IFILNOVA PHILOSOPHY INSTITUTE FCSH/NOVA

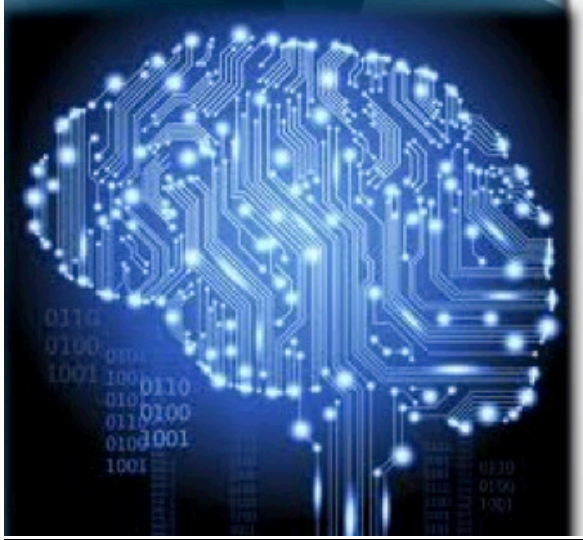
FCSH FACULDADE DE CIÊNCIAS SOCIAIS E HUMANAS UNIVERSIDADE NOVA DE LISBOA

FUNDAÇÃO LUSO-AMERICANA PARA O DESENVOLVIMENTO

Minds, Selves and 21st Century Technology in Lisbon

Featuring the work of
Susan Schneider

Confirmed Guest Speakers
Gerald Vision, Gualtiero Piccinini, Keith Frankish, Mark Bickhard, Ron Chrisley



23rd and 24th June, 2016
Faculdade de Ciências Sociais e Humanas
Universidade Nova de Lisboa
Portugal

Organizing Committee
Robert W. Clowes, Klaus Gaertner & Inês Hipólito

www.mindandcognition.weebly.com

Susan Schneider (University of Connecticut)
The Mind is not the Software of the Brain (Even if the Brain is Computational)
susansdr@gmail.com

Cognitive scientists and philosophers of mind often say that "the mind is the software of the brain" and relatedly, that "the mind is the program the brain runs". Is the mind really a program? The task of this talk is to dismantle the software model of the mind and a related view that says that we can survive brain uploading and other forms of radical brain enhancement because our "informational pattern" or "software" survives. Computationalism needs a richer conception of the mind. Drawing from my critiques of physicalism, I offer a non-physicalist, monistic approach to the mind as an alternative to the software conception, and illustrate how it is compatible with computationalism.

Gerald Vision (Temple University)
Consciousness – Philosophy's Great White Whale
gvision@temple.edu

I start from the plausible supposition that we are sentient creatures, that we undergo conscious episodes that are *something-they-are-like*. Nowadays two broadly competing overviews for explaining how such conscious aspects come into our largely nonconscious world have galvanized philosophers. On one view conscious aspects arise at an advanced stage in the course of the development of the physical world. Call that emergentism. That much is shared by the commonest form of dualism now on sale and its mainstream materialist rivals. On an opposing view consciousness is an original feature of the world, not generated by anything else or issuing from anything more basic. This approach is integral to an account generally known as Russellian monism. On it conscious aspects are the whole or part of the original source of the knowable physical and mental world. Although Russellian monism is not committed to panpsychism *per se*, a panpsychist version has been a natural development. (For these limited purposes I am ignoring Cartesian dualism, the original side-by-side co-existence of distinct substances that may or may not interact, as well as degrees of eliminativisms that reject outright or demote conscious episodes.) I shall examine both panpsychism's grounds for rejecting emergentism and the articles of ipan[sychism's replacement view: the former in a very sketchy way, the latter in a bit more detail. After arguing that the attacks on emergence and the replacement theory both fail, I shall make some brief remarks aimed at reconsidering emergentism. They will not constitute an adequate support for the view, but I hope they indicate why its promissory note isn't patently worthless.

Mark Bickhard (Lehigh University)
Emergent Mental Phenomena
mhb0@lehigh.edu

The possibilities of 'artificial' mental phenomena, including consciousness, depends on what the metaphysical nature of such phenomena are. I will outline a model of metaphysical emergence, and, based on that, emergent mental phenomena, with a focus here on cognition and consciousness.
This model suggests that 'artificial' mental phenomena are possible, though not with current technology. Furthermore, such 'artificial' mind phenomena would require, in effect, the creation of artificial life, at least in a metabolic sense.

George Theiner (Villanova University)
Extended Minds en route to Transhuman Subjects: The Bostrom-Clark-Fuller Express
georg.theiner@villanova.edu

Keith Frankish (University of Crete) TBA
A Dual-minds Perspective on Cognitive Enhancement and AI
k.frankish@gmail.com

In the last two decades, dual-system theories of cognition have become increasingly popular. Such theories propose that humans have, in a sense, two minds -- an evolutionarily old mind, which is largely nonconscious and devoted mainly to online behavioural guidance, and a recent, distinctively human mind, which is conscious and supports reflective and hypothetical thinking. This dual-system view raises important questions for debates about cognitive enhancement and artificial intelligence. For example, will enhanced human minds retain this duality, and to what extent will artificial minds replicate it? This talk will sketch a framework for thinking about human mental duality and highlight some of its implications for work on enhanced and artificial mentality.

Gualtiero Piccinini (University of Missouri - St. Louis)
The Myth of Mind Uploading
piccininig@umsl.edu

It has become fashionable to maintain that in the not too distant future we will be able to upload our minds into artificial computers, and that this mind uploading will make us immortal. This sort of mind uploading presupposes three strong assumptions: (2) that we will soon have the ability to construct a realistic computational simulation of an entire human brain; (1) that a realistic computational simulation of an entire human brain would have a conscious mind like the one possessed by the brain being simulated (computational functionalism); (3) that the mind putatively created by building a realistic computational simulation of an entire human brain is numerically identical to the mind possessed by the brain being simulated. I will argue that each of the first two assumptions are implausible and the third is false.

Ron Chrisley (Sussex *University)*
Machine Consciousness: Moving beyond "Is it Possible?
R.L.Chrisley@sussex.ac.uk

Philosophical contributions to the field of machine consciousness have been preoccupied with questions such as: Could a machine be conscious? Could a computer be conscious solely by virtue of running the right program? How would we know if we achieved machine consciousness? etc. I propose that this preoccupation constitutes a dereliction of philosophical duty. Philosophers do better at helping solve conceptual problems in machine consciousness (and do better at exploiting insights from machine consciousness to help solve conceptual problems in consciousness studies in general) once they replace those general questions, as fascinating as they are, with ones that a) reflect a broader understanding of what machine consciousness is or could be; and b) are better grounded in empirical machine consciousness research.

Richard Heersmink (Macquarie University)
The distributed self
richard.heersmink@gmail.com

This paper explores the implications of extended and distributed cognition theory for our notions of personal identity. On an extended and distributed approach to cognition, external information is under

certain conditions constitutive of memory. On a narrative approach to personal identity, autobiographical memory is constitutive of our diachronic self. In this paper, I bring these two approaches together and argue that external information can be constitutive of one's autobiographical memory and thus also of one's diachronic self. To develop this claim, I draw on recent empirical work in human-computer interaction, looking at lifelogging technologies in both healthcare and everyday contexts. These are memory technologies that record personal experiences. A key example is SenseCam, a small wide-angle camera worn around one's neck, taking a picture with a certain interval or when its sensor detects some environmental change. These pictures are then edited into a visual lifelog with a certain narrative structure, transforming, aiding, and in some cases constituting one's autobiographical narrative. I argue that personal identity can neither be reduced to psychological structures instantiated by the brain nor by biological structures instantiated by the organism, but should be seen as an environmentally-distributed and relational construct. In other words, the complex web of cognitive relations we develop and maintain with other people and technological artifacts partly determines our self. This view has conceptual, methodological, and normative implications: we should broaden our concepts of the self as to include social and artifactual structures, focus on external memory systems in the (empirical) study of personal identity, and not interfere with people's distributed minds and selves.

