The Locality Problem: balancing the need and dangers of subjectivity and its refusal

Abstract

Revolutionary technology challenges our use of subjective decision making. In the utilitarian calculation, the utility of an action scales with the amount of sentience affected by it. It follows that, for any agent performing a moral action, the smaller the ratio between the effect on others and the effect on that agent, the less the values of the agent should be considered. Very high connectedness risks therefore rendering agents insignificant. In unintuitive problems, logic tends to produce better outcomes than intuition. Consequentialism would, in these cases, urge for the logical approach and claim that that human cognition is maladapted for these problems. Both personal insignificance and personal maladaptiveness are reasons to adopt a more objective view. Problematically, many values seem to be intrinsically linked to subjective decision making. Removing subjectivity risks therefore eradicating sources of utility. I contrast the local (subjective) view with a universal (objective) view, and define the locality problem as the growing proportion of situations where utility is maximised by using a universal and rational method, instead of the local view. Two strategies are proposed. Following the principle that local decision making is best in local environments, the first is to isolate actions. However, various incentives for keeping many systems global make this difficult. Based on the general principle of precaution, the second strategy is extra caution when promoting new technologies. Unfortunately, their associated risks might be more hidden.

1. Locality

Values are local. That means that they exist solely in the minds of sentient beings. If life were to exist some hundreds of millions of light years away, values could look very different. If life were not to exist there, neither would any values. In this article, I will refer to locality, in contrast to universality, to the set of ideas and values that make up the moral view of sentient beings. I use the definition of value as being the importance, utility, usefulness, or other reason for regard that something is believed to have, in a given situation. Note that, with this definition, the only thing that prevents us from saying that an iron nail values colliding with a magnet is the character of experience that is interlinked with consciousness. From a universal perspective, values are just the biases or systemic tendencies that seem to preserve

the character of that being. From a local perspective, they are the preferences that we believe will make us happy.

I define religion by its failure to separate universality from locality; in religion, some superhuman figure or process is the origin of good and evil.[1] In accordance, many Christians would claim that the bible - being God's message - tells the absolute truth. Similarly, humanists believe that humans have an intrinsic value merely from existing. On the other hand, a failure to separate locality from universality defines (passive) nihilism. Since everything is ultimately meaningless nothing is ever meaningful, or so the argument goes. This separation and more are presented in table 1.

Table 1: Some examples of how perspectives can be categorised in the local or universal view. This dichotomy is simplified in that the continuity of the spectrum and multiple dimensions of judgement are neglected. For instance, codes of conduct are more universal than the dictum “do what you want” and a (passive) nihilist would not care about existential risks.

Moral view	Local	Universal
Ultimate source of meaning	A fiction	None
Nature of value	Religion	Nihilism
Model of ethics	Personal	Utilitarianism
Time perspective	Neartermism	Longtermism
Population ethics	Person-affecting view	Impartial view
Obligations	Codes of conduct	Apply consequentialism
Acceptance of the “Repugnant conclusion”[2]	No	Yes
Worries about	Misfortune	Existential risk

Moral philosophy is a rational attempt to answer the question of how we ought to behave. Every subject knows the importance of intuition and emotion when acting. Only when we try to agree on clear principles do we try the more objective way of ethics. In this thinking, discrimination in time and between beings does not make sense, and thus impartiality is key for the universal view.

[1] Religion is often defined in different ways. This definition does not use the even more variously defined term ‘God’ and this enables me to call some atheists religious. Interestingly it does not say that the claim that morality can be objectively studied, falls under religion. Further, even if morality is studied objectively, values are still only local since this is the study of local values.

[2] Notice that the word ‘Repugnant’ originates from the local view, hinting that the author (Derek Parfit) and probably most of us, ineluctably have some part of our heart in the local, person-affecting domain.

In principle, the more specific a law or regulation is, the fewer it should affect. Codes of conduct are specific for companies or communities. Rights are specific for societies. Appealing to universalisability, Kant's categorical imperative is closer to the universal view. Going further in this direction, consequentialism can be applied everywhere, if the agent is knowing and unbiased enough.

