researched, and their development will help in efforts to build an exocortex. Possibly one of the largest hurdles is the issue of creating connectors that are small enough to create millions of connections. This is likely to receive considerable amounts of funding regardless of whether anyone is interested in building exocortices as such.

We have so far only discussed connecting cortical areas, but it is clear that the brain is much more than just the cerebral cortex. However, much of what was said about understanding and being able to connect cortical regions also holds for the rest of the brain. In fact, many parts of the brain are better understood than the cerebral cortex. For instance, research of prostheses of hippocampus [Berger *et al.*, 2011] and cerebellum [Prückl *et al.*, 2011] is well under way.

Most of the approaches for a full uploading that are currently considered viable involve destructive uploading, i.e. cutting up the original brain to small slices and scanning them [Sandberg & Bostrom, 2008], which many people may feel uncomfortable with.

4. Barriers to Coalescence

So far, we have been presuming that the minds are willing and able to coalesce together. Yet the technical ease or difficulty of coalescence is only one of the factors that influences its adoption. There are numerous reasons for why people would not wish to coalesce, or employ exocortices.

4.1. *General integration difficulties*

This is a catch-all category for various technical problems that might crop up. Human brains did not evolve for the purpose of being easily merged, and the process may prove harder than anticipated. Errors and mistakes may prove hazardous to the subjects, and it is currently unknown what kind of a merging process is needed to ensure that the resulting mind will remain sane and functional. As noted in the introduction, we are intentionally glossing over most of the implementation details,

and much empirical work will be required before mind coalescence becomes a viable option.

Shulman [2010] notes that if uploads are willing to let themselves be deleted and experimented upon, many technical challenges can probably be overcome. Uploads can be copied and then modified and tested in various ways. Any failures can be deleted and computational resources reallocated to more successful copies, or for new attempts. Minds that have a flexible sense of identity, viewing the deletion of one copy as no different from a brief amnesia, might easily agree to such experimentation and deletion. If only some minds agree to such a treatment, the ones that do may gain a considerable advantage. Such experimentation is feasible if the upload's whole brain has been moved to a digital substrate, but less feasible if a considerable portion of the self still resides in a biological brain.

4.2. *Contrary preferences*

Two minds with a common goal may wish to coalesce in order to better pursue that goal. But if the minds think that they have incompatible or contradictory preferences or values, they might not want to merge together or to even share information with each other. Their preferences might be outright incompatible, as in the case of opposite ideologies or religions, or they might simply have few or no preferences in common.

Minds might also be mistaken about their true preferences, believing them compatible even when they aren't. Even if their core preferences were shared, coalescing may lead to unexpected interactions and give rise to entirely new preferences which cannot be anticipated in advance.

It's also possible for minds to only share preferences when observed at a rough level. For instance, two minds might share a preference for promoting communism. However, after they've coalesced and studied communism more, it becomes obvious that one of them wishes to promote Leninism and the other Maoism. Alternatively, one of them may become convinced that communism doesn't work as a system, while the other wants to stick to it. At such a point, the minds may wish to split.

A longer and deeper integration of minds may reduce the risk of contradictory preferences. On the other hand, even individual humans already exhibit plenty of contradicting preferences which they often have difficulty choosing between.

4.3. *Preferences opposed to coalescence*

There are some preferences which do not allow for coalescence, even if they were shared. For instance, two minds may both have a preference for preserving their personal identities. If they think that coalescing would involve their personal identity becoming lost, they will refuse to do so.

On the other hand, having this preference mightl not exclude the possibility of copying the mind and having those copies coalesce with another. If the preference is based in a strong sense of self-preservation, it may be likely that the copies will

also refuse to coalesce, but a mind could conceivably also have a more abstract preference for at least one copy of the original identity surviving. In such a case, the copy might have much less of an issue in coalescing, since another mind would still preserve its identity. An alternative route would be to only copy a part of the mind, if it was possible to do in such a way that its preferences did not carry over.

4.4. *Privacy-related preferences*

With coalescence, all of a person's thoughts and memories will potentially become available to another person. People may have thoughts or ideas that they want to keep secret, either because they are ashamed of them, or because they want them to be private out of principle. A person may also be unwilling to coalesce if he has been trusted with the secrets or private information of other people. Even if he was indifferent about his own secrets becoming available to the other person, he may not want to betray the trust that others have placed in him.