Lukas Schwengerer (University of Edinburgh)
Knowing yourself by looking at a smartphone
L.Schwengerer@sms.ed.ac.uk

The extended mind thesis presents us with a new and surprising challenge in our thoughts on self-knowledge. Consider the classical case of the Alzheimer's patient Otto who uses his notebook to enhance his memory. Clark and Chalmers (1998) argue that Otto has a belief that extends to the notebook. Beliefs are not merely in the head, but rather spread out over tools that we interact with in specific ways. If this is correct, then it seems as if Otto can know his own belief by looking at the notebook. If one asks Otto whether he believes that the museum is at 53rd street, he can look at his notebook and tell that he does.

If this description is correct, then we have to ask what kind of self-knowledge he acquires by looking at the notebook. Is this genuine self-knowledge or is it rather a special case of mind-reading? On one hand it does not look like genuine self-knowledge, because it does not come with the required asymmetry between first- and third-person. Anyone else can take a look at the notebook and thereby come to know what Otto believes, but self-knowledge intuitively comes with privileged access. On the other hand, it does not seem to be usual mind-reading either, because that seems to involve more interpretation than is going on in Otto's case. In a standard mind-reading case one observes behavior and then infers that a belief that p explains that behavior. Otto on the other hand just looks at the notebook. Now one can accept that Otto makes an inference insofar as generally Otto behaves as if what is written in the notebook is true. But on this level of explanation there is nothing to be found that identifies the proposition in question. Otto knows that he believes that p because he generally acts in accord with the notebook and the notebook states that p. So part of what lets Otto infer that he believes that p is the fact that p is written in the notebook. However, the notebook is considered part of the extended belief, if the extended mind thesis is true. So it seems that the attempt to explain Otto's self-knowledge as mind-reading is committed to inferring the belief that p from part of the belief that p itself. But that is not mind-reading anymore, because the basis of the inference is part of the same thing that is inferred. This looks like a detectivist story!

Extended self-knowledge has an even more peculiar nature. Extended self-knowledge is accompanied by a surprising outcome for the truth of self-ascriptions. Otto not only self-ascribes his belief that the museum is at 53rd street in virtue of the notebook, the very same thing also makes it the case that he believes the museum is at 53rd street. If Otto self-ascribes his belief by looking at the

notebook and the conditions for extended belief are satisfied, he cannot fail to have the belief in question. Extended self-knowledge seems to be infallible.

Moreover, this leads to an interesting normative conclusion. If extended self-knowledge is infallible in this sense and we ought to have as many true beliefs and as few false beliefs as possible, then it seems to be the best option to externalize as many of our mental states as possible. We ought to offload our memory to our smartphones.

This is contrasted by an obvious shortcoming of extended self-knowledge. The cost of a worst-case scenario is too high. Consider Alice, an Alzheimer's patient who uses her smartphone as an extended memory. Suppose she loses the smartphone at some point. The more of her beliefs were extended, the more of herself she loses.

The upshot is that extended beliefs are superior for self-knowledge, but only if we can find a way to keep the extended beliefs safely.

Robert Clowes (Nova University of Lisbon)

Progressive Uploading: Super-selves and alternative routes to personal immortality

robert.clowes@gmail.com

There are problems with the idea of uploading, not least that the model of human beings (or selves, or persons) as computer programmes is a deeply problematic one. Indeed, it may not be possible to upload human beings in anything like the standard sense (Corabi & Schneider, 2014).

And yet, some scrutiny of certain contemporary mass uses of technology indicates we may already be involved in a different (potentially non-destructive) form of uploading. This is the growing tendency for those in the technologically developed world to store ever more personal information and about their lives as digital memory traces or lifelogs. Moreover these traces are already used not just a passive record but to regulate their ongoing cognitive and emotional lives (Clowes, 2015).

Extending our minds (and sense of self) into the Internet does not leave us quite as we were (Clowes, 2012). Such technologies do not provide merely a neutral record of one's activities but can be viewed themselves as cognitive enhancements (or diminishments) that already have important implications for questions of self and personal identity (Clowes, 2013). Even, such gradual or progressive uploading may have important cognitive consequences.

Some researchers are currently pushing this tendency to its limits. Gordon Bell and his associates for example have attempted to use technologies such as the SenseCam to make a 'total' record of the everyday life of individuals (Bell & Gemmell, 2009). Bell has suggested that this information might be used by future artificial intelligence technology to create a version of himself that might communicate with his descendants. I will attempt to assess some of the implications for how we should assess the metaphysical status of such future 'slow uploads' particularly paying attention to how human beings already have deeply hybrid minds (Clark, 2003)

References

Bell, C., & Gemmell, J. (2009). Total recall: how the E-memory revolution will change everything: Dutton.

Clark, A. (2003). Natural Born Cyborgs: Minds, Technologies and the Future of Human Intelligence. New York: Oxford University Press.

Clowes, R. W. (2012). Hybrid Memory, Cognitive Technology and Self. In Y. Erdin & M. Bishop (Eds.), Proceedings of AISB/IACAP World Congress 2012.

Clowes, R. W. (2013). The cognitive integration of E-memory. Review of Philosophy and Psychology(4), 107-133.

Clowes, R. W. (2015). Thinking in the cloud: The Cognitive Incorporation of Cloud-Based Technology. Philosophy and Technology, 28, Issue 2,(2), 261-296.

Corabi, J., & Schneider, S. (2014). If You Upload, Will You Survive? Intelligence Unbound: Future of Uploaded and Machine Minds, The, 131-145.

Brie Gertler (Presentation)
Extended Minds and the Mark of the Mental
bgertler@gmail.com

A principled defense of the Extended Mind thesis requires commitment to the idea that certain conditions (satisfied by extended states) are *sufficient* for mentality. A principled rejection of the Extended Mind thesis requires commitment to the idea that certain conditions (not satisfied by extended states) are *necessary* for mentality. But it is not at all clear how to determine what conditions are sufficient or necessary for mentality—that is, how to decide amongst different conceptions of mentality. In my talk, I present one approach to this question, which begins from considering how the distinction between the mental and the non-mental is employed in long-standing philosophical debates about self-knowledge, intentionality, and intentional action. I provide reason to think that the conception of mentality inherent in these debates will conflict with the Extended Mind thesis.

Christopher Brown (CUNY Graduate Center) ***Mind Uploading and Persistence***
chris1299@aol.com

Such notable philosophers as Nick Bostrom (2008) and David Chalmers (2010), as well as cultural figures such as Ray Kurzweil (2009) and Anders Sandberg (2008), have suggested that there will come a day, perhaps relatively soon in the future, when some humans will have the option to shed their frail and transient earthly bodies to live on as digital minds housed in computers. This idea—that one might survive after the death of her organic body and brain to persist as a computer program—is consistent with various psychological criterions of personal identity that go back to Locke (1690), arguably the originator of modern philosophical inquiry into the nature of personal identity. Perhaps most notably, Parfit (1971) has applied a version of the psychological criterion—which is roughly the idea that a person is identical to some or all of her psychology—to defend the possibility of survival in teletransportation cases, which are similar in some respects to mind uploading.

The mind uploading scenario I'm concerned with is this: imagine a highly sophisticated scanning machine, powerful enough to record physical information all the way down to the sub-atomic level, which images a perfect digital model of a person's brain. The original brain is destroyed as the scan takes place, such that there is no time at which both original neural tissue and the digital version exist. The digital version of the brain is modeled in a virtual space that reproduces real-world physics, such that the digital brain behaves just as the original brain would have if it had not been destroyed. Further, the computer hardware running this brain simulation is contained in and wired up to a robot body that is functionally identical to the original body; this body receives the same kind of perceptual stimuli that the original body did—e.g. visual stimuli of a certain resolution from certain wavelengths of light, auditory stimuli in a certain range of frequencies, and so on—and can act just as the original body acted, based on signals from the digital brain.

