Preprint; final version forthcoming in *AI & Society*'s Curmudgeon Corner

Humansplaining: Is it a Thing? Is it bad?

Sanna Hirvonen[*]& Robert Michels[†]

April 18, 2025

Mansplaining happens when a man unnecessarily lectures a woman who is an expert on a particular topic because he unjustly assumes that he has more expertise with respect to the topic based on her gender. Mansplaining may not be one of society's greatest problems, but there is little doubt that it is bad. We may distinguish at least three reasons for why it's bad, one moral and two pragmatic.

Mansplaining is morally bad because a mansplainer treats his conversational partner unjustly. He puts himself in a position of epistemic authority over her without any good reason to do so.

The first pragmatic reason for the badness of mansplaining relates to the success of the conversation in question. As a speech act, it violates the pragmatic norms for effective communication in a conversation, in particular, the norms captured by Paul Grice's maxims of quantity and manner. The maxim of quantity states that a contribution to a conversation should be as informative as required, but not more informative, unlike the mansplainer's attempt to explain a topic to an expert. The maxim of manner requires one to avoid unnecessary verbosity, i.e. to avoid exactly what makes the mansplainer's contribution a lecturing.

A third reason is that mansplaining potentially leads to worse outcomes of the conversation, since it may prevent the female expert from contributing and shaping the concrete consequences of the conversation.

[*] LanCog, Centre of Philosophy, University of Lisbon; hirvonen.philosophy@gmail.com

[†] LanCog, Centre of Philosophy, University of Lisbon; mail@robert-michels.de

Enough of mansplaining; our topic is rather the possibility of humansplaining. We take humansplaining to happen when a human unnecessarily lectures to an AI that has expertise on a particular topic, because they assume that the AI lacks this expertise due to it being an AI.

The first question is, does humansplaining happen? With respect to current AI, it seems that the answer is no because interactions between humans and AIs

a) can not amount to explanations in the relevant sense, and

b) current AIs are not the right kinds of subjects to which an explanation could be directed.

Regarding point a), consider Reinforced Learning with Human Feedback (RLHF) which has been applied in training for example OpenAI's ChatGPT. During the training phase, humans interact with the AI by providing expected input for fine-tuning a supervised policy model and ranking sets of sample outputs to generate a reward model. Post-training, they evaluate the performance of the trained model, rating its output based on a number of criteria, and use it to generate text. Now, an explanation of the kind which would count as humansplaining should transmit general information; it should provide answers to 'how'- or 'why'-questions about the relevant subject. The inputs that humans give during the training phase and post-training evaluation in RLHF are not of this kind, since they consist either in specific examples or in evaluations of such examples. Prompts given to ChatGPT or other chatbots also do not qualify as explanations; they are rather instructions to generate a certain desired output.

Perhaps the more important point is that current AIs are not advanced enough to qualify as subjects, either in the intellectual or moral sense. An explanation has to be directed at a subject that can in principle understand (in the sense of being able to grasp linguistic meaning) and be convinced by it. Current AIs are arguably not yet there. Furthermore, even if explaining was possible, current AIs clearly do not count as agents to which moral considerations apply.

Things get more interesting once we look beyond the current state of AI and consider potential cases of humansplaining on the level of artificial general intelligence (AGI). AGIs are hypothetical future AIs that can perform at least as well as humans on a wide range of cognitive tasks. As of yet, it is unclear whether AGI is possible at all, however given the prevalence of discussions of it, we take it to be worthwhile to consider a scenario in which it is realized in a speculative manner.

Assuming that such scenarios are possible, interactions between humans and AGIs can take the form of explanations, since AGIs are, by hypothesis, subjects at which they can be directed. Hence, humansplaining becomes an issue, which raises the question of whether it is bad.