4.5. *Lack of mutual trust and memetic hazards*

It is not sufficient for two minds to both claim that they share preferences. They must also believe both that the other is honest, and that it is not mistaken. In an upload environment, purposefully flawed minds may be created with the express purpose of having them join a specific, enemy coalescence. The flaw may be a mismatch in the preferences of the mind and the target coalescence, incorrect information, or some issue causing problems in integrating the mind to the target.

If a particular piece of information is sufficiently valuable, the mind that considers sharing it also needs to consider the likelihood that the mind it's dealing with treats it with due care. Similar situations arise today. In considering whether to reveal someone state secrets, one needs to consider whether the recipient might accidentally slip the information to someone else, and whether the recipient is in danger of being kidnapped and tortured by the enemy. Such considerations become even more paramount when potential enemies have the capability to reliably extract all the knowledge that a mind has.

4.6. *Legal and ethical barriers*

The concept of mind coalescence creates new kinds of ethical questions. In the following discussion, we limit ourselves to the ethical questions that relate specifically to coalescence. Uploading in general also poses a number of important questions, such as whether the act of copying minds should be regulated or restricted [Bostrom, 2004; Hughes, 2004; Hanson *et al.*, 2007; Bostrom & Yudkowsky, 2011], or whether the ability to copy minds might make a state more willing to use weapons of mass destruction [Shulman, 2010]. We recognize that these are important questions, but do not discuss them here.

Possible ethical dilemmas involving mind coalescence include:

- If two minds coalesce fully and no longer exist as separate minds, did the pre-coalesced minds die to give birth to a new one? Is a forced coalescence, where one or more of the minds doesn't consent to the process, equivalent to murder?
- Under what conditions may a mind consent to coalescing with another?
- If two minds coalesce, is the resulting mind bound by contracts that either of the pre-coalesced minds had previously agreed to? What if a coalesced mind signs a contract and then splits back into two minds? Similar problems already exist with dissociative identity disorder, where it is not entirely clear whether contracts entered into by one personality should bind the others [Merckelbach *et al.*, 2002].
- If minds that are willing to coalesce receive a disproportionate advantage in, for example, the labor market, should coalescence be restricted so as to not create a pressure for everyone to coalesce?
- Might coalescence lead to extreme kinds of evolutionary scenarios (see the next section) and if so, are they desirable or should we try to avoid them?
- Does coalescing have military value—for example, by improving espionage efforts or battlefield coordination, and could it lead to arms races?
- Does the possibility for coalescence threaten some more intangible value, such as "human dignity" [Fukuyama, 2002; Bostrom, 2005]?

Many possible answers could be given to the above questions, and it is conceivable that legislatures might decide to restrict or even ban coalescence. On the other hand, factors such as possible military or economic value may make restrictions less desirable.

It remains unclear to what extent legal regulation, even if enacted, will be effective at preventing coalescences. If the advantages of coalescence are great, and policing it is hard, laws aimed against it may have little practical effect.

5. Evolutionary Pressures and Fluid identities

When information and memories can be freely shared and minds joined together, the boundaries of identity will necessarily fade. Some versions of the "psychological continuity" view of personal identity suggest that the coalescence of Albert and Bob ("Albob") now really *is* both Albert and Bob. Some of these theories may require that the coalescence preserves the original thought processes, while for others it is enough that the coalescence has the memories of both Albert and Bob. Other versions of the psychological continuity view hold that Albert is the same person as a future or past person, *only* if he is psychologically continuous with that person, and no other person is. In other words, Alfred can be the same person as Albob, or Bob can be the same person as Albob, but they both cannot be. Such views might consider Albob to be an entirely new person, created by combining two previous persons who no longer exist [Olson, 2010].

Regardless of which philosophical view one adopts, Albob may very well feel like both Albert and Bob, as he remembers doing both the things that Albert has done and the things that Bob has done. If he merges with more minds, his feeling

of having a unique identity may fade even further.

The situation of a coalescence with memories from many different individuals may be likened to that of bacteria. All bacteria can be thought of having a single core preference: to survive and replicate. To this end, they commonly engage in horizontal gene transfer, exchanging genes between each other [Ochman *et al.*, 2000]. For bacteria, any genetic information which may be used to further their preference of reproduction is readily accepted. It is not necessary, and could even be harmful, to preserve a strict sense of "genetic identity." Bacteria have evolved to exchange genes with each other because it helps them survive and replicate.