Two concepts help us decide how local a view to adapt. First, according to the assumed principle presented in the introduction, the smaller the portion you are affected by some action, the more you should tend toward the universal view when making that action. I will call this case "personal insignificance." Second, according to the principle that unintuitive problems require computing tools, the local view is better suited when our brains are suited for it. I will name this situation "personal maladaptiveness."

2. The problem

The locality problem arises when the conditions infrequently allow for the local view. If our brains are selfish and simplifying then personal insignificance and maladaptiveness are the markers of such conditions. As an example of personal insignificance, consider a pilot flying a large aeroplane. According to the principle that utility scales with "extent of sentience" (which, in this case, can be approximated as the number of people affected) a pilot should protect the life of the passengers even if that would mean extra work for him, temporarily depleting him of, for example, the chance to take a quick nap. When he later has entered his hotel room, his desire for rest rightfully dominates his duty to protect the same people that were on the plane earlier, even if they would be at a greater risk of harm now. In his hotel room, his actions doesn't affect other people, and he can therefore behave according to the local view.

Figuratively, I argue that emerging, revolutionary technology creates a world with more and more pilots flying bigger and bigger planes, forcing us to choose between egoism and self-abnegation, with increasingly severe consequences. A first example of this is that ardent social media posts affect many more than just the friends the user otherwise would venture their dismay toward. A second example is managing one's effect on the global climate. A third example is buying groceries: collectively, consumers steer conditions for the production and handling of food, and as such have an indirect effect on farmers and animals.

The third example is particularly interesting because it is a collective action problem. They arise when the system we partake in (e.g. food consumption) is so complex and of such a large scale that we can argue for no individual action (purchase) being morally wrong because no individual action (purchase) makes any meaningful difference. Of course, the problem arises from the mismatch between our intuition

and reality; individual actions must, somehow together, produce the total effect. The equation is made complicated both by the large number of participants (leading to personal insignificance) and the non-linearity of the system (leading to personal maladaptiveness). When considering such situations, people are often - in an ethical sense - personally maladapted.

As a more direct example of personal maladaptiveness, consider automation. Computers are much better at understanding data than human values. Therefore it is much easier to optimise some parameters and ignore others. Upgrading computers clearly improves the efficiency of handling valuable data, but it also, unclearly, forces humans to tackle choices that we are maladapted for. Here, the locality problem is the problem of how to avoid making analytically good decisions, given the visible data, which would create a world where any subsequent decision is bad. Alternatively, it is the problem of when the biases or preferences that make up our values become vices rather than virtues. This risks happening when the agent's personal preferences contradict the preferences of a bigger source of value (such as a bigger group of people), or simply when the consequences of the choice are important enough.

In summary, the locality problem arises because of the intrinsic link between people's innate striving for improvement and the consequent necessity to depress our desires. Undoubtedly, humans have a proclivity to improve, strive forward and reach for our desires. We want to do what is subjectively good. Indeed, whenever the "objectively good" it is merely as the subjective goods that everyone agrees on. But the better we become this, the greater and more complex the consequences become. Decision makers become personally insignificant and maladapted. Now, for safety reasons, and respect for moral patients, the decision makers should replace the local view with the universal view. The danger in this is that values are fundamentally subjective and only exist in the local view. Thus, actions affecting large-scale and complicated systems call for ignoring our desires. Commonality of such actions is therefore both dangerous (because it subjugates personal values) and desired (because striving for more power appears to be an universal and instrumental goal).

The locality problem can be illustrated through the following set of thought experiments contrasting the local view with the universal view.

<u>Thought experiment 1 - The classical trolley problem</u>
A runaway trolley is following the tracks without being able to stop. Further down the track five people are lying, tightly tied to the track. You are standing next to a lever. By pulling it, you can divert the trolley onto a sidetrack which has one person identically positioned on the rails. Do you pull the lever?

Most people would flip the switch in this example, saving the most lives.[3] Yet some people would not do this, because of the emotional consequences of actively deciding on killing that one person lying on the alternative track. To amplify this emotion, let us imagine that the person is someone we deeply care about.

Thought experiment 2 - The personal trolley problem
With the same setup as above, you notice that the person lying on the side rail is your mother. Do you pull the lever?

