	README: This is	**Categories** (explanati	**What can we learn from this ris**	**Outer/Inner/Intent**	**Sources of misalignment**		**Deception**	**Power-seeking**	**Capabilities**			**Existential risk**
	Takeaways												
	Typical use case:	Brief summary >			*What happened in the code?*	*Why was bad code written?*			*Propensity to have an objecti*	*Why can it influence the physi*	*What does the AI need to be able to think?*		
	Risk model	Brief summary v			Technical causes	Social causes			Agency	Power	Cognition		
		Description		An issue with the... Intent: The AI has nice intentions, Outer: A bad objective is put into t Inner: The AI reaches a bad goal i NA: The inner/outer framework do	Specification gaming: The AI us Goal misgeneralization: The trai Convergent instrumental goals: Crystallized proxy: The AI has a Any: Either cause would enable	Bad actors: Bad/careless peop Competitive pressure: AI deve Irrelevant: Social causes do no None: There need not be a so ? : A technical cause could not			YES --> Agentic: The AI pursu NO --> Not necessarily agenti	Received: The AI is mostly giv Acquired: The AI mostly acqui ? : The author does not specif	Agentic planning: Ability to make plans to reach Long-term planning: Ability to expect long-term c Situational awareness: Ability to understand info Strategic awareness: Ability to understand inform General purpose AI: Ability to be effective for a w	1: Inability to affect the AI. 2: Disempowerment. 3: Extinction. 4: S-risk. The number reflects the minimum level guaranteed
Class of scenario	**If you only read one thing, let it be this row.**	My quick takeaways	General takeaway: There are many paths to AI X-risk. Some are more likely or dangerous than others, but none can be neglected if we are to avoid disaster.	Along the many scenarios, misalignment comes from many places. No single solution will solve everything, though inner misalignment is significantly more common than other types.	It is not enough to have a solution for a specific cause, as there are many technical reasons for misalignment.	1- The most common cause of misalignment is competitive pressure. Our society *must* learn to solve coordination problems. 2- Even if we coordinate perfectly, there are technical hurdles that could still lead to failure.	There is a strong correlation between deception, power-seeking and agency.		1 - Ensuring that AIs are not agentic would go a long way towards preventing most AI takeover scenarios. 2- But even if no AI were agentic, a disaster could happen because of societal failure.	Most scenarios assume that an advanced AI will be able to seize power, so denying it access to resources such as banks and factories is not enough of a safeguard.	AI-takeover-like scenarios usually require specific capabilities that we can detect in potentially dangerous AIs through capability assessment and interpretability. In many scenarios, these capabilities come from general purpose training and are not externally visible, so it is a misconception to assume that "we'll see it before we get there".	Yes, AI *is* risky. While not all scenarios are guaranteed to lead to human extinction, it is important to remember that even if we survive, disempowerment and other societal issues will need to be addressed. There is much uncertainty about the severity of the risks.
"Society matters"	Christiano1 - You	We make AIs to optimi	Intent alignment is a thing. Societ cannot be negliged to solve align	Intent	Specification gaming	Competitive pressure	☐	☐	☐	Received	?	2
	Christiano2 - Influ	AIs that try to acquire		Inner	Convergent instrumental goals	Competitive pressure	☑	☑	☑	Acquired	?	3
	Critch1 - Producti	AI enables processes	Identify control loops as points of to prevent a Critch scenario.	Any	?	Competitive pressure	☐	☑	☐	Acquired	?	3
	Critch2 - Flash Ec	AI enables processes		Any	?	Competitive pressure	☑	☑	☐	Acquired	General purpose AI, agentic planning	3
"Mining too deep"	Hubinger - How lil	The only AIs we allow	Deceptive alignment must be saf against. Do not just hope it won't -> security mindset? Distinguish high and low path-de	Inner	Crystallized proxy	?	☑	☑	☑	?	Situational awareness	?
	Soares - A centra	Capabilities will gener	ASI is coming. Don't expect it to t by default. We should figure out h an ASI at all (the hard bits of alig	?	Any	Any	☐	☐	☑	Acquired	General purpose AI	3
	Cohen et al. - Adv	Advanced AI strives to	Beware trying to tame something intelligent than you. -> Vingean re	Inner	Crystallized proxy	Any	☐	☑	☑	?	Agentic planning, situational awareness	1
"Skynet by default"	Cotra - Without sp	The world is on a path	We are on our way to a bad futur Action needs to be taken, for real - stop the AI arms race - don't use naive safety protocols - don't assume it'll never work too	Outer	Any	Competitive pressure, bad act	☑	☑	☑	Any	General purpose AI, agentic planning, situational	3
	Ngo - The alignm	By default, advanced A		Inner	Convergent instrumental goals	Any	☑	☑	☑	Acquired	General purpose AI, agentic planning, situational	?
	Shah - AI Risk fro	Searching for an effici		Inner	Convergent instrumental goals	Any	☑	☑	☑	?	Agentic planning, situational awareness	?
	Carlsmith - Is Pov	Power-seeking AIs tak	Be careful of power-seeking AIs, there's a lot to unpack.	Inner	Any	Competitive pressure, bad act	☑	☑	☑	Acquired	Long-term agentic planning, strategic awareness	2