Bacteria only have a genetic level of information. Humans also have a knowledge-based, or memetic [Dawkins, 1976; Blackmore, 2000] level of information. Coalescences may view the trading of memories and other memetic information the same way that bacteria "view" the trading of genetic information. Attempts to strongly preserve a strict sense of personal identity by restricting the memetic trades that are entered into may hinder promoting the coalescence's preferences. A coalescence may seek to absorb all the memories and memetic information that it may have even a potential use for.

Among uploads, preferences such as ones relating to the preservation of personal identity may be much more strongly subject to evolutionary pressures than in humans. Biological evolution is based on differential reproduction: genes which increase the amount of surviving offspring will increase in relative frequency. In developed countries, the amount of surviving offspring is more influenced by an individual's desire to have many offspring than on his talents in any particular field. While particular preferences or other traits may give individual humans an advantage in a particular field, this doesn't usually convey an advantage in evolutionary terms. (A top-earning lawyer might choose to have no children while somebody working the minimum wage might have three.) Even if it did, most preferences are less than perfectly heritable and might not be adopted by the individual's offspring, so a preference increasing one's earning potential might not convey an evolutionary advantage to one's offspring.

In contrast, an upload's ability to replicate may be directly proportional on its ability to earn wages. A wealthy upload can choose to rent or buy the computing power needed to copy itself. Alternatively, an employer might offer the upload money in exchange for the right to make a copy of it and employ the copy [Hanson, 1994]. While the upload's willingness to create copies of itself is still a major factor, a copy of an adult mind does not need constant attention and care the way a child does. If someone else offers an upload money in exchange for the right to make a copy, the upload will agree to the deal as long as it is indifferent to the prospect of being copied, or finds the compensation adequate. This might lead to a small number of exceptionally productive or otherwise skilled uploads being copied in large numbers. Hanson [1994] speculates on a scenario where a large fraction of the employers in a major profession, such as contract law, might be copies of a single upload.

Even small differences in traits such as skill or the willingness to be copied could

translate into a vast reproductive advantage. If the best worker in a field was one-tenth of a percentage more effective than the second best worker, and was willing to let himself be copied an arbitrary amount of times, then all the employers might choose to copy him instead of the second best worker. Depending on the size of the profession, the one-tenth of a percentage difference in skill could then translate into a multimillionfold advantage in differential reproduction, much greater than anything seen in current human evolution.

As a result, the creation of uploads may result in very rapid selection that strongly favors particular kinds of uploads. If minds that are indifferent to preserving their own personal identity get a competitive advantage as a result of being willing to coalesce with other minds, such "coalescence-ready" uploads may end up dominating overall. The end result could be that a large fraction of humanity would be indifferent to preserving their personal identities, and the borders of their personalities would thus grow increasingly weak.

In addition to full-scale coalescence, it might be possible to merely transfer knowledge between individuals. This could be done by copying the contents of the exocortex or part of it, or if a mind exists in a purely digital form, by copying a part of the brain. This might allow for the transfer of memories and skills, without also bringing over all the desires and preferences. Standard interface exocortices, as discussed above, might also make it easier to only share information.

To the extent that this is feasible, some individuals may choose to make their memories or parts of the memories, freely available for others. An educated or otherwise experienced person may wish to altruistically make his knowledge freely available for others to learn from. Another may desire to promote his ideology by more effectively communicating his reasons for believing in it. Yet another may hold a memory-centered view on personal identity, and seek immortality by spreading his memories. Whatever the reason, various repositories may come to hold the memories and knowledge of many different minds. If such repositories contain specialized skills or otherwise useful information, absorbing some of their information may become commonplace. This will lead to the borders of personal identity fading further, if a large fraction of minds carry with them many memories from entirely different minds.

At one extreme, things might proceed to a point where it is no longer meaningful to talk about individuals, or even specific coalescences at all. By trade or theft, several coalescences have gotten to the point where they all share almost all knowledge and memories that can be shared. What differentiates them is not their knowledge, but their core preferences and ultimate goals. To the extent to which this is plausible depends on the degree to which knowledge and memories can be transferred without also transferring preferences.

