AI and Ethics
https://doi.org/10.1007/s43681-025-00698-5

ORIGINAL RESEARCH

Human biases and remedies in AI safety and alignment contexts

Zoé Roy-Stang[1] · **Jim Davies**[1]

Received: 23 May 2024 / Accepted: 26 February 2025

Abstract
Errors in judgment can undermine artificial intelligence (AI) safety and alignment efforts, leading to potentially catastrophic consequences. Attitudes towards AI range from total support to total opposition, and there is little agreement on how to approach the issues. We discuss how relevant cognitive biases could affect the general public's perception of AI developments and risks associated with advanced AI. We focus on how biases could affect decision-making in key contexts of AI development, safety, and governance. We review remedies that could reduce or eliminate these biases to improve resource allocation, prioritization, and planning. We conclude with a summary list of 'information consumer remedies' which can be applied at the individual level and 'information system remedies' which can be incorporated into decision-making structures, including decision support systems, to improve the quality of decision-making. We also provide suggestions for future research on biases and remedies that could contribute to mitigating global catastrophic risks in the context of emerging, high-risk, high-reward technologies.

Keywords Cognitive biases · Existential risk · AI safety · Alignment · Remedies · Decision-support systems

1 Introduction

Cognitive biases are 'cases in which human cognition reliably produces representations that are systematically distorted compared to some aspect of objective reality' [1, p.968]. Political biases are systemic influences within social systems. Cognitive biases and political biases can be considered the joint sources of behavioural biases, which are reliable patterns of suboptimal or irrational decision making [2].

With the rise of AI developments and AI experts identifying concerning possibilities of catastrophic outcomes, including human extinction this century [3], it is worth thinking about how human biases could lead to inaccurate representations and suboptimal or harmful decisions that undermine the ongoing global project of aligning these cutting-edge technological developments with human intentions and values. This paper explores how human biases may affect people's perceptions of AI developments and risks from advanced AI. Throughout this paper, the general term 'advanced AI' refers to AI systems with narrow or general capabilities matching or surpassing those of humans. The term artificial general intelligence (AGI) refers to the more specific milestone where an AI system has general capabilities exceeding those of the average human across nearly all domains. We focus on how AI developers, safety researchers, and policymakers may be prone to biases and potential effects on their prioritization of resources and decision-making in contexts relevant to safety and alignment. Remedies and future research suggestions are explored.

Over 200 human behavioural biases have been identified [2], although the research lacks parsimony [4]. To list a few issues, there are different names for the same biases, it is unclear which underlying processes give rise to them, and a systematic, scientific way to cluster them is lacking. There are lists of biases, but the big picture of how they connect is unclear as the interrelations between biases remain largely unexplained [5]. Cognitive biases are sometimes clustered based on the heuristics that are thought to give rise to them [6], or levels of the social sphere (e.g., individual, interpersonal, and intergroup levels), or as arising from various psychological needs such as cognitive closure, self-esteem, and

✉ Jim Davies
Jim.davies@carleton.ca

Zoé Roy-Stang
zoe.roy.stang@gmail.com

[1] Department of Cognitive Science, Carleton University, Ottawa, Ontario K1S 5B6, Canada

Published online: 14 May 2025

social belonging [7]. A comprehensive review and scientific classification of biases is still missing. For this analysis we selected the most important biases related to advanced AI based on perceived relevance to alignment.

Eliezer Yudkowsky reviewed more than a dozen biases that could have devastating consequences in the context of assessing and responding to catastrophic risks [8]. Errors in judgment could be catastrophic in decisions that can drastically affect humans and other sentient beings. Further research on bias reduction and elimination in high-stakes contexts would be highly beneficial to ensure a safer future.

