On *Against Purposeful Artificial Intelligence Failures*

James D. Miller, Smith College, Northampton, MA USA
jdmiller@smith.edu

The topic of the paper is highly relevant, and the analysis presented is well-conceived and articulated. However, there are several areas where the paper could benefit from further refinement to enhance its impact and clarity.

The paper would benefit from a focused discussion on the impact of nuclear accidents on societal and political attitudes towards nuclear power. Specifically, it should detail how such accidents have not only significantly reduced public support for nuclear energy but have also led to a marked decrease in investment in this sector. Analyzing these incidents provides a clear illustration of how technological mishaps can influence public perception and subsequently affect policy and investment decisions.

The author should consider whether cyber criminals have had the unintentional effect of strengthening cyber security in a way that promotes AI safety.

The use of the term "fringe" in the discussion should be reconsidered. This term may unintentionally convey a sense of belittlement toward the ideas being discussed, which can detract from the objective of presenting a robust critique through steelmanning.

The author is known to hold a very high P(doom). Given this belief, shouldn't the author be in favor of us taking wild chances even if they are unlikely to work? Might it not be rational to take significant risks or even encourage purposeful failures as a means of breakthrough or change if you have a P(doom) as high as the author does? The arguments in the paper all make sense if you have a P(doom) of, say, <70% but not if above 99.99%.