In such a situation, it might be most meaningful to talk about the assets and positions of the core preferences themselves, rather than of the minds acting as carriers for the preferences. While current-day humans could also be viewed as mere carriers for memes and preferences, today it is still clearly meaningful to talk

about specific humans with their unique personalities and memories, whereas in a highly-coalesced society it might not be.

One might even go as far as to suggest that in the long run, it is the details of the core preferences themselves that determine which ones win out at the end. General, open-ended preferences will fare better than narrow, specific ones. A coalescence with no ultimate desire other than to replicate as much as possible will have a definite advantage over one which has its options constrained by wanting to protect a specific landmark. All other things being equal, this might be the deciding factor that tips the balance in the favor of the replicator.

On the other hand, minds with relatively easily achievable preferences may find it beneficial to merge even if they would previously share no preferences. The resulting mind has all the knowledge and preferences of both original minds. This can be viewed as a trade or an alliance, where both minds agree to further the other's core preferences in exchange for having their own core preferences likewise furthered. The difference to an actual alliance is that once the minds have coalesced, there is no risk of either party betraying the deal.

Specific preferences may then be at an advantage or disadvantage, depending on how easy they are to implement without contradicting other preferences and how likely it therefore is that they will be accepted in exchange by others. For example, a mind wishing to protect the economic interests of Finland may agree to coalesce with a mind wishing to protect the economic interests of Sweden, as the countries are geographic and cultural neighbors and furthering both goals at the same time is relatively easy. Coalescing with a mind promoting the interests of a geographically and culturally distant country, such as Uganda, may require more consideration.

6. Conclusions

We have argued that exocortices are a plausible development in the foreseeable future. We have outlined three assumptions which, if true, would seem to make the development of exocortices technically feasible. First, that there is a unified cortical algorithm which can be replicated with relative ease, without needing to study many different cortical algorithms in detail. Second, that consciousness works according to a biased competition model, and that it is possible to integrate the exocortex to that process. Third, that the ability to move information across different areas is an inherent property of the cortical algorithm. This can be utilized to transfer knowledge from the existing brain to the exocortex. If all three of these assumptions hold, and the other technical challenges are overcome, exocortices could then be connected together in order to coalesce minds.

While it could be possible for minds to coalesce even without exocortices, it would seem more desirable to have exocortices as mediating components. Although a great deal of work remains to be done before exocortices can be developed, none of the challenges seem unsurmountable in principle. Various neural prostheses that act as artificial senses are already in use, while prostheses for motor outputs, mem-

ory, and cerebellar function are under active development. As such prostheses have obvious medical uses, their development is likely to remain funded in the future. Exocortices, while arguably a more radical intervention, are a plausible outgrowth of the current work. As they are prostheses that can be used to alleviate the damage caused to an aging cortex, they might be developed even without anyone having a particular interest in mind uploading or mind coalescence.

Should coalescing turn out to be possible, there are a number of reasons for why minds might want to do so. Most importantly, coalescing allows for better co-operation and sharing of information. Even if minds did not choose to explicitly merge together, the possibility to copy and transfer memories might turn out to be highly useful. People might also choose to coalesce with others for the sheer experience of it, and possibly to achieve an inner richness which would be impossible without combining the life experiences of several very different individuals. In addition to the possibilities we have already discussed, a very speculative possibility would be to partially coalesce with an animal such as an octopus, in order to get a taste of a very alien way of thought.

On the practical side of coalescing minds, many questions remain, including but not limited to the following ones:

- How can connections necessary for building an exocortex be made as small as is necessary?
- How easily, and to what extent, can knowledge be moved from an existing cortex to a fresh exocortex?
- How long will it take for an exocortex to integrate with an existing brain?
- How can the processes of exocortex integration and mind coalescence be made safe?
- Is it possible for two minds to split after coalescing together? Is there some stage after which splitting becomes impossible without causing serious damage?
- Is it possible to transfer memories and knowledge without also transferring things such as values, preferences and goals?
- Is the basic cortical algorithm really the same everywhere?
- To what extent are standard interface exocortices possible?
- To what extent is it possible to safely coalesce with non-human animals?
- To what extent is it possible for a person to share a skill such as a strong visualization ability with someone with a weaker ability, without losing their own skill?
- What is needed for the same mind to be able to possess several conscious thought processes at once?
- How should the process of coalescence be treated from a legal perspective?
- In the long run, are the outcomes of mind coalescence being possible actually desirable?