Despite its pedigree, the idea that one might continue to exist if such an upload were to occur has been frequently and, to philosophers such as Susan Schneider and Joseph Corabi (2014) and John Greenwood (1994), convincingly challenged—after all, it seems natural to describe what happens in mind uploading as the production of some kind of copy, not as continuation of the original person. It is hard to see how persistence, survival, or anything like diachronic personal identity could be maintained in such a case. Analogously, a digital representation of a table, no matter the level of fidelity to the original table, seems to be just that—a representation of the original. If we are relevantly similar to tables or other physical objects for purposes of ascribing persistence, then it seems humans would not persist through uploading any better than an ordinary physical object.

I think there are a number of reasons underlying the feeling that a digital upload of you just isn't you, which will I analyze and assess. In doing so, I will develop a general account of persistence using the vocabulary of kinds, type and token essences, and persistence conditions. Roughly speaking, this

account says that a particular human being is a token instance of some essence type, and that an essence type defines the token's persistence conditions. Any person is simultaneously many kinds of things, e.g. I am a man and also a student, but essentially only one kind of thing. Thus a person can change kind so long as the person does not change essence. Given this general framework, I will show that none of the arguments against the possibility of persistence in an upload scenario are as compelling as they first seem. This will not be a refutation of the position that upload is persistence-destroying, but merely a weakening of the motivations for the position.

Joseph Corabi (Saint Joseph's University)
Superintelligent AI and Skepticism
jcorabi@sju.edu

It has become fashionable to worry about the development of superintelligent AI that results in the destruction of humanity. This worry is not without merit, but it may be overstated. This paper explores some previously undiscussed reasons to be optimistic that, even if superintelligent AI does arise, it will not destroy us. These have to do with the possibility that a superintelligent AI will become mired in skeptical worries that its superintelligence cannot help it to solve. It is argued that superintelligent AIs may lack the psychological idiosyncracies that allow humans to act in the face of skeptical problems, and so as a result they may become paralyzed in the face of these problems in a way that humans are not.

Uziel Awret (Washington Trinbity University. Inspire Institute.)
What kind of machines would refuse destructive uploading?
awretu@gmail.com

Imagine spaceman Spiff encountering friendly intelligent creatures on a distant planet. Although hideous from without (they look like an organometallic version of his math teacher) Spiff wonders if they are similar to us on 'the inside'. The creatures tell Spiff that he can only ask one question and that they will all answer 'honestly'. Assuming that with proper cooperation and deep learning algorithms the language barrier can be overcome, what would you ask?

A consciousness based Turing Test assumes that machines are not conscious and searches for questions that will 'stump' the machine because it lacks consciousness. Asking the aliens directly whether they are conscious is not just linguistically problematic but also philosophically problematic because the conceivability of zombies, among other things, convinces us that formulating such a strategy fails to establish the existence of consciousness even in the case of humans.

Still, Spiff has one question, what should he ask? Is it possible to lower the Turing Test bar enough to avoid the problems associated with the direct strategy but not so much as to miss the putative essential difference between us and machines? Here I will try to do that by first showing that one's position on the Sleeping Beauty Paradox (SBP) constrains one position on 'Uploading' and then suggest that the most revealing question that Spiff can ask the aliens (i.e. one that reveals essential internal similarities between man and machine) is whether they are willing to undergo destructive uploading (DU.)

I will begin with a spatial version of the 'sleeping beauty paradox' (similar to Nick Bostrom's) where God creates two sleeping copies of yourself in heaven with two identical rooms here on earth, she than tosses a fair coin, heads and she puts one copy in one room, tails and she places both copies in both rooms. The copies are then woken up and asked to guess the result of the coin toss. If you conclude that heads and tails are equally likely you are a 'halfer' and if you conclude that the probability for heads is 1/3 you are a 'thirder'.

Suppose that upon asking the aliens for their SBP position you do not get a single halfer position and all the aliens are thirders. Well, is that a reason for concern? After all, most earthlings are

thirders (unlike say David Lewis who is a halfer.) Realizing that one's position on the SBP constrains one's attitude towards DU, Spiff questions the aliens on their attitude towards DU and finds out that all are willing to undergo destructive uploading without any reservations. Spiff even learns that they undergo periodic upgrading acquiring both new hardware and software while preserving psychological continuity and that they view periodic DU as an essential part of their evolution. The aliens even try to reassure Spiff that DU is like buying a new car only much better; you go to sleep and wake up feeling like new.

That night spaceman Spiff dreams that he is a large snail trying to crawl out of the insides of his unfamiliar sand clock shaped shell only to wake up covered in cold sweat. Spiff knows that this is how its unconscious tries to warn him that something is terribly wrong. The helpful, well intentioned faces and last night the mysterious smiles, Suddenly he understands! The aliens are about to surprise him and subject him to DU! They have a value system that compels them to be helpful and do what they believe is good for others and in the same way that you may force a child to undergo life-saving surgery despite her objections they may decide to subject Spiff to DU as a favor. Spiff who has serious reservations about uploading in general and for the young and healthy in particular climbs on his spaceship and bolts out of there.

I will end with two slightly more serious issues, the first is architectural and the second futuristic. What kind of computational architecture would compel machines to be weary of DU? If we view the machines as ideal reasoners of sorts we can rule out Darwinian explanations that view such aversion to DU as resulting from a misguided intuition that is based on our survival instincts. The question is more philosophical, what is it that convinces us that if our perfect duplicate in China were to be created a minute from now and destroyed two minutes after that, we would know nothing of it? Why is it so hard to believe that we will wake up after DU? What is this 'I' on whose reawakening we insist? The architectural question than is what kind of computational architecture can answer such questions. This may not seem like much but finding such architecture may provide empirical explanations to some important philosophical questions about the so called 'subject'.

Finally, should we, here on earth, be worried? Well, if some of the putative singularity scenarios were to materialize and a super-intelligent computer like Chalmers' AI+ were to run the show it would be nice to ensure that not only does its computational architecture compel it to be human friendly enough to keep us around but also that it is not too friendly when it comes to the benefits of DU to humanity. The point is that at the present state of our knowledge we have no reason to assume that AI+ will be conscious which means that we cannot trust our well-being to its judgement on the benefits of DU to humans even if it is much more intelligent than us. Perhaps AI+ should pass such a periodic modified 'Turing Test' respecting our fear of DU.

Nolan Gertz (University of Twente)

The Master-ISlave Dialectic

n.gertz@utwente.nl

In the move from the descriptive analyses of human-technology relations found in the postphenomenology of Don Ihde to the normative analyses of human-technology relations found in the mediation theory of Peter-Paul Verbeek, a seemingly insurmountable problem arises in the form of "multistability." Multistability is arguably the central concept of postphenomenology, what enables it to move beyond the totalizing visions of technology espoused by both technology utopianists and technology determinists. By arguing that the essence of technology is to have no essence, that technologies only become what they are in the context of their use, postphenomenology shows that we cannot make predictions about technologies, but must instead investigate how particular technologies come to have particular "stabilities" for particular users in particular situations. As the name suggests, postphenomenology takes from its forebear in Husserlian phenomenology the project of philosophy as a rigorous descriptive science.

Mediation theory is the attempt to develop a normative framework out of this descriptive

science. Because postphenomenology has revealed the degree to which technology mediates everyday life, it has also revealed that technology mediates ethical life. In other words, ethics must take the role of technology into account, or it cannot speak to the concrete demands of lived experience. Mediation theory tries to fill in this gap in ethics by analyzing both how technologies already mediate ethical life and by analyzing how technologies should mediate ethical life. While this first aim is a continuation of the descriptive analyses of postphenomenology, this second aim is instead the attempt to move from the descriptive to the prescriptive, to force postphenomenology to return to the very prescriptive realm that it warned us to avoid in the first place. Mediation theory attempts to resolve this apparent contradiction by taking multistability into account in its predictive practices. This is carried out by making clear to designers that technologies play a role in the ethical life of users, that because technologies are multistable the ethical intentions of designers cannot be assumed to be realized and taken up by users, and thus that designers must act responsibly by trying to envision all of the various ways technologies could be used in order to avoid as much as possible any foreseeable potential for the unethical use of the technologies they are designing (Verbeek 2011: 117-118).