This paper extends Yudkowsky's work by covering a more extensive list of relevant biases and specifying how they could affect judgments in relevant contexts, with examples. We also cover associated remedies, including those from more recent research, and how they could be applied to the alignment of advanced AI in the coming years. Specifically, we address how AI developers, governance/policy teams, and the general population may exhibit biases in the context of AI developments leading up to and beyond AGI. This includes the general public's perception of AI-related risks, opinions on prioritization of resources, and decision-making or propensity to action. Based on the most potentially relevant biases identified, we suggest empirical studies to determine whether these biases apply as predicted in given contexts, as well as remedies that could potentially counter these biases in high-risk, high-impact contexts. Each section covers one bias and after a brief description, follows this general content format (in paragraph form):

- Potential effects on a person's perception of AI developments
- Potential effects on a person's perception of risks associated with advanced AI
- Potential effects on a person's prioritization of resources and propensity to action (decision-making), especially in the contexts of AI development, safety, and governance work
- Potential remedies
- Future research suggestions, including ways to test predictions of how biases could apply in relevant contexts

Remedies can be classified in four categories: broadcast, personal communication, information consumer remedies, and information system remedies. Broadcast remedies refer to ways that governments and NGOs can present information in ways that help their audiences reduce bias in their thinking. Personal communication remedies are similar, except the suggested actions focus on how other individuals can account for an audience's bias. Information consumer remedies are actions that people can take to reduce their own bias as they receive information. Finally, information system remedies are debiasing strategies implemented in decision-making structures, including decision-support systems and ways to present data within these systems. AI organization leaders, project coordinators, managers, and people working on improving institutional decision-making are probably best positioned to implement information system remedies at scale, though other workers may also be able to implement and test them in their workflows.

Throughout this paper, we will focus on information consumer and information system remedies that could be beneficial to the general public, AI developers, and policymakers in the context of improving the quality of reasoning and decisions related to AI safety and alignment. For an overview of broadcast and personal communication remedies, Lewandowsky et al. [9] provide an excellent graphical summary of findings on solutions to address common misinformation effects. Further research could investigate how broadcast and personal communication remedies can best be applied in AI safety and alignment contexts. For example, reducing bias in media and conversations could help bridge the gap between views on AI.

2 Methodology

This literature review involved consulting lists of human biases relevant to AI safety and alignment from multiple sources, including papers on biases in change management [10] and project management [2], as well as the Decision Lab's list of cognitive biases [11], and Yudkowsky's work on biases in existential risk contexts [8]. We selected biases that could influence how individuals and institutions perceive and think about unprecedented catastrophic risks from future AI advancements. This initial exploratory approach allowed for the identification of biases that might shape AI governance, risk perception, and decision-making in alignment contexts.

Following this broad survey, we conducted a targeted literature review on select biases and their remedies. As the research progressed, we refined the selection of biases by removing those for which we found less convincing evidence or could not identify meaningful examples to establish clear relevance to AI safety. This iterative process resulted in a curated analysis of biases that are more theoretically robust and practically significant for AI safety discourse.

To improve readability and organization following reviewer feedback, we used Claude 3.5 Sonnet (an AI language model) to assist with copy editing the structure of bias mitigation sections. Specifically, this editing involved reformatting text to include clear topic sentences introducing types of remedies (individual vs. systematic), while maintaining all original content, citations, and evidence.

The AI was used solely for readability improvements and formatting changes, with no generative editorial work or autonomous content creation. Both authors reviewed and verified that the restructured sections accurately reflected the original work. We also used ChatGPT (another AI language model) to help identify and correct minor typographical errors.

3 Human biases applied to AI contexts

3.1 Availability heuristic

The availability heuristic is the tendency to consider information that is readily available to the mind when making judgments [6].

The information available to people when they think about AI could lead to biased perceptions of AI developments. For example, if people have recently watched a movie about killer robots destroying the world, they might be more likely to perceive AI developments as dangerous, and vice versa if they instead watched a movie that depicted AI through an extremely positive lens.