In general, the ability to copy minds seems to lead to very strong evolutionary pressures. While various reasons exist why not everyone might want to coalesce,

or why legislatures might want to restrict coalescence, it also seems possible that coalesced minds could quickly outcompete uncoalesced minds. The possibility to share memories and information would weaken the borders of individuality even further. In the long run, the sense of identity of a great deal of people might become considerably more fluid than it is today.

Acknowledgments

The authors would like to thank Louie Helm, Peter Scheyer, Vladimir Nesov, several pseudonymous commenters on `http://lesswrong.com/`, as well as two anonymous reviewers, for their feedback on the article.

References

Berger, T. W., Hampson, R. E., Song, D., Goonawardena, A., Marmarelis, V. Z. and Deadwyler, S. A. [2011] A cortical neural prosthesis for restoring and enhancing memory, *J. Neural Eng.* **8**(4).

Blackmore, S. [2000] *The Meme Machine* (Oxford University Press).

Bloom, J. S. and Hynd, G. W. [2005]. The role of the corpus callosum in interhemispheric transfer of information: excitation or inhibition? *Neuropsychology Rev.* **15**(2), 59–71.

Bostrom, N. [2004] "The future of human evolution," in Tandy, C. (ed.) *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing* (Ria University Press), pp. 339–371.

Bostrom, N. [2005] In defense of posthuman dignity, *Bioethics* **19**(3), 202–214.

Bostrom, N. and Yudkowsky, E. [2011] "The ethics of artificial intelligence," in Ramsey, W. and Frankish, K. (eds.) *Cambridge Handbook of Artificial Intelligence* (Cambridge University Press).

Buxhoeveden, D. P. and Casanova, M. F. [2002] The minicolumn hypothesis in neuroscience, *Brain* **125**(5), 935–951.

Chalmers, D. J. [1995] Facing up to the problem of consciousness, *J. Consc. Studies* **2**(3), 200–219.

Dawkins, R. [1976] *The Selfish Gene* (Oxford University Press).

Deco, G. and Rolls, E. T. [2004] A neurodynamical cortical model of visual attention and invariant object recognition, *Vision Research* **44**(6), 621–642.

Dehaene, S. Changeux, J.-P. Naccache, L. Sackur, J. and Sergent, C. [2006] Conscious, preconscious, and subliminal processing: a testable taxonomy, *Trends in Cognitive Sciences* **10**(5), 204–211.

Desimone, R. and Duncan, J. [1995] Neural mechanisms of selective visual attention. *Ann. Rev. of Neurosci.* **18**, 193–222.

Feuillet, L., Dufour, H. and Pelletier, J. [2007] Brain of a white-collar worker, *The Lancet* **307**(9583), 262.

Fukuyama, F. [2002] *Our posthuman future* (Farrar, Strauss and Giroux, New York).

Gazzaniga, M. S. [2005] Forty-five years of split-brain research and still going strong, *Nature Reviews Neurosci.* **6**(8), 653–659.

Hanson, R. [1994] If uploads come first, *Extropy* **6**(2), `http://hanson.gmu.edu/uploads.html`.

Hanson, R. [2008] Economics of the singularity, *IEEE Spectrum*, 37–42, June 2008.

Hanson, R., Hughes, J., LaTorra, M., Brin, D. and Prisco, G. [2007] The Hanson-Hughes debate on "The crack of a future dawn," *J. of Evolution and Technology* **16**(1).

Hochberg, L. R., Serruya, M. D., Friehs, G. M., Mukand, J. A., Saleh, M., Caplan, A. H., Branner, A., Chen, D., Penn, R. D. and Donoghue, J. P. [2006] Neuronal ensemble control of prosthetic devices by a human with tetraplegia, *Nature* **442**, 164–171.

Hughes, J. [2004] *Citizen Cyborg: Why Democratic Societies Must Respond to the Redesigned Human of the Future* (Westview, Cambridge MA).

Jacobs, G. H., Williams, G. A., Cahill, H. and Nathans, J. [2007] Emergence of novel color vision in mice engineered to express a human cone photopigment, *Science* **315**(5819), 1723–1725.

Lassalle, J-M., Bataille, T. and Halley, H. [2000] Reversible inactivation of the hippocampal mossy fiber synapses in mice impairs spatial learning, but neither consolidation nor memory retrieval, in the Morris navigation task, *Neurobiology of Learning and Memory* **73**(3), 243–257.