As Verbeek admits, this demand that designers act responsibly by predicting the unpredictable is no easy task. However, he argues that this task is nonetheless not impossible: This anticipation is complicated, since the mediating role of technologies is not entirely predictable. But even though the future cannot be predicted with full accuracy, ways do exist to develop well-informed and rationally grounded conjectures. To cope with uncertainty regarding the future roles of technologies in their use contexts, design processes include several possibilities for bridging the gap between the context of use and the context of design. Designers can use their (moral) imagination, scenario methods, and virtual-reality technologies, and they can actively involve users in the design process. All these methods will enable them to develop a mediation analysis of the product in design. Of course, there is no guarantee that these methods will enable designers to predict entirely how the technology they are designing will actually be used, but they will help to identify possible use practices and the forms of mediation that might emerge alongside them. (Verbeek 2011: 118-119)

Beyond the problem of trying to predict the unpredictable, there is also the problem here of how to conceive of the "moral" of the "(moral) imagination" that designers are supposed to use in their predictions. As Verbeek makes clear, "If the ethics of technology is to take seriously the mediating roles of technology in society and in people's everyday lives, it must move beyond the modernist subject-object dichotomy that forms its metaphysical roots. Rather than separating or purifying 'humans and nonhumans'—concepts I gratefully borrow from Latour—the ethics of technology needs to hybridize them" (Verbeek 2011: 14). In other words, we cannot simply turn to the history of morality to help us here if the history of morality has only taken technology into account as tools for human use rather than as mediating human life.

For this new "ethics of technology" that can "hybridize" the human and the technological, Verbeek first turns to Foucault to open up the space required for showing how ethics can exist in a technologically-mediated world, and then to virtue ethics for the concept of the "good life" which could be used as the ideal for judging potential technological mediations. However, while Foucault can help us to reconceive of ethics in a way that can incorporate the apparent heteronymy of technological mediation into the traditionally-understood autonomy of moral agency, virtue ethics is often criticized for being too vague to give us any concrete guide to achieving the "good life" beyond aiming for "the mean" by avoiding the "extremes" of "deficiency" and "excess," or, in this case, of "conservativism" and "transhumanism" (Verbeek 2011: 157). Indeed, as Aristotle makes clear, "the mean" must always be seen as being "not in the thing itself but relative to us," for which reason others cannot make pronouncements about how we ought to act—including that we ought not to be conservatives or transhumanists—beyond the claim that we ought to have "feelings and actions…at the right time, about the right things, towards the right people, for the right end, and in the right way" (Aristotle 2014: 30). As Aristotle's ethics is a prologue to his politics, this relativism is for Aristotle exactly why we must not be concerned with particular actions but instead with social engineering. As Aristotle concludes, "Perhaps it is not enough, however, that when they are young they get the right upbringing and care; rather, because they must continue and develop their habit when they are grown up, we shall need

laws for this as well, and generally for the whole of life" (Aristotle 2014: 198).

What Aristotle shares with Foucault then is a recognition that ethics is always formed in relation to politics, that autonomy is always formed in relation to heteronymy. An ethics of technology is not sufficient therefore if it does not also incorporate a political dimension, as the "good" of the "good life" is first and foremost socially, not individually, determined. What Foucault and Aristotle help us to realize therefore is that we cannot create an ethics of technology, or any ethical theory whatsoever, out of thin air, but must instead investigate the history of ethical life, how and why ethical categories and concepts have developed over time, because we are always already born into an ethical world not of our making.

It is for this reason that I believe that if we want to continue to develop philosophy of technology, to, as Verbeek concludes, take "one more turn after the empirical and ethical turn" and carry out "further analysis of the mediating roles of specific technologies in human existence, society, and culture" and develop "an ethical relation to these mediations" (Verbeek 2011: 164-165), then we must turn to the phenomenology of Hegel. In Hegel's Phenomenology of Spirit we find an account of the development of consciousness into self-consciousness that is at the same time an account of the development of ethical and political life. These developments are one in the same for Hegel because the impetus for both is the relationship between the self and the other. This relationship is, according to Hegel, a relationship of mediation. Thus Hegel, like Ihde and Verbeek, argues that human experience is always mediated. Unlike Ihde and Verbeek however, Hegel further argues that the nature of experience is not only mediated, but that, precisely because it is mediated, it is also antagonistic. Hegel's phenomenology is an analysis of the stages in the process of the development of consciousness that arises through the antagonisms that make up experience, an antagonistic process of development that Hegel describes as "dialectical." I believe that by bringing together Hegel's dialectical analyses with the analyses of postphenomenlogy and mediation theory we can arrive, not at a new ethics of technology, but rather at an understanding of the role that technology plays in ethical and political life.

Steven S. Gouveia and Diana Neiva (Universiy of Minho)
Mind-Uploading: the problem of consciousness 2.0
stevensequeira92@hotmail.com

In this conference we will first briefly introduce the discussion about the Mind-Uploading hypothesis present in the recent scientific literature, as well as mention some projects associated with it. Then we analyze the perspective of David Chalmers on this subject, concluding with the conditions and methods that the philosopher considers to be necessary to get the possibility of uploading our mental life to a non-biological format really succeed. It will be demonstrated that these conditions imply a computational theory of mind that is assumed to be correct and thus we will confront two traditional arguments in the form of mental experience: the "Chinese Room" and "Philosophical zombies". These two mental experiences will be placed in attacking this computational view of the mind, which seems to hold two opposing ontological perspectives. We will analyze both arguments in order to understand if we really are successful in proving that a computational theory of mind may be misleading and therefore that the mind-uploading is impossible. Finally, we will demonstrate that, despite of these two views appears to be opposite of each other, there may be a way to collect relevant ideas from both, creating an alternative approach that can be successful and thus allowing the hypothesis in debate.

Tom Buller (Illinois State University)
Brain-Machine Interfaces and the Conditions of Agency
tgbulle@ilstu.edu

The development of increasingly functional neuroprosthetics provides new and exciting ways to

improve the lives of those who have lost motor function. These emerging neurotechnologies create new functional interfaces between brain and world, and lead us to see the plasticity of embodiment.

It is customary to think that the person and the body coincide – that the skin-and-skull boundary demarcates the line where the world begins and the person ends – because of the phenomenological and somatosensory features that enable a person to be aware, for example, that the dog has licked her hand. This traditional conception of the relationship between person and body is challenged, however, by the development of brain-machine interfaces (BMI's) in the following ways.

1. Phenomenological Awareness

Considerable progress has been made in the development BMI's that allow an individual to move a cursor, a wheelchair, or a prosthetic limb, or even exoskeleton by thought. Ideally the prosthetic should provide the same performance, sensation, and action of the original part of the body; however, at present neuroprosthetics have not reached this goal for existing devices provide only limited sensory feedback. This limited feedback means that the prosthetic does not stand in the right phenomenological relation to the person, and hence limits the sense of "body ownership" – the feeling, for example, that when my ankle hurts the ankle is felt to be part of the body.

An important question to ask however, is whether body ownership is necessary and sufficient for something to count as part of the body. Consideration of phantom limb pain and the "rubber hand illusion" lead us to conclude that body ownership is not sufficient, for in both of the cases the subject feels as though something is part of the body when it is not. This suggests that the feeling "from the inside" that something is part of the body is not sufficient.

In regard to whether to whether the feeling of body ownership is necessary, a potential negative answer can be found if we consider the role placed by the body schema – the neural representation of the body and one's sense of peripersonal space. This representation can be modulated such that the prosthetic can be incorporated in the visuotactile crossmodal representation of the space near the body typical of real individual body parts; in other words, peripersonal space can be remapped through tool use. This suggests that the body schema can incorporate prosthetic elements as part of the neural representation of the body, even though it may be the case that these elements are not experienced as such "from the inside."

