Corporantia: Is moral consciousness above individual brains/robots?

Abstract: This article calls out the common assumption that moral consciousness occurs at the level of individual brains and robots. It explores the alternative, evidence against the assumption, and provides a means to further test the assumption. It also discusses the consequences of making or abandoning this assumption, especially the consequences for the further evolution of robots.

Keywords: Ethics, robots, social technology, consciousness, social epistemology, corporantia, evaluative diversity

1 Introduction: The less-anthropocentric model

We are told that life appeared on Earth billions of years ago in the form of single-celled organisms and that only recently did any cells relinquish their independence to form into bodies [1]. Thereafter, the cells in bodies specialized more and more so that modern humans are bodies composed of cells of hundreds of specialized types ranging from neurons to bone cells to muscle cells and skin cells. This specialization brings humans great power, but prevents individual human cells from surviving independent of the body. The ancient Latin name for entities which form into a body, "corporantia," would apply to the cells of the first bodies, but the term is more likely to be used to refer to humans and robots who relinquish their independence to form interdependent parts of a social body, such as a corporation, church, family, or nation.

Arguably, humans and robots are already interdependent, but many modern humans willfully shift from one social body to another, more like a parasite than like corporantia. Unlike a bone cell which would never attempt to fill the role of a neuron, many modern humans seek ability to fill other roles in society. In other words, many modern humans aspire not for equality, but for interchangeability. Most importantly, we each expect to develop our own independent capacity for moral consciousness as shown in Figure 1.

The vertical dimension of this figure comes from integrated information theory [2]. According to this theory, removing a piece of a person's brain might not prevent that person from having consciousness, but could nonetheless restrict that consciousness, so consciousness must come in grains of different size, $\Phi_{\text{Max}}$. For example, to be conscious of a book being a book, a person or machine would need to understand the phenomenon of written language; as we gradually reduce the brain, the ability to manifest such consciousness would disappear long before the ability to manifest mere consciousness that one is warm, so consciousness of one's warmth has smaller $\Phi_{\text{Max}}$. Arguably, even a single molecule has sufficient complexity to manifest consciousness of its own warmth. The vertical dimension of figure 1 is $\Phi_{\text{Max}}$.

Panpsychists believe that every part of the universe has some degree of consciousness, so they would see figure 1 as a picture of the entire world, but others would see Figure 1 as excluding molecules beyond bodies they believe could manifest consciousness (e.g. beyond pets, cyborgs, future robots, etc.). Those who give much thought to hives commonly believe hives have greater $\Phi_{\text{Max}}$ than in-
individual ants and bees (i.e. can solve more complex mental problems). Our habit of deferring certain problems to legislative bodies, juries and teams implies that we likewise think certain problems demand greater $\Phi^{\text{Max}}$ than individual humans can manifest. We do not collaborate merely because many hands make for faster work but because we do not trust individual doctors to understand the full complexities of optimizing our health, economy, ecosystem, etc. Integrated information theory may seem controversial to anyone who thinks some part of his/her own consciousness could manifest without a physical host, but it is a theory worth considering until we can find a better explanation for the relationship between consciousness and brains.

The point being expressed by figure 1 is that each modern human tends to believe that his or her consciousness is just barely sufficient to make him or her a moral agent, thus privileged over his or her own components (which are too dumb to know better), yet morally deserving of rights which privilege him or her over the entities of which he or she is a component. Tests of moral education are administered to individual human bodies. Voting rights are allocated to individual human bodies (often one vote per body). Individual human bodies are put on trial and can be compensated in courts of justice. Most robots are currently counted at the level of cells/circuits, which could be components of a human body - as in the case of a pacemaker. Science fiction confronts the possibility of developing robots as complex as human bodies, and wonders whether such robots might have moral agency and deserve the same rights we do.

Reality doesn't necessarily align with this model. It is possible that moral consciousness requires a complexity greater than that of an individual human body, as shown in Figure 2.

If we analyze the mechanisms of moral understanding into its functional components, we notice that different bodies tend to align with different functions in that mechanism [3]. This may be why radicals so consistently oppose conservatives: because one’s function in the corporation is to provide novelty while the other’s is to provide fidelity. Moral consciousness includes both devising new ethics and defending established ethics, so both kinds of bodies participate in moral consciousness (as do neurons and DNA), but figure 2 suggests neither could be sufficient alone.