The illusory truth effect (also known as the availability cascade) is the tendency to believe information that is repeated multiple times, even when it contradicts prior knowledge [12]. People tend to underestimate the effect of this bias, especially on themselves [13].

People are likely to perceive information about AI developments as true if it regularly shows up on their news feed, social media feed, or in day-to-day conversations. For instance, if people see many headlines about existential risks from AI, they may think that AI poses an existential risk, regardless of whether the information is based on evidence or expert opinions. If individuals are constantly exposed to bad news and doomsday predictions of AI developments, they may believe AI is more dangerous than it is. Conversely, if companies repeatedly assert and advertise that their products are robustly tested and safe to use, people could come to believe that instead, regardless of whether there are good reasons to think it is disinformation. In any case, seeing repeated information about AI and the risks of advanced AI may lead to individuals misjudging which issues are most pressing and how to allocate resources optimally to achieve desired outcomes.

In the context of existential risks, including those potentially associated with advanced AI, the availability heuristic might lead to underestimating unprecedented global risks as no one can recall an event that has led to the extinction of humanity [8]. Availability could negatively affect the decisions of teams of AI developers or policymakers if they only consider options that are immediately available to memory or obvious in a given context, failing to consider unprecedented outcomes. These people may have some initial ideas of how AI might lead to poor outcomes based on past experiences and knowledge, but they might fail to consider that these outcomes are not necessarily the most relevant or likely to occur. They might only act on ideas or plan for outcomes that came to mind initially, giving less consideration to other equally or more relevant outcomes. For example, if a product has some obvious potential harms that need to be addressed before launch, teams of developers might set narrow goals around these issues but fail to set safety goals broad enough to capture other harms worthy of prevention. This can lead to suboptimal strategic planning and resource allocation.

If there is a strong emotion associated with certain events, these events will be recalled more easily [14]. For example, the headlines “AI Detects Disease Faster” or “Killed by Generative AI” might be emotionally arousing and remembered more easily than neutral information. This is known as the affect heuristic. If people are asked about how dangerous they think AI developments are, their judgments of the commonness and probability of mishaps will likely rely more heavily on this salient remembered information, including emotionally arousing memories, than other events. This could partly explain the polarized views on AI in the general population and among AI experts. People evaluate equal negative outcomes differently depending on the cause and its associated emotional valence, which makes it harder to accept cost-benefit analyses [14]. This means that if the same negative outcome occurs as a result of two different AIs, one with positive valence and one with negative valence (in someone’s mind), risk perception and concerns will likely be higher for actions and consequences caused by the AI associated with the negative valence.

Multiple strategies can mitigate the influence of the availability heuristic and related illusory truth effects. A first remedy is to intentionally consider more options than the ones that initially come to mind when deciding what to work on or which solutions to implement. Generating more options is recommended in high-stakes contexts because the availability heuristic can lead to considering too few options before committing, which is a common, limiting mistake [15]. Thinking of only a few initial ideas to solve a problem before starting to work on an issue can lead to missing the opportunity to think of much better options (design fixation). In contrast, generating a longer list of possibilities or solutions before narrowing them down to make more in-depth evaluations is a low-cost option with high potential benefits, especially in the context of innovative projects that have a high potential for negative or positive impact.

Using carefully selected data in analyses while avoiding information overload can lead to better strategic

decision-making [10]. Decision-makers such as team managers or AI safety researchers are susceptible to the availability bias in the pre-analytical steps of choosing which analyses to conduct and what data to collect. Thus, this remedy could prevent the failure to identify more pressing risks or more promising research avenues. This remedy applies at the information consumer and system levels. For example, individual researchers or policymakers can think of many ideas before committing to a project, or an organization can require that teams of AI developers document multiple potential risks before selecting which capabilities or safety mechanisms to work on.