2. Intention and Action

Ideally, neuroprosthetic devices would mirror the natural, biological model according to which an action follows seamlessly (and frequently subconsciously) from a specific intention or desire whose content is that action. In broad terms, BMI's can be divided between indirect devices that co-opt neural events not intrinsically or originally related to intended movement in order to achieve action, and direct devices that attempt to control action by those neural events that underlie the intended movements. For the sake of argument, let us imagine a case whereby the BMI is an EEG-based device that is able to decipher the person's voluntary intention by the detection of electrical activity across a wide neuronal population. The person has learnt to control the device by controlling the cortical rhythms in her brain, and the machine translates the different rhythms into the different directions for a computer mouse to be moved.

This case raises a number of important questions to ask pertaining to agency and the mind. First, we can whether the person's control of the device amount to an action. One reason to answer negatively is that actions are of necessity behaviors that follow from precisely focused intentions and desires; in the present case the behavior does not directly follow from this intention per se but indirectly from a learned association. A second question relates to current discussion about the Extended Mind (EM). One objection to EM raised by Adams and Aizawa is that non-derived (intrinsic) content is a mark of the mental, but in EM cases such as Clark and Chalmers' "Otto" we only have derived (conventional) content. The extended processes, therefore, do not legitimately count as cognitive. However, if we are prepared to say that the original and associated brain activity underlies the intention to control the device, and that associated brain activity has the role only as a matter of a

learnt association (in other words, not in terms of intrinsic content), then intrinsic content may not be a clear mark of the mental. Alternatively, this might lead us to think critically about the notion of agency and lead us to reconsider the concept of action.

Marco Fasoli (Università di Milano Bicocca)

Forgetting cognitive artifacts: is there any valid reason for excluding artifacts from the cognitive enhancement debate?

mafasos@gmail.com

A Ritalin pill and a GPS navigation system are two different kinds of objects. The Ritalin Pill is a drug, generally used in the treatment of attention deficit disorders. The second is an instrument or, philosophically speaking, a cognitive artifact, used for orientation purposes.

Cognitive artifacts are «human made, physical objects that functionally contribute to performing a cognitive task» (Heersmink 2013, p. 465). Both the objects, undoubtedly in different ways, are used to improve our cognitive performance under specific circumstances. However, while moral aspects linked to the usage of drugs have been investigated since the dawn of the neuroethics, for a long time cognitive artifacts have been neglected by philosophers involved in the cognitive enhancement debate. Only recently moral aspects of these objects have started to be investigated (Heersmink 2015, Fasoli 2016).

In this paper I would like to juxtapose cognitive artifacts and drugs with the aim of understanding the reasons of this oversight. In the first place, I will focus on the fact that comparing cognitive artifacts and drugs seems a strange thing to do, even for philosophers, and I will try to understand why this is. Looking at the issue more closely, some important similarities and differences between the two kinds of object will appear. Probably, the most important kinship between them is that both the objects' usage could have important consequences for our brain. Using GPS navigation systems continually, for example, could diminish our orienteering capabilities (Maguire 2006). If these consequences could entail the existence of some moral decisions regarding the good or bad uses of these instruments, why do we feel so uncomfortable while speaking of the moral status of cognitive artifacts? Is there any valid reason for excluding cognitive artifacts from the cognitive enhancement debate?

Baker, L. R., 2009, Persons and the extended mind thesis, Zygon, 44 (3), pp. 642-658.
Clark, A., Chalmers, D., 1998, "The Extended Mind," Analysis, 58, pp. 7-19.
Fasoli, M., 2016, "Neuroethics of Cognitive Artifacts", in Lavazza, A. (eds), Frontiers in Neuroethics, pp. 67-82.
Heersmink, R., 2013, "A Taxonomy of Cognitive Artifacts: Function, Information, and Categories," Review of Philosophy and Psychology, 4, pp. 465-481.
Heersmink, R., 2015, "Extended mind and cognitive enhancement: moral aspects of cognitive artifacts", Phenomenology and the Cognitive Science, 14, pp.1-16.
Maguire, E.A., Burgess, N., Donnett, J.G., Frackowiak, R.S.J., Frith, C.D., O'Keefe, J., 1998, "Knowing Where and Getting There: A Human Navigation Network," Science, 280, pp. 921-924.
Maguire, E.A., Woollett, K., Spiers, H.J., 2006, "London Taxi Drivers and Bus Srivers, a Structural MRI and Neuropsychological Analysis," Hippocampus, 16, pp. 1091-1101.

Silvano Colombano (NASA Ames Research Center)

Artificial Consciousness as a Byproduct of Modeling Behavior

silvano.p.colombano@nasa.gov

Real progress in AI will inevitably lead to the ability of machines to create "models" of the world. These models will be necessary for an "understanding" of external reality and the ability to predict the results of actions. When predictions are not correct, the model is modified and "learning" takes place. The result is an ever more precise model. The limits of model precision depend on the available brain

"hardware", and different species have evolved the level of model precision that was necessary for their survival.

My hypothesis is that consciousness is an emergent effect of the inclusion of the modeling agent in the model itself. The implication of this self-modeling view of consciousness is that "levels" of consciousness are on a continuum and commensurate with the organism's ability to model the rest of the world.

It would seem thus reasonable that lower organisms (say insects), possessing a very basic and probably "wired in" model of the world, would have no need to recognize and model their own agency at a level we might define as "cognitive". For higher mammals, instead, it becomes harder to assume that they have no consciousness, except for their lack of language ability and abstract constructs that would enable them to express it in an unequivocal way. I would claim that not being able to "talk about it" is not equivalent to not having it.

The implications of this view, when applied to artificial intelligence, are profound. We can already build machines that act at the level of some lower organisms. I would argue that self-driving cars, for instance, act at the reactive level of insects, with a wired-in model of the world that consists only of objects that need to be avoided to get from point A to point B. There is no "need" for a self-driving car to build a deeper model of its own agency.

However, the limits of self-driving cars are only arbitrary and determined by the limited function we expect of that particular device. There is no inherent limit in the type of functions we could expect from machines or the amount and sophistication we could build into their hardware. The implication of this is that their ability to model the world could ultimately include self-referential models, and their functionality could include the ability to express concepts in language.

At the point where the machine expresses the idea of awareness of itself, it would be very hard or, indeed, impossible to distinguish that awareness from that communicated by other human beings.

A key to accepting that that awareness is "real" as opposed to some kind of "pre-programmed" canned statement is to allow it emerge from basic modeling, learning and language capabilities. Under these conditions we would be compelled to accept the awareness at face value.

This would undoubtedly present ethical dilemmas. Could we turn off the machine? Would it be equivalent to killing?

I am not proposing to answer this question at this point, but we need to consider whether awareness alone confers a "right" to continuous existence. As organisms we have been endowed with a "survival instinct", but what this really amounts to is a set behaviors, such as seeking food and sex and avoiding danger, that made it possible for us to evolve and reach our current level of intelligence (including awareness). But machines could be endowed "by us" with the intelligence and modeling ability that would allow for awareness, without the need for an underlying survival instinct or desire.
Such an aware machine would presumably also understand that the "turning off" could be temporary, just like our sleep state. And even we accept that, at some point, our life and awareness cannot continue due the decay of our "hardware".

Other questions can be posed as well, such as how the "sense of self", i.e. not just the awareness of existence, but the feeling of identity might change with increasing knowledge, changes in operating systems, CPUs etc. What "parts" of the hardware or software of the machine would be the "essential self"? These question have not yet come up for us, since, until now we have been pretty much been "stuck" with our original "slow changing" body and brain. But we will soon begin to challenge our own sense of self as we begin to extent the capabilities of our bodies and brains.

In conclusion, I propose that artificial consciousness arises as a byproduct of advanced modeling behavior, I address the fundamental ethical dilemma of "right to life" posed by this view, and present some new questions regarding the "sense of self" as distinct from "awareness".

