Times that Eliezer has mentioned dying with dignity in the late 2021 MIRI conversations:

Eliezer Yudkowsky
The first reply that came to mind is "I don't know." I consider the present gameboard to look incredibly grim, and I don't actually see a way out through hard work alone. We can hope there's a miracle that violates some aspect of my background model, and we can try to prepare for that unknown miracle; preparing for an unknown miracle probably looks like "Trying to **die with more dignity** on the mainline" (because if you can **die with more dignity** on the mainline, you are better positioned to take advantage of a miracle if it occurs).
https://intelligence.org/2021/11/11/discussion-with-eliezer-yudkowsky-on-agi-interventions/

Before that AI grasps the big picture and starts planning to avoid actions that operators detect as bad, there will be some little AI that partially grasps the big picture and tries to avoid some things that would be detected as bad; and the operators will (mainline) say "Yay what a good AI, it knows to avoid things we think are bad!" or (**death with unrealistic amounts of dignity**) say "oh noes the prophecies are coming true" and back off and start trying to align it, but they will not be able to align it, and if they don't proceed anyways to destroy the world, somebody else will proceed anyways to destroy the world.

https://intelligence.org/2021/11/22/yudkowsky-and-christiano-discuss-takeoff-speeds/

In the Overt Plotting Phase, which is not the main phase you're asking about, the AI is visibly plotting to take over the world and hasn't realized it ought to hide the fact. In the default expectation where we die with very little dignity, the operators smile to each other and come up with a rationalization for why it's totally fine to proceed, either with or without tossing on some kind of fig leaf like training away the visible manifestations of failure. I am not going to predict the particular rationalizations and arguments for proceeding anyways, because I don't want to give them even more ideas.

…

Operators on the mainline, **dying without dignity**, will say, "Oh, yay, it stopped plotting, the latest corrigibility training intervention we tried must've totally worked!"

The Law of Even Less Dignified Failure suggests that in fact they will not be trying any corrigibility options and will assume the AI just got smart enough to be nice; or that they will have shrugged about the AI's earlier antics and not think much of the disappearance

of those antics, since this is a way to **die with even less dignity** and before getting a chance to fail in a more interesting way.

Going in the more improbable direction of **death with greater dignit**y, if we have somehow achieved vastly vastly more transparency into the AI's thoughts than is possible with present ML technology, and if the AI models the operators as modeling its actions before the AI models the operators as having that transparent access to its thoughts, we might get to explicitly see the AI thinking about how the operators model its actions and conforming those actions in such a way as to manipulate the operators.

…

Operators on the mainline, **dying without dignity**, will say, "Oh, yay, it stopped plotting, the latest corrigibility training intervention we tried must've totally worked!"

The Law of Even Less Dignified Failure suggests that in fact they will not be trying any corrigibility options and will assume the AI just got smart enough to be nice; or that they will have shrugged about the AI's earlier antics and not think much of the disappearance of those antics, since this is a way to **die with even less dignity** and before getting a chance to fail in a more interesting way.

Going in the more improbable direction of **death with greater dignity**, if we have somehow achieved vastly vastly more transparency into the AI's thoughts than is possible with present ML technology, and if the AI models the operators as modeling its actions before the AI models the operators as having that transparent access to its thoughts, we might get to explicitly see the AI thinking about how the operators model its actions and conforming those actions in such a way as to manipulate the operators.

…

A way to **die with less dignity** than that is to train directly on what should've been the validation set, the more complicated domain where plots to kill the operators still seem definitely detectable so long as the AI has not developed superhuman hiding abilities.

A way to **die with even less dignity** is to get bad behavior on the validation set, and proceed anyways.

A way to **die with still less dignity** is to not have scaling training domains and validation domains for training corrigibility. Because, like, you have not thought of this at all.

https://intelligence.org/2021/11/29/soares-tallinn-and-yudkowsky-discuss-agi-cognition/

I am not shocked by the AGI stuff being a gigantic megaproject

it's not above the bar of survival but, given other social optimism, it permits **death with more dignity** than by other routes

https://intelligence.org/2021/11/25/christiano-cotra-and-yudkowsky-on-ai-progress/

I replied asking if Gwern's 3.8x estimate sounds right to them.

A 10x improvement could power what I think is a jumpy AI timeline. I'm currently trying to draft a depiction of what I think an **unrealistically dignified but computationally typical end-of-world would** look like if it started in 2025, and my first draft of that had it starting with a new technique published by Google Brain that was around a 10x improvement in training speeds for very large networks at the cost of higher inference costs, but which turned out to be specially applicable to online learning.

https://intelligence.org/2022/03/01/christiano-and-yudkowsky-on-ai-predictions-and-human-intelligence/

vaguely plausible rough scenario: there was a big ongoing debate about whether or not to try letting the system trade stocks, and while the debate was going on, the researchers kept figuring out ways to make Something Zero do more with less computing power, and then it started visibly talking at people and trying to manipulate them, and there was an enormous fuss, and what happens past this point depends on whether or not you want me to try to describe a scenario in which **we die with an unrealistic amount of dignity**, or a realistic scenario where we die much faster

I shall assume the former.

https://intelligence.org/2022/03/02/shah-and-yudkowsky-on-alignment-failures/