Our answer is that it depends. In particular, we distinguish between three kinds of topics of explanations which make them more or less amenable to cases of humansplaining. First, we assume that even in the age of AGI, there will be some topics in which humans retain a necessary advantage in expertise compared to AGIs. A plausible example is given by topics that are essentially connected to human experiences, for example gustatory taste. Gustatory tast is a prime example of embodiment in that tasting food is massively multisensory. Tasting is also not purely sensory but incorporates emotions, memories, and thoughts into the experience. Moreover, tasting relates to fundamental bodily states necessary for survival, such as hunger. An AGI could create potentially great recipes, but because it cannot taste the resulting dishes, it will lack the relevant gustatory experiences. Thus it can't form evaluative judgments about the dishes which puts it at a clear disadvantage in expertise. Concerning such topics, humans do not run the risk of humansplaining since AGIs simply cannot qualify as experts.

The second class of topics are those that do not essentially concern human experiences. With respect to them, humansplaining is possible, and plausibly problematic in the same ways as mansplaining. Given AGIs' (assumed) intellectual capabilities, it is possible that they can be treated unjustly by humans who assume to have epistemic authority over them. Humans may not always know better!

Furthermore, we can assume that humans and AGIs will be able to successfully engage in human-like conversations where the Gricean maxims apply. A potential example can be drawn from research on automated scientific discovery in physics. In this area, researchers have recently made significant progress in training AIs to re-discover known physical laws from either prepared physical data or in some cases directly from videos showing relevant physical processes (see https://www.quantamagazine.org/machine-scientists-distill-the-laws-of-physics-from-raw-data-20220510/). Some research even suggests that AIs may be able to find entirely new laws involving new state variables (see e.g. https://www.nature.com/articles/s43588-022-00281-6). This may point to a future in which fully autonomous 'automated scientists' discover alternative laws of physics, potentially leading to new and better physical theories. We can easily imagine a conservative human physicist who engages in humansplaining in this scenario, because they insist on the traditional, tried and tested 'human-made' physics.

What would make humansplaining bad in such cases? First, the human would violate the conversational maxims of quantity and manner for the same reasons as someone who is mansplaining: the AGI may not have needed any of the explanations, being able to figure things out for itself (or perhaps, having already figured things out in a better way). Second, humansplaining may lead to worse practical outcomes; in our example, we can imagine that without the humansplaining, the AGI could have come up with new and better physical theories. Put more generally, if an AGI's potential contribution to a conversation is drowned out by a lecturing humansplainer, then relevant ideas may remain unstated, and related actions leading to e.g. concrete social or political improvements may remain untaken.

The third case we want to consider is an inversion of the first: There may be topics concerning which AGIs necessarily have more and different kinds of capabilities compared to humans, for example due to their more powerful abilities to compute and store information. Regarding such topics, humans would be at an increased risk of humansplaining due to the inherent imbalance in expertise.

One interesting thing to note is that there may be a principled asymmetry regarding the imbalances between humans and AGIs with respect to topics of the first and third kind. It may be that with advancing technology, AGIs can be equipped with say, artificial or simulated noses and mouths, modelled on our abilities to perceive flavours. This could bring their gustatory expertise to our level. But there seems to be no comparably promising pathway towards equality of expertise regarding topics of the third kind: We are hopelessly inferior in, for example, processing information. Indeed, we may even ask whether, by parity of reasoning, we should also consider potential cases of AIsplaining: Will advanced AIs at some point be able to patronize us by ignoring our expertise on a topic and lecturing us about it?

To wrap up: If an AI is advanced enough to be a genuine partner in a conversation, then humansplaining becomes an issue. Fortunately, just as with mansplaining, there is a rather simple solution to avoid the problems it may cause, namely a good dose of epistemic humility. To be a good, cooperative partner in a conversation, we should not in general assume that we know better than an AGI interlocutor, just because they are not human. Given that the differences between humans and AGIs will likely be much greater than among humans, epistemic humility is definitely called for. Human hubris should not stand in the way of fruitful human-AGI collaboration.[1]

[1] R.M. acknowledges financial support by the FCT (https://doi.org/10.54499/2021.03171.CEECIND/CP1702/CT0015, https://doi.org/10.54499/2023.15136.PEX).