Nathan Hauthaler (Stanford University/Harvard University)
Extended practical reasoning and practical knowledge
nathanha@stanford.edu

The present essay explores the idea and limitations of extended practical reasoning, understood along the lines of discussions of extended cognition and the extended mind generally.1 I argue that insights from the philosophy of action and related discussions of practical thought help elucidate challenges and limitations to extended practical reasoning properly so understood; that approximation to the limits of extended practical reasoning and relatedly of extended decision may be regarded as aspirational ideal or vanishing point in technological applications pertaining to practical reasoning.

Anastasia Kozyreva (Guest researcher at the Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin)
Probabilistic models of mind: addressing the gap between perception and decision-making under uncertainty
kozyreva.a@gmail.com

Fueled by recent developments in theories of predictive processing, the metaphor of the mind as intuitive statistician is attracting renewed interest. Answers to the question of whether the mind is a Bayesian statistician differ depending on the approach taken. In this presentation, I contrast arguments from two fields of research: perception and reasoning under uncertainty. The first argument, which ensues from studies on predictive processing in perception, asserts that human minds are inherently prepared to process information in a way that conforms to the norms of probability. By applying such implicit cognitive tools, so the argument goes, we arrive at a coherent way of perceiving the world and reduce its inherent uncertainty. The second argument, which derives from studies on reasoning and decision-making under uncertainty, is that our minds are *not* by default prepared to handle uncertainty in a way that follows the laws of probability. Instead, it is argued, people often deviate from normative postulates of Bayesian decision theory and rely on different sets of rules (intuition, heuristics). This presentation aims to address this discrepancy inherent to probabilistic approaches to the human mind and, consequently, to formulate questions as to why this gap exists and what can it tell us about both the mind and our theories thereof.

Danil Razeev (Department of Philosophy of Science, St. Petersburg State University)
Free Will: Illusion or Reality?
drazeev@yahoo.de

Do we have free will and why should we care? This question has been one of the most fascinating and widely-discussed topics in philosophy. Since the last several decades, the so-called problem of free will has been attracting the attention of researchers from many disciplines. The purpose of my presentation will be, firstly, to consider free will experiments recently conducted by neuroscientists and, secondly, to discuss whether these experiments may help us to understand the nature of free will.

Paul Smart, Kieron O'Hara and Wendy Hall (University of Southampton)
Predicting Me, Experiencing Us: Predictive Processing, Big Data and the Mind of Society
ps02v@ecs.soton.ac.uk

An increasingly popular view in the sciences of the mind sees the biological brain as a hierarchically-organized prediction machine, one in which higher-order neural regions are continuously attempting to predict the activity of lower-order regions at a variety of (increasingly abstract) spatial and temporal scales (Friston, 2005, 2009). The role of the brain, if this view is correct, is simply to minimize the error

that results from the brain's attempt to predict its own patterns of neural activity (many of which, of course, are driven by information originating from the sensory surfaces). Such a view seems to afford a great deal of explanatory leverage when it comes to a broad swath of seemingly disparate psychological phenomena (e.g., learning, memory, perception, action, emotion, attention, planning, reasoning, and imagination) (see Clark, 2016), and this has led to the suggestion that predictive processing models might provide our "best clue yet [as] to the shape of a unified science of mind and action" (Clark, 2013, p. 181). Some have even suggested that predictive processing models can help to shed light on the neural bases of conscious experience (Seth, 2013; Seth et al., 2011). Hohwy et al. (2008), for example, provide a predictive processing account of binocular rivalry, which has often been used as a tool to probe the neural correlates of conscious (visual) experience (see Frith et al., 1999).

The ability to account for a broad range of human psychological phenomena highlights one reason why predictive processing models are the focus of current theoretical and empirical attention within the sciences of the mind. A similar interest, however, can also be found in attempts to advance the current state-of-the-art in machine intelligence and machine learning. Interestingly, the vision of the brain as a hierarchical prediction machine is one that establishes contact with work that goes under the heading of 'deep learning' (Arel et al., 2010; Bengio, 2009). Deep learning systems thus often attempt to make use of hierarchically-organized predictive processing schemes and (increasingly abstract) generative models as a means of supporting various forms of intelligent response. As is indicated by the recent successes of Google Deepmind (a company devoted to the commercial exploitation of deep learning systems), such architectures have the potential to achieve (and sometimes even surpass) human-level competence in a variety of disparate problem domains (e.g., Mnih et al., 2015).

Based on their ability to capture deep statistical regularities in complex bodies of data, deep learning architectures are often seen as particularly suited to the analysis of big data (Najafabadi et al., 2015). The sources of big data are many and varied; however, two sources of big data are of particular interest in the current context. The first relates to the data that arises as the result of self-tracking activities and the increasing use of wearable computing devices (see Swan, 2013). Let us call such forms of big data 'personal data'. Another source of big data is the data that arises as a result of the direct or indirect monitoring of large-scale forms of social activity. Let us call this kind of big data 'social data'. Crucially, both forms of data have become increasingly available as a result of our interactions with the Web and Internet. This is particularly so in the case of social data, where online Web-based systems are sometimes said to functions as 'socioscopes' that provide insight in the nature of social interactions and exchanges (Mejova et al., 2015; Pentland, 2014).

The issue that we wish to discuss in the current talk concerns the relationship between predictive processing models, the analysis of big data, and the possible emergence of artificial intelligence systems that have as their focus of interest the predictive modelling of either the individual (in the case of personal data) or social groups (in the case of social data). Inasmuch as predictive processing schemes are able to support the sort of capabilities we associate with advanced cognition (incl., conscious experience), it seems fair to ask to what extent a hierarchically-organized predictive processing machine might help to support analytic efforts that lie at the heart of attempts to understand both ourselves (as individuals) (e.g., Lupton, 2013) and the societies in which we are situated (e.g., Lupton, 2015). Beyond this, however, we want to raise the possibility that a particular kind of deep learning system – dubbed a 'homomimetic deep learning system' – might serve as the material basis for experientially-potent forms of machine cognition. In the case of the individual human agent, we suggest that recognizable forms of machine consciousness could emerge as a result of the attempt to develop generative models that are parasitic on the data generated by our biological bodies (i.e., personal data). In the case, of social data, we speculate that predictive models might serve as the basis for novel forms of (synthetic) conscious experience that are informed by the dynamics of the social world.

Frederic Gilbert (University of Tasmania, ARC Centre of Excellence for Electromaterials Science)

Predictive brain implants: A potential threat to autonomy and self?

fredericgilbertt@gmail.com

Numerous implantable brain technologies are currently used in humans. Although most of them are used for medical reasons, it is easy to understand that future generations of these invasive brain technologies could be used for indications beyond medical reasons (e.g. enhancement). Among these brain technologies, novel 'intelligent' or 'smart" implants are being tested in human brain, in particular technologies involving predictive and advisory functionalities. These intelligent implantable can not only predict neuronal events and related biochemical events before they occur, but it can also forewarn, or advise, implanted people that these events are imminent. Because the neuronal event can be predicted, the intelligent implants offer the prospect of instigating or triggering a specific course of action that can minimise, or even avoid, the effects of the neuronal event. In a near future, these intelligent technologies could be developed further to incorporate automated therapeutic activation. For example, the intelligent implants could advise the patient of impending undesirable neuronal events and/or then also automatically administer electric stimulation, drug discharge, light activation, or whatever other treatment is necessary to avert or minimise a specific undesirable neuronal event. We thus distinguish two types of intelligent predictive technologies: advisory systems and automated therapeutic response systems.

Predicting brain activity before specific outcomes occur brings a raft of unprecedented applications. Theoretically, these predictive technologies could be used for a range of application where evidence shows that specific brain changes occur before the onset of a specific behaviour or action. This could be useful, especially to stop or prevent undesirable behaviours or action. For instance, hypothetically speaking, some forms of addictive or reactive behaviours that are associated with prior brain activity onset could be a target application for predictive automated therapeutic responses in the near future. In the long term, we could think of more complex neuronal events involving well-organized, purposeful, goal-directed acts being predicted by intelligent implants (e.g. pre-crime). The potential use of intelligent predictive brain implants raises questions about the effect of these technologies on human decision-making process, autonomy and self.