Individual and systemic remedies could also prevent or reduce illusory truth effects (such as availability cascades). First, in high-stakes decision-making contexts, people should avoid exposure to repeated, uncertain, or false facts. Fact-checking and related epistemic health practices can be applied at the information consumer and system levels given that information pools can be contaminated. Large amounts of misinformation or disinformation is known as epistemic flooding. Fact-checking, however, is costly [16].

Moreover, information consumers or platform managers can use systematic reminders to send debiasing reminder messages. For instance, warning people that the information they are about to receive could be misleading can reduce misinformation effects [9] (broadcast and information system remedy). Similarly, a study by Pennycook et al. [17] suggests that reminding people to consider whether information is true improves their assessments of the information's accuracy.

Finally, to prevent illusory truth effects in the general population, social media platforms or add-ons could track how many times a user has seen posts with similar messages and provide the option to reduce the number of times the algorithms show posts with repeated messages in the future (broadcast and information system remedy).

3.1.1 Order effects: primacy and recency

Serial position effects, such as primacy and recency, are order effects associated with the availability heuristic. Primacy bias refers to people's tendency to remember the first part of an event or the first item in a series better than other parts in the middle. Similarly, recency bias is the tendency to recall the last part of an event or list better [18].

People may be more likely to remember the information they received the first time they heard about current AI developments or the possibility of advanced AI, and this first impression may disproportionately affect their perspective. People's judgments of their first and most recent encounters with AI (e.g., on the news, in movies, or at work) might disproportionately influence their judgment of how much weight to assign to potential risks or benefits. In decision-making contexts, if important options or arguments for or against options are stated at the beginning or end of a communication, they are more likely to be remembered and acted upon.

Systematic reordering is an information system remedy for order effects. For example, changing the order of items in a list such as names of candidates on voter ballots can systematically cancel out the emphasis on the first and last items [18]. This remedy could be applied to recommender algorithms for videos or reading recommendations on websites designed to disseminate AI safety and alignment content to help learners form accurate opinions about AI safety and risk mitigation (e.g., a list of proposed technical solutions to specific AI safety concerns on a website or recommended sources for students to learn more about an area). Moreover, those who write reports in the governance space could present lists of options in a different, randomized order to each policymaker or board member involved to reduce the influence of order effects in collective decision-making.

To test whether primacy bias applies in this way in the context of AI perception, researchers could ask research participants in what context they first and last heard of the possibility of advanced AI, as well as instances in between, and rate them on a scale of most negative to most positive, then assess whether the first and last instances are more predictive of their current perception ratings of AI trajectories. It might also be worth investigating whether this perception influences people's likelihood of taking concrete actions such as sharing news articles, promoting or discouraging various tools at work, donating to AI safety initiatives, or voting for political representatives who prioritize AI safety issues.

3.2 Representativeness heuristic

The representativeness heuristic is a mental shortcut that involves relying on a prototype of a given situation when making a probability judgment under uncertainty [6]. Similarly to the availability heuristic, using the representativeness heuristic can lead to a restricted view of the landscape of possibilities, thus biasing judgments and decision-making.

When thinking about AI developments, people may rely on an example of AI to make judgments about developments in the whole field. For instance, if mishaps related to algorithmic bias make the news more than other mishaps for a given period, people might use this type of issue to represent problems associated with AI developments generally. This failure to account for progress and issues in various subfields and subprojects leads to an unrepresentative picture of the whole field, which leads to certain issues getting

disproportionate attention and resources. For instance, if a policymaker only thinks of a commonly used large language model (such as the current free version of ChatGPT) while making judgments about the safety of AI developments generally, they would fail to consider risks from a wider range of AIs, including possible future developments leading up to and beyond AGI. AI policies are more likely to cover relevant areas and effectively prevent harm if governance teams can properly account for the wide range of differences across AI systems and organizations.

