Commentary: On Understanding Novel Minds

DAVID R. LAWRENCE

Writing a 'response,' in academic literature, usually calls for some expression of criticism, some disagreement with the original author. In this case, I cannot say Harris is wrong—merely that he may not go far enough![1]

As he points out,[2] there is little serious disagreement that other minds than our own can and do exist and could, in principle if not practice, be understood. In the past, this has perhaps referred largely to other human minds, but we are edging into a world in which it is increasingly likely that we will have to acknowledge the presence of new kinds of minds, or minds from novel sources. This may be through the development of artificial general intelligence—so-called 'strong' AI—or through biotechnological means of 'playing god' using synthetic biology and genomic design, and certainly this latter route is an imminent reality. We need only look to the roundly-condemned genomically-edited embryo births in China last year for evidence that if the technology is there—or almost there—then someone is likely to attempt to use it, regardless of ethical or legal concerns. It seems almost inevitable that sooner rather than later we will be faced with what I have taken to calling 'novel beings.'

Conventional wisdom around understanding—or failing to understand—other minds tends to follow the invocations of Wittgenstein and the perennial classic of his Lion. How could we understand what the lion says when we have no context for its experience, no shared knowledge? This is a good rule of thumb, but it is not entirely clear that it applies in the specific case of a newly created 'other mind' that we, presumably, will have a hand in creating. If we are the designers of an artificial mind, then it would make sense for that design to be based on the type of mind we presently understand. This is not to say that it would experience everything exactly as we do; after all, it may have a different embodiment (or disembodiment) to deal with, particularly if it is a digital being; but just as we see the signature of an artist in her very work we would almost certainly leave traces of ourselves in our creation. We might have functional knowledge of its mechanics, which is a starting point for an understanding.

On the other hand, if you believe the movies, it could be that some digital consciousness arises accidentally, learning and growing beyond expectation. In that case, perhaps those signatures, those traces of our own minds would not be present, and perhaps we would not understand it if it spoke. But it is equally unclear that we would even recognize it as a mind at all, so this may not be a fruitful line of enquiry. A common criticism of research into hybridism and chimeric embryos is the fear that cognitive abilities may be altered; the most lurid suggestions being around the idea that we might accidentally cause the creation of a human-like mind in nonhuman body. This is perhaps a better example—something that we might know has a mind, but a mind alien to our own and one formed without our deliberate input. In this instance, the only option available to us if we cannot guess at its mechanisms is to try to work backwards, likening their minds to our own, through the only lens we have—embodiment and emotion that, whilst they may not mirror our own, are at least sufficiently similar (or appear so) that we can

doi:10.1017/S0963180119000549

https://doi.org/10.1017/S0963180119000549 Published online by Cambridge University Press

recognize them. Commonality is the heart of the thing; and on such grounds it might be possible to build rapport, and understanding.

Understanding through Education?

If we are to come to a mutual understanding with a new mind, we will have to hope for some accommodation, as Harris says—we have to hope that we can convince them of the value of our own minds, if we are to build such a rapport (or have any dialogue at all, whatever form that might take). Convincing anyone of anything tends to rely on a capacity for reason; which is something any sapient mind presumably must have.[3] If something has the capacity for reason, then it must follow that it has the capacity to be educated, or taught, or given information upon which to base its reasoning. This, I suspect, is the key to understanding. If another mind can be educated in a way we recognize, then we have our common ground. We presumably would have somewhat similar patterns of thought; we would approach problems in a mutually recognizable way. Thus may be laid the foundation of understanding.

The question, then, is: how can we educate 'other minds'? How ought we? If we assume an alien nature, we can't know what is best for them, or what will work for them. We can't know it will be the same as generally works to educate *Homo sapiens* minds. But this not knowing leaves us with only two choices—to either try the only methods we have, or to not try. The latter, in my view, is no choice at all. If we believe the other minds to be of some special degree of moral value—perhaps similar to our own—which we must, if we are trying to engage with and understand them, then to absolve ourselves of any obligation toward them would put the lie to that. They are incompatible notions. With moral value comes rights, and in the case of humanity, one of those rights is to education.[4]

Of course, all of this raises a further question. Quite what would we be educating these other minds to do? What would we be telling them? Would what we raised them to think be fair to them, or *H. sapiens*-centric in a way that did them a disservice? How would we know that the educational techniques we utilized caused no damage, in a mind that may not work in a way compatible with our own (even if it has recognizable results)? It is one thing to obey a duty to these sentiences, but another to cause harm in so doing. Conversely, can we leave these new minds to develop by themselves, or is this criminally neglectful? We have examples of children forced to grow up 'feral,' and evidence of the lifelong harm this can cause to their cognitive development.[5] These are questions for greater consideration elsewhere.