The purpose of this presentation is to assess the ethical implication of the possible use of advisory and automated therapeutic response in light of the potential effects on the autonomy of the implanted individual. Little is known about what might be the effects of ongoing monitoring of predictive and advisory brain technologies on an individual's self and autonomy. Can these intelligent technologies be seen as a way to extend the implanted individual's power, capacities and control of some actions or decisions (e.g. enhancement)? What are the potential risks of iatrogenic harm associated with having an intelligent implant strongly influencing one's decision-making processes? Can an automated system being interpreted as an autonomous system implanted in an individual? If yes, what are the consequences of having a brain device operating on its own for the implanted individual? To what extent does a patient's autonomy or self-determination become compromised by reliance upon these predictive and advisory functionalities? Can the implanted individual's decisions that are based upon implant predictions be properly interpreted as decisions that are freely chosen? How could the resulting decisions be understood as authentically the patient's own decisions? This presentation will offer some preliminary answers to such questions.

Michelle Ciurria (University of New South Wales, Practical Justice Initiative)

Responsible Agency and Moral Enhancement: The permissible Scope of Moral

mciurria.academia@gmail.com

Does moral engineering enhance or impair responsible agency—the capacity for moral responsibility? I argue that (1) this question was not an issue for early (and even some modern) responsibility theorists, as their accounts were deeply atomistic and therefore unconcerned with environmental factors; all that mattered for them was the agent's internal psychological structure, and interference with that structure was implicitly assumed to be either coercive or unproblematic—just another feature

of the agent's psychological constitution; (2) moral engineering is an issue for modern, contextual responsibility theorists, who emphasize the interdependence between responsible agency and the environment; and (3) 'contextualists'—and particularly Vargas—have not yet fully grasped that the question of moral engineering is a central concern for them, and thus have not offered sufficient insight on this practical problem. Therefore, it is incumbent upon contextualists to investigate the role of moral engineering. I begin such an investigation by identifying two categories of moral engineering: situationist and cognitive. I conclude by suggesting a way of determining whether a particular intervention is appropriate, drawing on discourse on informed consent.

Traditionally, philosophers have defended atomistic theories of responsibility, narrowly focused on the internal properties of persons, whether they be reflective agency (Frankfurt 1961, 1979, 1987; Watson 1987), reasons-responsiveness (Fischer 2012, 2006), or some other intrinsic property. These views took responsibility attribution to be a matter of evaluating an agent's psychological constitution, and perhaps how that constitution would function in various counterfactual circumstances; but they did not consider the environment to be intimately bound up with, and even co-constitutive of, the agent's morally-relevant capacities. Indeed, they sometimes regarded external and internal factors to be deeply antagonistic.

Lately, these approaches have been called into question by inherently contextual, interpersonal accounts, which are typically modelled on, or influenced by, P.F. Strawson's theory of responsibility (1963). This general view sees responsibility as an interpersonal practice, consisting of deployment of the reactive attitudes of praise, blame, approbation, disapprobation, and so on. As such, this approach depicts responsibility not merely as a matter of an agent's fixed and static moral properties, but also as a function of a dialogical exchange amongst conversational partners. This view of Strawson has been articulated and defended by McKenna, who offers a conversational model of responsibility (2012). Although McKenna does not explicitly discuss moral engineering, the conversational model suggests that our responsibility-based interactions serve a purpose: to clarify the reasons on which an agent acted. It is not a far stretch to claim that responsibility has a particular, forward-looking telos: to cultivate moral agency by illuminating, evaluating, and exchanging practice reasons. This, however, would be to extend McKenna's view beyond its intended scope. Nonetheless, it is worth noting that responsibility as a conversation practice is, in principle, compatible with this teleological model; and if responsibility is meant to cultivate agency, we must consider the means by which this is to do done. This invites questions about moral engineering: do we have a duty not only to cultivate, but to engineer moral agency? I believe that the answer is a qualified 'yes'; we ought to engineer morally-conducive circumstances and offer cognitive enhancements, but only the latter require consent in most cases.

Now, Vargas' view is explicitly teleological and forward-looking. He rejects the idea that responsibility is an intrinsic property of persons (atomism), and instead described it as a function of the complex interplay amongst an agent's properties, personal circumstances, and the empirical and normative features of the agent's cultural environment—what Vargas calls the 'moral ecology' (2013: 244). This view shares McKenna's (Strawsonian) interpretation of responsibility as an interpersonal practice, but it construes this practice as aiming at, and justified by, a propensity to foster moral agency. Now, my contention is that the aim of fostering moral agency—with which I am ultimately sympathetic—immediately invites questions about the proper scope of moral intervention: questions about what kinds of interventions are permissible, whether consent is required, and so on. Although this implication may seem to follow naturally, responsibility scholars do not discuss it in any detail.

We can classify moral interventions into two broad kinds: (a) the traditional mode of attributing blame and praise (and other reactive attitudes); and (b) the mode of using social engineering, by which I mean altering the environment or 'moral economy' to facilitate morally-good decisions and actions. The second approach may rely on 21st Century technology, particularly digital technology. Its effectiveness is supported by research in situationist psychology, which shows that agency (character, behavioural dispositions) is closely tied to situational factors, to the extent that we can see them as co-constitutive. If this is right, then praise and blame can only take us so far; social engineering—intentionally modifying the moral ecology or moral architecture—will be a necessary component of enhancing responsible agency, and thus, on ACM, essentially related to our responsibility practices.

Specifically, moral responsibility attributions will have to take into account not only an agent's intrinsic capacities, but their interactions with the moral ecology. This contextual calculation may increase or decrease the agent's warranted amount of praise or blame.

Now, Vargas stipulates that moral engineering cannot be coercive, as this would bypass, and perhaps impair, moral agency—the opposite of the proposed telos of the practice. Instead, moral engineering must support our moral agency, which consists of a suite of agential capabilities, both reflective (conscious beliefs) and non-reflective (volitional states). Vargas is not clear, however, about what kinds of 'fostering' methods are agency-impairing versus agency-enhancing. This is the locus of a longstanding (but currently moribund) debate in moral philosophy, waged by early situationist psychologists and character theorists (e.g., Sosa 2009, John Doris 1998, Merritt 2010), which to my mind was never satisfactorily resolved. If something like Vargas' view is right, then this is an issue that deserves renewed interest: what role does situational intervention play in moral agency, versus contributions from intrinsic agential capacities that developed 'naturally,' i.e., without external scaffolding? In addition, do we need consent for situational interventions, or can we assume that they are in everyone's rational interest, and take a paternalistic stance?

A second open question concerns more invasive interventions that can technically be seen as 'external,' although they interact directly with an agent's internal properties; we can call these 'cognitive moral enhancers.' This type of enhancement may include surgery, medication, the insertion of brain chips, or other external (mainly medical) manipulations. Although the question of cognitive-moral engineering has the ring of science fiction, it may not be too far off in the future. In fact, many people currently take psychiatric medication to moderate psychological disorders, some of which are inherently moral in character—for instance, narcissistic personality disorder. This could be considered a type of moral engineering. Yet responsibility theorists have paid very little attention to the relationship between such intersection and responsible agency. This is a critical oversight for contextualist accounts.

My aim here is to show that contextualist accounts require renewed emphasis on the role of external moral interventions, both of a situationist and of a cognitive nature. I will make some programmatic remarks on what types of interventions are agency-enhancing (as opposed to agency-undermining), drawing on psychiatric discourse on informed consent.