Many humans have switched functions, for example, proposing a new ethic on one occasion but defending an established ethic on another, and humans are even more flexible than stem cells in that regard because humans switch in both directions; however, it is doubtful that individuals are as effective as teams at filling all functions simultaneously. The coincidence of specialized evaluative functions across cultures is evidence for the existence of at least some objective evaluative truths that would make such functions useful - eyes do not evolve if there is nothing to see - but the persistence of evaluative disagreements within cultures evidence inability to recognize those truths as individuals. The situation of modern humans seems very much like that of the cells which formed the first bodies: specializations are still in flux and many individuals can shift between them, but collaboration has proven superior to singular being.

Corporantia care whether moral consciousness occurs at the level of individual bodies, because that is the premise which justifies human rights which protect even the worst humans from being treated as parasites, but the thing that raises the question of moral consciousness today is the development of robotics. First of all, if the moral consciousness of individual humans is illusory, then the quest to develop moral consciousness in individual robots seems less reasonable. Rather than expect it to manifest in superintelligent AI, we would expect moral consciousness to manifest across a society of intelligences, which might include humans as well as robots. Some strategies for machine intelligence already look like internally-conflicted societies (e.g. genetic algorithms, neural nets, etc.), and adding humans to those societies (a complement of human augmentation) would make that machine intelligence interdependent with humans. Second, machines are often designed to do things faster than humans, and one thing machines might do faster than humans is to evolve specializations. In many ways, human evolution has cleared a path for robots, but the development of robots could clear a path for human evol-
ution in this regard, causing humanity to specialize into types much more quickly. That may already be occurring.

Notice that humans and robots occupy different cells in the first figure, but all types of body in figure 2 can be either human or robot. If we believe that moral consciousness is above individual brains/robots, then distinctions between humans and robots are less meaningful. Especially because there are pressing decisions to make about how to design robots, the nature of moral consciousness deserves our consideration.

2 Evidence that Anthropocentrism Produces Political Polarization

One explanation for disagreement over such moral questions as whether abortion is permissible is that no objective moral truth exists to ground agreement [4]. That is not to say that abortion is OK so long as we disagree, but that statements using terms like "permissible," "praiseworthy," "right," "wrong," "ought" and "should" lack real world referents.

This article has two reasons to quickly skip past that explanation: First, because that explanation obviously supports the claim explored in this article that objective moral consciousness in an individual robot (or human) would be illusory. Second, it seems implausible. For example, the fact that so many people do not bother to attempt impossible tasks (e.g. walking through walls) implies remarkably widespread agreement with the moral claim that "we ought to attempt only what we can achieve," and the best explanation for such coincidence of behavior is that an objective truth about this claim exists and coordinates us (whether we admit it or not).

If we do not take the persistence of moral disagreement as evidence that moral consciousness never occurs, we can at least take it as evidence that it occurs at a level higher than bodies. If bodies were individual moral agents, then each of us would be individually responsible to develop moral knowledge, and we could discover our errors by watching for disagreement - disagreement is a red flag warning us that someone is not aligning with truth. If we were individual moral agents, then we could respond to each moral disagreement by engaging opponents in education until one side had discovered its error.

Some disagreements get resolved that way, but not all. It turns out that certain biological differences among humans predispose us to different values which, in turn, shape our beliefs even before we realize it. For example, experiments reveal that 40% of variation in political ideology is explained by genetics [5], yet over 70% of us would reject the most qualified candidate for a scholarship if the candidate happened to affiliate with the political party that opposes our own [6]. The injustice and inefficiency measured in the scholarship experiment was not conscious; participants did not realize that their judgements had anything to do with politics. We are not very self-aware and it is unlikely that we ever will be.

The claim that some disagreements are impossible to resolve because they stem from differences in values is known as "evaluativism" [7, 8]. For example, a disagreement between a theist and an atheist may result from differences in epistemic practices which ultimately stem from differences in values. Thus, the theist and atheist might never be able to achieve agreement about whether God exists.