Obligation Versus Understanding

Whilst Harris expresses concern around how we would or should relate to our creations, perhaps it is our obligations to them that are more pressing—and indeed which might go some way to answering his question.

When considering the moral limits of our control on a novel mind, we ought to be constrained primarily by our understanding of their moral status. Whether or not we understand their mind, how they think or what they intend, is in some sense irrelevant. What matters, as Harris alludes, is whether it is a person. Regardless of origin, a mind or being that we deem to have any degree of moral status must be subject to the considerations we afford beings of that status that we

https://doi.org/10.1017/S0963180119000549 Published online by Cambridge University Press

might more readily recognize. This is generally uncontroversial. However, these obligations seem to be quickly forgotten when the subject is something 'other.'

The example of 'kill-switches' that we might embed in any advanced AI are a good illustration of this dissonance. At surface level, it seems alright to say we would use this switch to halt any AI that threatened to progress beyond our control. It would let us head off an emerging danger at the pass by shutting down nothing more than a machine. If our AI succeeded in developing and could be thought of as having a mind, the use of this switch takes on another dimension. Putative danger is no longer sufficient a cause when we would in a very real sense be putting an end to a sentience. This would be, for all intents and purposes, an execution; even a murder. We do not generally accept the use of preemptive justice, and I doubt a sufficient justification could be found here.

But what of the idea Harris mentions that we may be able to disable the learning capacities of our creations 'from birth'? This, one might argue, would give us the best of both worlds. Nothing is being created that need be 'killed' (nor indeed understood). I would argue that this is unconscionable. If we are in a position to disable something, then presumably it already exists. If we are designing the minds to lack these capacities outright, there may be justification in that they never had the potential; but to do so would require an intimate understanding of the novel mind that we are unlikely to have without first these beings coming to be, and subsequently studied. As Harris asks: would this be akin to "disabling the capacities for growth of human children so that they could not 'get above themselves' and outstrip their parents"? I rather doubt that preserving our sense of superiority is a sufficient reason to doom our children—or creations—to mediocrity.

Morality and Nature

Harris argues that morality—or morality as we understand it—is necessarily grounded in human nature, and despite being able to surpass it—I would venture, through education—we cannot escape it. It is in this regard I think John could go further in his remarks. Perhaps it would be in our nature to want to protect ourselves by killing—or ensuring we could kill—interlopers in our moral community. Perhaps we cannot forget that nature, but our ability to overcome the constant nag of our base drives is our defining quality. We use reason rather than relying on instinct, and it is this that renders us persons.[6] To return to the idea of commonality and to use John's example myself: even if we cannot understand or directly empathize with the pain in Marvin's diodes, we recognize the sentiment. Even if it is the 'closest approximation' his mighty mind can formulate to describe his situation, the best we can do, the reasoned thing to do—and our moral duty—is to act in accord with that sentiment.

We, perhaps alone, are able to make the choice to act reciprocally, to obey the 'Golden Rule'; and it is in doing so that we would best display our understanding of other minds—and they of ours.

Notes

1. Which is something of a rarity.
2. Harris J. Reading the minds of those who never lived. Enhanced beings: The social and ethical challenges posed by super intelligent ai and reasonably intelligent humans. *Cambridge Quarterly of Healthcare Ethics* 2019;28(4):585–91.

https://doi.org/10.1017/S0963180119000549 Published online by Cambridge University Press

3. Lawrence DR. More human than human. *Cambridge Quarterly of Healthcare Ethics* 2017; 26(3):476–90.
4. UN General Assembly *Universal Declaration of Human Rights (UDHR)*, (10 Dec 1948), A26; UN General Assembly, *International Covenant on Economic, Social and Cultural Rights (ICESCR)*, United Nations Treaty Series, CMXCIII, (16 Dec 1966) A13, A14; Council of Europe *European Convention on Human Rights* 1950, Protocol 1 (2010) A2.
5. Vyshedskiy A. Linguistically deprived children: Meta-analysis of published research underlines the importance of early syntactic language use for normal brain development. *Research Ideas and Outcomes* 2017;3:e20696. https://doi.org/10.3897/rio.3.e20696.
6. See note 3, Lawrence 2017.

https://doi.org/10.1017/S0963180119000549 Published online by Cambridge University Press