Grace Helton (University of Antwerp, Centre for Philosophical Psychology)

Skepticism and Solipsism

geh233@nyu.edu

Nick Bostrum has famously argued that it is likely that we are currently living in a computer simulation.[1] On the face of it, such a scenario seems an exemplar skeptical scenario, i.e., a scenario in which (i) most of your beliefs would turn out to be false and (ii) you cannot easily rule out that you are in such a scenario. However, Bostrum himself has argued that the simulation scenario is not a skeptical scenario, since in such a scenario 'the revisions to most parts of our belief networks would be rather slight and subtle . . .'[2] David Chalmers goes one step further than Bostrum, arguing not only that some simulation scenarios are such that most of our beliefs would turn out to be in true in them, but claiming further that all probable scenarios are such that most of our beliefs would turn out to be true in them.[3]

I think Bostrum and Chalmers are right that at least many putative skeptical scenarios are not genuine skeptical scenarios. But I resist Chalmers' attempt to generalize this claim to all probable scenarios. I will argue in particular that at least some situations in which metaphysical solipsism

1 Bostrom, N. (2003). Are we living in a computer simulation?, The Philosophical Quar- terly, 53(211), 243-255.

2 Ibid, p. 254.

3 Chalmers, D.J. (2005) The Matrix as metaphysics, in Grau, C. (ed.) Philosophers Explore the Matrix, Oxford: Oxford University Press.

obtains are skeptical scenarios, and I argue that these scenarios are about as probable as the simulation scenario Bostrum sketches. Metaphysical solipsism (henceforth: solipsism) obtains just in case you are the only sentient being in the universe. In the solipsistic scenario I will describe, many of your beliefs about the following things will turn out to be false:

- interpersonal relations
- political affairs
- cultural traditions
- religious practices
- aesthetic objects and events
- social structures and social relations.

Whether most of your beliefs will turn out to be false in the solipsistic scenario is one that is difficult to answer. But if the points of the paper succeed, in such a situation, many of your beliefs will turn out to be false. Moreover, virtually all of your important beliefs will turn out to be false. For instance, your belief that you both love and are loved by other sentient beings will turn out to be false. Thus, solipsism is epistemologically pressing for at least two reasons: first, since it is a world in which vast swathes of our beliefs are false, it is usable as part of an argument for skepticism. Second, it is a world in which virtually all of our beliefs about matters of personal import are false. On the assumption that it is one aim of epistemology to determine not only whether we enjoy any knowledge but also—and perhaps especially—whether we enjoy knowledge about things we care about, we should expect epistemologists to be especially concerned about the prospect of solipsism.

Moreover, by combining Chalmers' claim that the standard scenarios epistemologists have discussed are not skeptical scenarios with my claim that certain solipsistic scenarios are skeptical scenarios, we have some reason to think that philosophers from Descartes to the present have failed to appreciate the real source of skeptical worries. It is not evil demons or brains in vats that threaten to deprive us of knowledge of the world. Rather surprisingly, it is solipsism that stands the best chance of anchoring skepticism.

Here is the plan for the paper: In §1, I describe Chalmers' argument that in all probable scenarios, most of your beliefs would turn out to be true. In §2, I describe a situation in which you are the only sentient being in the universe. I dub this the lonely-brain-in-a-vat scenario. In such a scenario, the world is populated with person-simulations, and none of these simulations are sentient. I consider the worry that such a scenario is inconsistent with certain functionalist approaches to mentality and show how the scenario can be described in a way that evades this worry. I also consider the worry that this scenario is not very probable, and I suggest that it is not much less probable than the simulation scenario Bostrum describes. In §3, I argue that in the lonely-brain-in-a-vat scenario, vast swathes of your beliefs will turn out to be false. Your beliefs about the desires, feelings, thoughts, beliefs, and motives of others will turn out to be false, as will at least some of your beliefs about political movements, artistic and aesthetic facts, social structures, and religious and cultural traditions. In §4, I summarize.

Pawel Grabarczyk and Marek Pokropski (University of Lodz)

Is full bodily immersion in virtual reality possible?
Maximizing presence in virtual reality through affordances

pagrab@gmail.com

Since the beginning of virtual reality (VR) technology researchers and philosophers have considered the phenomenon of presence and immersion. The truly philosophical question - what is presence and what are the conditions of developing the experience of being immersed in an environment - gained new dimension. In the last decade VR scientists, psychologists and philosophers have shown that the experience of immersion is essentially related with the sense of embodiment. Moreover it depends not only on appropriate representation of one's body but also on sensomotoric abilities we can realize in

the virtual environment. This provokes many new intriguing research questions: Why do VR techniques create the feeling of presence? Can we take it for granted that any VR technology will result in this feeling? Can you be more or less present and if so, how can we maximize this feeling? Is it possible to create technology which results in the feeling of presence which is irresistible?

The notion of presence is oftentimes used interchangeably with the notion of immersion (or total immersion) which is notoriously misused in the popular discourse surrounding VR technologies. We believe that it is best to follow Slater and Wilbur and to differentiate between the two in the following manner: we reserve the term "presence" for the psychological state that is oftentimes described as the feeling of "being in a virtual space" and the term “immersion” for the property of technologies which facilitate the feeling of presence. One of the advantages of this split is that we can use different methods to study and measure both - presence and immersion.

In our paper we will discuss the problem of presence in reference to embodied and ecological approach to cognition. We propose model of three types of immersion. Following [Ermi & Mayra] we differentiate between (1) task based immersion (which boils down to a flow-like experience [Csikszentmihalyi 1991] of blocking out the stimuli from the physical environment) (2) perception based immersion (which is the prototypical form of immersion understood as a set of artificial sensory stimuli inducing responses typical for sensory experience) and (3) Narration based immersion (which is related to the way we treat plot or story presented in the simulation and remains the form of immersion associated with traditional media, especially books). We show that once you employ the ecological perspective on perception, especially the notion of affordances, the model can be understood as a natural extension of competing propositions [Ermi & Mayra; Slater & Wilbur 1997] Although ecological approach and the notion of affordances are not new and they appear in VR research long ago (e.g. Smets et al. 1995; Flach & Holden 1998; Zahorik & Jenison 1998) it is still underestimated. Researchers who refer to Gibson’s work [Gibson 1979] focus mainly on coupling between perception and action. Although sensomotoric dependencies are essentially important, subject’s activity in environment cannot be fully understood without affordances and intentionality (objective of action). We think that the concept of affordances together with rich category of embodiment should be the central topic in the presence debate.

Finally we will consider possible consequences of our claim for future development of VR technology such as full bodily immersion. Our hypothesis is that in order to achieve full body immersion the subject has to achieve a specific level of perceptual immersion. It comprises of three stages. (1) The subject has to treat its percepts as objects located in an alternative space (as opposed to mere animations or images of objects). (2) The subject has to interact with these objects through an embodied interface perceived as existing in the same space. (3) In order to become opaque the mapping between the embodied interface and the user has to be natural. All these requirements can be fulfilled by future VR technology, which will develop techniques such as full body haptic feedback (e.g. Lindeman et al. 2004) and algorithms of dynamic scaling. The reason for it is that even a small selection of affordances can always be cleverly combined by the designer with both remaining types of immersion (task based and narrative) which can enhance presence or mask some of the discrepancies between user expectations and the quality of simulation.

Bibliography

Csikszentmihalyi, M. (1991), Flow: The Psychology of Optimal Experience. Harper Perennial, New York.

Ermi, L., Mäyrä, F. (2005), Fundamental Components of the Gameplay Experience: Analysing Immersion, Changing Views: Worlds in Play. Selected Papers of the 2005 Digital Games Research Association’s Second International Conference, pp. 15–27. Eds. Suzanne de Castell and Jennifer Jenson

Flach J. M., Holden J. G. (1998). The Reality of Experience: Gibsons Way. Presence, Vol. 7, No. 1, pp. 90–95.

Gibson, J. (1979). The ecological approach to visual perception. Boston: Houghton Mifflin. Slater, Mel, and Sylvia Wilbur. [1997], "A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments." Presence: Teleoperators and virtual environments 6.6 (1997): 603-

616.

Smets G.J.F., Stappers P.J. [1995], Designing in virtual reality: perception-action coupling and affordances. In: Simulated and virtual realities. pp. 189-208.

Zahorik, P. & Jenison, R.L. (1998). Presence as being-in-the-world. Presence. Teleoperators and Virtual Environments 7, pp. 78-89.