Like many other -isms, evaluativism supplements its descriptive claim with a prescriptive claim: How should one respond to the impossibility of reaching agreement? Evaluativism is the prescription that one should discriminate on the basis of evaluative diversity in these cases. Harty Field described it this way: "...in dealing with a follower of the Reverend Moon, we may find that too little is shared for a neutral evaluation of anything to be possible, and we may have no interest in the evaluations that the Moonie gives. The fact that he gives them then provides no impetus whatever to revise our own evaluations." In other words, the evaluativist tries to resolve disagreements with people who share his or her own values, but not with people who don't.

Here's why evaluativists need to segregate: In addition to acknowledging the impossibility of resolving moral disagreements that stem from evaluative diversity, evaluativists acknowledge their potential to make moral errors and their need to discover such errors by sorting through disagreements with others. Around people who share their own values, evaluativists retain the possibility of sorting-out which is in error; however, if evaluativists take seriously the intelligence of people with opposing values, then their error-catching strategy can get hung-up on irreconcilable disputes. To evaluativists, political polarization is a practical necessity.

One objection to evaluativism is to assert that disagreement indicates error even when it stems from evaluative diversity, so segregation merely allows such error to persist outside the evaluativist's awareness. Here the evaluativist can respond that the principle of ought-implies-entails that it is OK to let such errors persist. This principle tells us that we have no moral obligation to catch errors that we cannot possibly catch. If we locate moral responsibility at the level of the individual, and the dispute
cannot be resolved at the level of individuals, it follows that there must be no moral responsibility to resolve the error.

Notice that this defense of political polarization is premised on locating moral responsibility at one’s own level. If we hold robots responsible individually, we will lead robots into that same reasoning. Evaluativists will design robots to share their own values, liberal or conservative, competitive or compassionate. Robots will reach different conclusions about how to behave, just like humans do, and each robot will dismiss the objective moral knowledge of any human that doesn’t share its values. Thus, most robots will treat most humans the way human conservatives and liberals currently treat each other. This is the logical consequence of the belief that one has individual moral agency.

Robots would not have this problem if they were corporantia. Modern robots are open to correction; they do not take responsibility for catching their own moral errors. Since modern robots locate the “ought” reasoning at a level higher than themselves, they cannot cite the ought-implies-can principle to defend their right to make moral errors. Robots already contribute to moral consciousness, just as neurons contribute to thought, and that would continue even if robots reach levels of consciousness exceeding that of individual human bodies, but being corporantia would stop robots from treating us with the same disrespect modern humans have for their political opponents.

3 Testing the assumptions

Robots will be able to tell when their intelligence exceeds that of humans; then convincing robots to remain corporantia would require convincing robots that even humans lack sufficient intelligence to achieve moral consciousness individually. We can test human intelligence empirically by manipulating the evaluative diversity of human social groups and measuring the impact of such manipulation on group intelligence. If depriving societies of certain kinds of evaluative diversity turns-out to morally handicap those societies, then we have evidence that people are not sufficiently intelligent individually.

Perhaps the closest we have come to running such experiments is the teamology work pioneered by Douglass Wilde at Stanford [9]. Wilde controlled the composition of student teams in design classes, and showed that teams engineered for greater evaluative diversity significantly outperform self-selected teams. Variations on his experiment were replicated at Carnegie-Mellon, Loyola University of Los Angeles, Oregon State, Shanghai Jiao-Tong, Sungkyunkwan University, U.C. Berkley, U.C. San Diego, the University of Florida, and U.T. Austin. Left to their own devices, modern students fall into evaluativism as a form of discrimination and this significantly decreases their intelligence.

The teamology work was not designed to sift good people from bad, but designers of potential robots have a moral obligation to sift good designs from bad, so we need to run experiments designed to do that. If we do not run the experiments in a lab, the deployment of robots in real life will be our experiment.

The research needs theories of specializations expected to serve specific functions in moral consciousness. Such theories allow for the possibility of multiple kinds of good robots and people, but also allow for the possibility of bad robots and people—robots and people who fail to serve any function well. A measure which classified people or robots by kinds of goodness would leave some people and robots unidentified. One such theory is the GRIN theory, named after four functions serving moral consciousness: Gadfly, Relational, Institutional and Negotiator. The GRIN Self-Quiz has not yet been used in teamology-style experiments, but the validation study did find that the instrument detects bimodal distributions for all four types among humans [10].