Martin, K. A. C. [2002] Microcircuits in visual cortex, *Current Opinion in Neurobiology* **12**(4), 418–425.

Marzullo, T. C., Lehmkuhle, M. J., Gage, G. J. and Kipke, D. R. [2010] Development of closed-loop neural interface technology in a rat model: combining motor cortex operant conditioning with visual cortex microstimulation, *IEEE Transactions on Neural and Rehabilitation Systems Engineering* **18**(2), 117–126.

Merckelbach, H., Devilly, G. J. and Rassin, E. [2002] Alters in dissociative identity disorder: Metaphors or genuine entities? *Clinical Psychology Review* **22**, 481–497.

Moravec, H. [1988] *Mind Children: The Future of Robot and Human Intelligence* (Harvard University Press).

Nelson, S. B. [2002] Cortical microcircuits: diverse or canonical? *Neuron* **36**(1), 19–27.

Newton, J. R. and Sur, M. [2004] "Plasticity of cerebral cortex in development," in Adelman, G. and Smith, B. H. (eds.), *Encyclopedia of Neuroscience* (Elsevier, New York).

Normann, R. A., Greger, B. A., House, P., Romero, S. F., Pelayo, F. and Fernandez, E. [2009] Toward the development of a cortically based visual neuroprosthesis, *J. Neural Eng.* **6**(3).

Ochman, H., Lawrence, J. G., Groisman, E. A. [2000] Lateral gene transfer and the nature of bacterial innovation, *Nature* **405**, 299–304.

Olson, M. [1965] *The Logic of Collective Action: Public Goods and the Theory of*

Groups (Harvard University Press, Cambridge MA).

Olson, E. T. [2010] "Personal identity," in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy (Winter 2010 Edition)*, `http://plato.stanford.edu/archives/win2010/entries/identity-personal/`.

Ong, J. M. and da Cruz, L. [2011] The bionic eye: a review, *Clinical & Experimental Ophthalmology*, doi:`10.1111/j.1442-9071.2011.02590.x`.

Peterson, N. R., Pisoni, D. B. and Miyamoto, R. T. [2010] Cochlear implants and spoken language processing abilities: review and assessment of the literature, *Restor. Neurol. Neurosci.* **28**(2), 237–250.

Prückl, R., Grünbacher, E., Ortner, R., Taub, A. H., Hogri, R., Magal, A., Segalis, E., Zreik, M., Nossenson, N., Herreros, I., Giovannucci, A., Ofek Almog, R., Bamford, S., Marcus-Kalish, M., Shacham, Y., Verschure, P. F. M. J., Messer, H., Mintz, M., Scharinger, J., Silmon, A. and Guger, G. [2010] "The application of a real-time rapid-prototyping environment for the behavioral rehabilitation of a lost function in rats," in *Proc. IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind and Brain (CCMB 2011)* (Paris, France), pp. 1–8, doi:`10.1109/CCMB.2011.5952121`.

Rattenborg, N. C., Amlaner, C. J. and Lima, S. L. [2000]. Behavioral, neurophysiological and evolutionary perspectives on unihemispheric sleep, *Neurosci. Biobehav. Rev.* **24**(8), 817–842.

Sandberg, A. and Bostrom, N. [2008] Whole brain emulation: a roadmap, Technical Report #2008-3, Future of Humanity Institute, Oxford University. `http://www.fhi.ox.ac.uk/reports/2008-3.pdf`.

Shulman, C. [2010] Whole Brain Emulation and the Evolution of Superorganisms, `http://singinst.org/upload/WBE-superorganisms.pdf`.

Simons, D. J. and Chabris, C. F. [1999] Gorillas in our midst: sustained inattentional blindness for dynamic events, *Perception* **28**(9), 1059–1074.

Sotala, K. [2012] Advantages of artificial intelligences, uploads, digital minds, *Int. J. of Machine Consciousness.*

Strout, J. [2007] Mind uploading home page, `http://www.ibiblio.org/jstrout/uploading/MUHomePage.html`.

Vuillerme, N. and Cuisiner, R. [2009] Sensory supplementation through tongue electrotactile stimulation to preserve head stabilization in space in the absence of vision, *Invest. Ophtamol. Vis. Sci.* **50**(1), 476–481.