If this would change your mind about pulling a lever, that should contain some information about your values. In the first dilemma, why is it better that five people survive than one? I would argue that it is because people, on average, are expected to feel positive emotions: they have values that can be satisfied. Most often, one of these values is their relationships: they can care and love. So, when the utilitarian suggests that it is best to save the most lives, is it not the stimulation of values causing utility such as happiness, joy and meaning, that he wants to preserve rather than the lives themselves? In other words, is it not utility itself rather than the vessels of utility he wants to maximise? I think that caring for our fellow human beings is one of the things we want to preserve when pulling the lever. In short, we should try to both maximise utility and preserve the opportunities for it to be maximised, and it seems utility requires values in the same way values require life. The locality problem arises when we maximise utility without maintaining the chances for new utility to be produced.

I can now present a twofold critique to the simplified utilitarian view that is so often prescribed to the classic trolley problem. First, true utilitarianism is only about maximising utility and human lives have no intrinsic utility. Second, true utilitarianism is impartial in time and is therefore concerned about maintaining the chance for future utility to be produced. If the existence of values is connected to the local view, this means that utilitarianism could actually advocate not needing to be utilitarian. Of course, the more that is at stake, the more difficult it is to argue for directly “worse” decisions. The motivation for my writing this essay is that I predict a future where consequences become larger, making it increasingly difficult to argue for subjective decision making.

3. The values we lose

Moral intuitions have developed for the benefit of the group. The problem of “I versus us” was largely solved by automatic common-sense morality including biases, emotional reactions and subconscious processes. Examples include the odium toward acting violently, the uncontrolled blushing when we do something awkward,

[3] For statistics and a deeper explanation of why this is, see Greene, J. (2013). Moral tribes: Emotion, reason, and the gap between us and them. Penguin Press.

and predilections displayed in implicit association tests. Global interconnectedness creates a new problem of “us versus them”, contrasting any sect of humanity with all others.

The utilitarian solution is an unintuitive, culture-independent meta-morality based on reason. In the same way that theories have to become more objective when applied to a greater number of people, the moral view has to be more universal in this situation. On this ground, science must assume the existence of an external reality to motivate the belief that our laws of nature can converge to the truth.[4] Likewise, ethics must assume moral realism for universal views to make sense. On a theoretical level, objectivity must mean both general theory of relativity over classical mechanics, and utilitarianism over "doing what you feel like." On a factual level, objectivity must mean both that stones generally fall down rather than up, and that it is generally bad to hit someone. This does not imply that scientific facts, or good moral decisions are situation-independent, but it does mean that every situation can be analysed rationally for a better moral decision.

In line with this, Derek Parfit writes that “we may need to make some changes in the way we think about morality. … Common-Sense Morality works best in small communities. … Until this century, most of mankind lived in small communities. What each did could affect only a few others. But conditions have now changed. We can have real though small effects on thousands or millions of people.” Parfit points out that our common-sense morality is maladapted to handle cumulative effects. What goes unsaid in his quote is the loss of values. Only when both the necessity for meta-ethics and the problem with it are understood can one understand the locality problem. In this section I will explain that the danger of choosing the universal view over the local view is that it requires the suppression of the biases and preferences that make up our values.

Historically, suppressing biases has become increasingly important. One example of this is racism. In opposition to race impartiality, racism is an evaluation based on the race of a person. Much like why young children prefer toys from people who speak their dialect, looks have been a historically important criteria to determine in-group from out-group. Immigration, tourism and global cooperations has today made this bias unfair. However, few people would see the resistance to racism as something bad. What I would like to draw attention to in this essay is, however, the mechanism behind this history. Slavery and colonisation were not motivated by impartiality or equality. Rather the opposite: it required ambivalence toward a great deal of racism and inequality to take place. Yet, the increased efficiency to act in a racist manner made it less tenable, in the long term, to do just that. When the slave trade had gone

[4] This does not mean that science finds truth, only that it systematically approaches it (if the assumption holds true).

on for a long enough time, it became harder to refute the misery that had been caused, and resistance by the exploited people became stronger.