The GRIN model offers social science something akin to what the periodic table of elements did for chemistry. It aims to identify stable types that we will inevitably encounter again and again, in robots as well as in human societies. The periodic table of elements developed over the course of many years and many experiments. The earliest versions left much to be desired, and the GRIN model may be just as rough, but serves as an example of the kind of work that would be required to test our assumptions about our relationship to moral consciousness.

4 Conclusion

Bodies have components and are components in larger entities, so Integrated Information Theory implies that the consciousness manifest in a body is neither maximal nor minimal. We tend to assume that bodies are sufficiently complex to manifest moral consciousness, but the persistence of moral disagreements between bodies suggests limits on the ability of individual bodies to discern objective moral truth. Much as a neuron can be conscious of a book only in the sense of being part of a brain that is conscious of a book, individual bodies might be conscious of moral-
Corporantia: Is moral consciousness above individual brains/robots?

Corporantia is only in the sense of being part of a team or society that is conscious of morality.

Moral disagreement will not be resolved through education - it correlates with biological differences. That leaves robots and humans with three kinds of ways to respond to moral disagreement: They can ignore its evidential weight with respect to the veracity of the disputed claims, they can count it as evidence merely that those who do not share their own values fail to discern objective moral truth (i.e. evaluativism), or they can count it as evidence also against their own individual ability to discern objective moral truth (i.e. acting as corporantia).

Just like one can empirically measure where consciousness of a book occurs and what kinds of cells help or hinder that consciousness, one can empirically measure where moral consciousness occurs and what kinds of robots would help or hinder that consciousness. On the other hand, rather than seek such evidence and use it to test the premises of corporantia, one could choose to be evaluativists or corporantia for practical reasons. Modern systems of justice and moral education are built around the assumption that moral agency occurs at the level of individual bodies - so being corporantia would present the practical challenge of reforming those institutions. On the other hand, evaluativism forces political polarization, so building evaluativistic robots would introduce the practical risk of irreconcilable political divide between them and one’s human children (which frequently have different evaluative predispositions from their parents).

Mass deployment of technology can impact the way a society functions much as a drug can impact the way a body functions. An individual consumer may personally boycott the use of any given robot design, but cannot control whether that design is used by the rest of society - especially by his or her political opponents. If moral consciousness occurs at the level of society, then mass deployment of robots could improve or degrade our moral consciousness (against our will!), and we should want development of robots to go through something like clinical trials for new mind-altering drugs. If we aren’t confident that we each have an individual moral consciousness, we might demand such testing just to be safe.

References

[1] S. J. Gould, The evolution of life on the earth, Scientific American, 1994, 271(4), 84-91
[2] G. Tononi, M. Boly, M. Massimini, C. Koch, Integrated information theory: From consciousness to its physical substrate, Nature Reviews Neuroscience, 2016, 17(7), 450-461
[3] C. Santos-Lang, Moral ecology approaches to machine ethics, in S. P. Rysiewy, M. Pontier (Eds.), Machine medical ethics, Springer International Publishing, Cham, Switzerland, 2014, 111-127, DOI: 10.1007/978-3-319-08108-3
[4] J. L. Mackie, Ethics: Inventing Right and Wrong, Harmondsworth: Penguin, 1977
[5] P. K. Hatemi, S. E. Medland, R. Klemmensen, S. Oskarsson, L. Littvay, C. T. Dawes, B. Verhulst, R. McDermott, A. S. Nørgaard, C. A. Klofskar, K. Christensen, M. Johannesson, P. K. Magnusson, L. J. Eaves, N. G. Martin, Genetic influences on political ideologies: Twin analyses of 19 measures of political ideologies from five democracies and genome-wide findings from three populations, Behavior genetics, 2014, 44(3), 282-294
[6] S. Iyengar, S. J. Westwood, Fear and loathing across party lines: New evidence on group polarization, American Journal of Political Science, 2015, 59(3), 690-707
[7] H. Field, A priority as an evaluative notion, in P. Boghossian, C. Peacocke (Eds.), New essays on the a priori, Oxford University Press, 2000, 17-149
[8] R. Martin, The past within us: An empirical approach to philosophy of history, Princeton University Press, 1989
[9] D. J. Wilde, Teamology: The construction and organization of effective teams, Springer Science & Business Media, 2008
[10] C. Santos-Lang, Measuring evaluative computational differences in humans, Figshare, 2016, DOI: 10.6084/m9.figshare.4003407.v1