Multiple remedies can help mitigate biases arising from the representativeness heuristic in technological forecasting and AI development. A study of operations managers found that cognitive training involving making people aware of representativeness biases (with a definition and example) yielded a significant reduction of relevant biases [19]. This information consumer remedy addresses several specific biases that tend to arise from the representativeness heuristic, including base rate neglect, insensitivity to sample size, misconception of chance, insensitivity to predictability, the illusion of validity, and the misconception of regression [6]. This remedy could be particularly valuable in AI safety and alignment contexts. For example, when making judgments of risks and capabilities, individuals need to consider a sample of AIs that includes the most recent cutting-edge developments across subfields.

Information systems such as cognitive training programs for AI researchers and policymakers could also present the representativeness heuristic with examples of its influence on AI risk perceptions. Other systematic remedies for representativeness bias in AI development focus on ensuring comprehensive consideration of reference classes when making judgments about technological progress. For general forecasts about technological developments, AI developers might fail to consider developments outside of their subfield of expertise if they rely on an unrepresentative reference class. AI safety organizations could implement systematic processes for tracking AI progress with a representative set of pre-established evaluations to avoid overweighing a subset of the capabilities relative to others in risk assessment and mitigation.

However, even with a representative sample of recent developments, outliers may not receive appropriate consideration in forecasts or capabilities and risk assessments because the AIs with the most exceptional and advanced capabilities may have orders of magnitude more impact (positive or negative) than the representative cases. It is especially difficult to accurately depict predictable future changes when a new technology consistently yields small benefits but also has a small chance of creating disproportionately large harms that have not yet occurred (see Sect. 3.4.1 for a deeper discussion of this issue).

3.2.1 Base rate neglect

One bias that was originally thought to arise from the representativeness heuristic is base rate neglect [20]. It refers to the tendency to ignore given information about the base rate frequency (also known as prior probability) of an event when judging the probability of outcomes. A more accurate conclusion based on later reviews is that participants tend to underweigh base rates, rather than ignoring them fully [21, 22]. Experimental evidence also shows that people's reliance on base rates changes based on the perceived trustworthiness of the base rate data [23].

As people hear about AI, for example, through conversations, they may construct an idea of what it is based on their experiences, and use the representativeness heuristic to make judgments and predictions. People may give more weight to a specific example of AI success or failure, neglecting less publicized but statistically significant risks or potential benefits, even if they are explicitly given those statistics. For example, when judging the probability of an AI project's success, people might fail to consider the typical success rate of similar projects. People's judgments of whether and when milestones (e.g., the creation of safe AGI) will be reached might not account for relevant base rates, such as the proportion of successful and unsuccessful previous, similar attempts and the success rate of subgoals. This could lead to underestimating or overestimating the complexity of the challenges involved in reaching these goals as well as associated risks. Neglecting to take the prior probability of developments into account can lead to allocating more resources to less pressing issues and fewer resources to more pressing issues.

Information system remedies to reduce base rate neglect include visual and interactive approaches to probability representation. Experimental evidence from Roy and Lerch [24] suggests that using graphs as visual representations of probabilities can reduce base rate neglect and lead to better judgments. Additionally, Hayes et al. [25] found that providing opportunities to draw samples from probability distributions and offering causal explanations for false positives (a signal that something happened when in fact it did not) improved the accuracy of probability judgments. Adding visual representations of AI risk scenarios or threat models could help policymakers understand the quantifiable nuances of different situations. Moreover, interactive tools to draw samples from a probability distribution of AI development trajectories could help teach students or practitioners how to think about possible futures (e.g., in AI safety courses or training programs).

3.3 Clustering illusion

People tend to perceive patterns or associations between two variables in random data. This bias is the clustering illusion. When interpreting random data, people's subjective probability judgments don't tend to align with probability theory, but some evidence suggests judgments align with the representativeness heuristic [26]. See the confirmation bias section for a similar discussion on the tendency to perceive inaccurate patterns.