But was the subjugation of racism really a suppression of value and not just a change of value? Instead of appreciating certain homogenous groups, do we now not simply appreciate diverse groups? I would say yes. And the reason is that we are not racially impartial yet. Both a racist and someone following contemporary anti-racist ideals would derive a level of value solely out of the race, ethnicity or background present in a group. Impartiality would mean (fundamentally) indifference towards this. An optimally organised group would maximise utility, which can only be, weakly or strongly, correlated with one or many group arrangements. Creating this group should therefore, ideally, engage complete impartiality.

High impact actions require responsibility or control. Limiting the speed limit and usage of fossil fuels only became necessary after it became possible, and repressive after it became feasible. If (some) actions are emotional expressions, then restricting them restricts our self-expression. This is very different from never having had the opportunity to perform the action in the first place. Let me thus ask: is it possible that when we give ourselves increased opportunities for freedom, communication, power, etc., acting on this becomes so dangerous that it is wrong? Value resides in self-expression. The locality problem is particularly difficult because it impedes at least three ways of expression: freedom of speech, freedom from oppression, and freedom of action.

Each of the three mentioned freedoms is based on a fundamental need. Beginning with freedom of speech, the matching central need is to be heard or understood. As words, mostly through the internet today, inadvertently travel to people that may misinterpret and be hurt by them, we add responsibility and complexity to this essential process of self-expression. In the U.S. the portion of people who feel that they can't say what they want to has tripled over the past 60 years. Cancel culture is another example of a both necessary and "need-inhibiting" phenomenon, resulting from this propagation of words. So, making communication more available consequently diminish the freedom we value it for.

Similarly, the freedom from oppression and freedom to do what one wants follows from the needs to feel autonomous and employ our personal capacities. An obvious side-effect of mitigating coercions and bringing meaning to our skills is the increased risk for abuse and malice. Surveillance and governance is a straightforward counter force against people who would harm others, intentionally or not. For individuals, more freedom increases the demand for considering those around them. In this way, the effect of communication technology can be generalised to freedom itself: our process of making freedom more attainable also makes it harder to attain it safely.

I must here repeat my counter argument: changing lives also adds new values. For instance, the IT revolution has made many people value data much more, which can be shown through the increased data obtained through photos and smartwatch sensors, for example. Video games and career opportunities have also provided new details and strategies to appreciate. While this is true, the locality problem is not so much about changing values *per se*, but rather their overall decimation. That means that even these newly acquired values can only exist for as long as their implementation does not clearly lead to systematically and significantly worse outcomes for people.

When something other than the process of deciding is to be maximised, appeals to stories is at best equal to appeals to reason, but usually worse. This type of situation, therefore, could demand as clear and objective an understanding as possible. That would mean that religion, mysticism, and fictions regarding free will and the self become damaging. The tragedy of this is that these stories act as personal sources for meaning and value, despite leading to less rational decisions.

In conclusion, for values to matter - for us to matter - we need to afford to make the wrong decisions. If actions are much less local than our values, we cannot afford this. Then the subjective (decisions) would have to become more objective.

4. Looking for a solution

Underlying mechanisms or driving forces are complex; it is difficult to identify one point of causation that should be addressed. In this essay, I have assumed a causation starting with curiosity, that, in combination with the scientific method, leads to knowledge. Creative use of this uncovers opportunities. If the will is strong enough, the opportunities will be acted upon, which leads to development and change. Due to the character of motivation, this tends to make something that is desirable more easily attainable. If other people are affected, directly or indirectly, which is often the idea, this results in increased interconnectedness and larger communal effects. This means more responsibility. In order to avoid discrimination between groups, the responsibility say that we should prioritise rationale over emotions. More and more, emotions become superfluous or even damaging. Naturally, this produces a sense of us as being bad for the world.

Here a potential solution appears: reducing the effect on others. Playing video games, watching movies, reading books and going to concerts effects, mostly, ourselves. Isolating activities or, at least, making them more local can be done for social media as well. One such method would be to restrict uploads to only the nearest friends. More generally, subjectivity could be kept in fictional playgrounds such as games and online, but only if the playground were safe (meaning it was

isolated against the external world). Local decision-making is best in local environments.