People might think they see a pattern in AI developments and make inaccurate predictions based on a series of events that appear consistent but in reality are largely dependent on highly variable factors. This clustering illusion can lead to inaccurate judgments of the possibilities of different outcomes associated with AI developments, including highly beneficial or catastrophic ones. For instance, seeing nonexistent trends can give a false sense of security, and, as a result, people in charge of safety and governance may fail to identify the areas that need the most attention and resources. It could also give a false sense of danger from a coincidental series of mishaps, leading to similar resource allocation issues. Interpretability researchers or evaluators might see false trends in the internal working or external output of an AI system and waste time following intuitive but false leads.

The use of robust statistical tests is crucial to avoid this bias and correctly identify the presence of patterns of potentially risky developments or successful safety efforts. To test the effectiveness of encouraging this strategy in various alignment-related contexts, study participants could receive basic statistical tools, training, and relevant example problems. They could then decide to use the tools or rely on their intuition to make judgments about AI-related trends from a dataset of random data, and then rate the accuracy of their judgments. If the perceived accuracy is closer to measured accuracy in the experimental group than in a control group who underwent non-statistical training instead, this would be evidence of awareness of bias. If the accuracy of people who decided to use robust statistical methods is higher than the accuracy of those who did not, this would suggest that using statistical tests is an effective debiasing technique in the relevant contexts.

3.4 Hindsight bias

Hindsight bias refers to people's tendency to overestimate the probability of outcomes they believe have already occurred [27, 28]. It's also known colloquially as the I-knew-it-all-along effect, and it might lead to overestimating the predictability of future events [8].

After an advanced AI-related outcome has occurred, there might be a tendency to believe that the outcome was obvious or inevitable. If a person mistakenly thinks that they have a track record of accurately predicting past outcomes, they might be overconfident in their ability to foresee future AI developments. As such, public perception and trust in emerging AI technology may be unreasonably influenced by past successes or failures.

For AI developers learning from past mistakes, hindsight bias may lead them to think that an alignment strategy's failure was obvious and should have been avoided, resulting in harsh judgment of previous efforts, and potentially discouraging similar innovative or risk-taking approaches in the future regardless of how much sense they made with the information initially available to make the decision.

On the governance side, hindsight bias might lead policymakers to focus on addressing issues related to outcomes that have already occurred, assuming these events were predictable and preventable. For example, policymakers might focus on implementing responsible AI policies and guidelines around known harms such as algorithmic bias. In hindsight, it may seem obvious that algorithmic bias was going to be an issue, so they might be surprised that nothing was done about this issue in foresight. This perceived predictability of outcomes can lead to considering too few potential futures, risks, and solutions. Policymakers might fail to proactively anticipate other potential outcomes, such as global catastrophic risks associated with the race to AGI. They could also fail to consider and prepare for unforeseen challenges in AI. To avoid the dangers of having a reactive approach to policymaking in the context of emerging technologies, it may be helpful for AI governance teams to anticipate a range of potential futures, many of which may be unprecedented and not easily inferred from past events [29].

Multiple information consumer remedies can help reduce hindsight bias in high-stakes forecasting and decision-making contexts. One evidence-based approach involves maintaining foresight records of thought processes and predictions to consult after outcomes are known. Laboratory research has shown that providing participants with a list of their foresight reasoning significantly reduced hindsight bias [30]. Another effective strategy, particularly valuable in policy analysis contexts, is the consider-the-opposite approach where decision-makers explicitly explain plausible alternative outcomes beyond what actually occurred [31]. These remedies could be particularly valuable for AI safety and alignment efforts where accurate assessment of past predictions and outcomes is crucial for improving future forecasting. For example, AI developers and safety or governance team managers could implement processes to keep track of foresight records of the reasons behind

decisions by default, including any predictions or forecasts, and to use those records after the plans have been implemented to accurately reflect past thinking during learning, evaluation, and improvement processes (e.g., postmortems).