However, while isolating actions might work to decrease the problem, it will not solve it. In the case of social media, influencers and interesting material attracts viewers even when they would be happier not viewing it. Other actions, like releasing carbon emissions or developing runaway AI, cannot in any clear way be made to have only local consequences. Moreover, participation in systems with these dynamics is likely to be practically mandatory. I see two reasons for this. First, someone not participating could be left so much worse off that he would not be able to stand the inequality. For instance, in a world where brain-computer interfaces increase people's IQ to 200, someone without the interface would not be able to compete. Second, not using the product could appear obviously immoral. For example, if self-driving cars were to dramatically reduce the number of car accidents, exactly how the ethical decision making is programmed is a secondary problem. Only afterwards might we realise that we value (potentially unpredictable) autonomy. Returning to person-driven cars could then be structurally infeasible, especially if other people do not share your concerns.

Simple consequentialism fails to distinguish between a value that is satisfied when making a choice and one that is satisfied after the choice is made, as a consequence. This is a crucial distinction to make if the existence of one value relies on its need. In that case, values that guide decisions will vanish even if they are stimulated following that decision. Just like the permanent destruction of all life is especially bad since it eradicates all potential for creation of utility, so the quenching of values removes one channel of utility production.

The problems of collective impact and personal maladaptiveness arise from a mismatch between human cognition and reality. Thus, individuals can contribute to a solution by not just changing their own individual behaviour but also the system in which they operate. Apart from the eventual shielding of activities it is to me unclear how this could be done without defying the drive for progress. To the extent that this drive is caused by societal structures, such as capitalism, a path of defiance might help against the locality problem.

A general principle of precaution could be advised. Problematically, this is more difficult to follow when the risk is less apparent. Unpredictable side-effects are thereby often not treated with the same level of precaution. In the case of complex new inventions that affect complex societies this appears more relevant than ever. Anyone calling for a search for knowledge rather than inventions, should realise that knowledge in combination with creativity leads to opportunities that can harbour fantastic promises and hidden dangers. I therefore imagine that there is special value in philosophy, ethics, art, psychology and neuroscience as these are the

exploration and clarification of thought, moral considerations, emotions, psyche and brain, respectively.

There is another point to be made: the locality problem is a paradox. In Section 2 I wrote that commonality of powerful actions is both dangerous (because it subjugates personal values) and desired (because striving for more power appears to be an universal and instrumental goal). Essentially this is the instrumental convergence thesis: regardless of the end goal, instrumental goals will always converge to self- and goal preservation, and power. However, if the goal is local (it exists in the subjective experience that is intrinsically linked to experienced behaviors), then the instrumental goals seem to contradict the end goal. Pursuing the instrumental goals (of using large-scale and complex systems) renders the end goal (of increasing happiness) more difficult. This is not a complete objection to the instrumental convergence thesis because the contradiction holds, at first, the other way too: not pursuing the instrumental goal (of increasing our perceived chances of happiness) renders the end goal (of more happiness) more difficult.

The only solution here is to change our intuitions or beliefs. First, if the large-scale and complex systems become intuitive then our moral values may adapt so that we can act instinctively as important decision makers with the same consequences as adapting a careful analysis first. I deem this highly unlikely. The second method is to stop believing that increasing the scale or complexity of an action will result in more happiness. This also seems unlikely to me. If I am correct about this, then the locality problem is a ticking time bomb which can be slowed down by the principles of isolation and precaution, but will eventually oppress all sources of human happiness. Of course, as the example of racism shows, we can change our values a lot and postpone their eradication.

5. Conclusion

The locality problem asks us to choose between bad decision-making and the suppression of values. This problem rests on the claims that values emerge from subjectivity, or the local view, and that good decisions are, more and more, made by appealing to objectivity, or the universal view. Apart from recognising the connection between less reliance on subjectivity and less attachment to emotions and values, two general strategies are proposed. First, shielding and isolating environments leaves subjectivity with greater relevance. This is based on the principle that local decision-making is best in local environments. Second, general precaution is advised, specifically in the methods of philosophy, ethics, art, psychology and neuroscience.