3.4.1 Black swans

Black Swans can be described as opportunities that are perceived to yield consistent gains but involve disproportionately large, low-probability risks that tend to be more obvious in hindsight after the occurrence of harm. The harms of Black Swans outweigh the benefits in terms of expected value over time. Investors often make decisions based on recent data without accounting for potential risks not visible in a dataset [32].

There are no uncertainty-free data to rely on when making predictions about unprecedented technological innovations, and this inherent variability can be seen as a risk. Advanced AI might seem like a great opportunity to improve the world (e.g., cure diseases, solve climate change), but the bet may not be worth it due to the possibility of extreme risks (e.g., loss of control, extinction). A series of AI developments could yield a streak of great outcomes, appearing as a stable trend, until a disastrous event occurs. A visible streak of consistent success may lead people to overlook or underestimate the potential risks from AI advancements. Investors, companies and governments could over-invest in these misleading developments with too little caution before negative and unforeseen (but preventable) consequences occur.

Research suggests promising potential for information consumer remedies to improve awareness of Black Swan events, although significant research gaps remain. A laboratory experiment demonstrated that people's learning about Black Swans followed a Bayesian probability model when confounding information and incentives were absent [32]. This finding suggests that under appropriate learning conditions, both the general public and key decision-makers could develop better awareness of extreme risks from advanced AI. Creating clear reports or contexts free of confounding variables and incentives might be best for learning about Black Swans for AI forecasting and policymaking tasks that require foresight (e.g., generating a comprehensive list of valuable if-then commitments for what to do about powerful capabilities if or when they emerge and analysing the relative strength of proposed solutions). However, real-world implementation faces challenges due to the complex interaction of multiple factors and difficulty identifying which variables will affect target outcomes.

Future systematic research could help validate and improve Black Swan awareness remedies. Key areas for investigation include testing the comparative effectiveness of different training modules and tools, such as warning participants about Black Swans through descriptions and example solutions before investment or planning tasks. Researchers could also examine how different experimental conditions affect remedy effectiveness, particularly comparing scenarios where Black Swans are present in available datasets versus cases where they are not. For instance, studies could investigate participant responses to remedies (e.g., warnings, definitions, historical examples, or interactive probability modeling tasks) when told of expert predictions about low-probability risks (e.g., 2% chance of project failure) in cases where historical data show no failures due to small sample size or normal variation.

3.5 Conjunction fallacy

The conjunction fallacy refers to the mistake people make when they overestimate conjunctive probabilities [33, 34]. A conjunctive probability is the likelihood of outcomes occurring jointly (e.g., winning two separate bets). In an experiment, participants were given a description of a person that was intended to be representative of feminists and unrepresentative of bank tellers. Participants were tasked with ranking statements in order of likelihood. 85% of the participants ranked the statement "Linda is a bank teller and is active in the feminist movement." as more probable than the statement "Linda is a bank teller" [33]. This is incorrect reasoning because the former is a subset of the latter, so logically the latter must be more likely. A replication of this experiment found a similar effect, this time with a 58.1% error rate [35].

In the context of risks from advanced AI, an example of the conjunction fallacy would be to attribute a higher likelihood to the possibility of company Z developing AGI than to the possibility of AGI being developed by anyone. The tendency to make this fallacy could lead to an irrational perception of the distribution of AI safety concerns and suboptimal prioritization of resources in this space. In public governance contexts, if decision-makers implicitly rank the likelihood of "AGI will be created by company Z in the next century" higher than the likelihood of "AGI will be created in the next century", for example, they may be more likely to implement policies and safety measures that target company Z rather than implementing general safeguards that apply to a group of existing and future companies (including company Z).

To test whether this bias would apply to AI perception in this way, empirical studies could compare whether people believe specific AI developments and safety concerns are more likely to occur in conjunction with another statement or on their own. Even if people recognize that many outcomes have to happen in conjunction (in other words, multiple things have to go right in a series of events) for safe

AGI to be developed, they might tend to overestimate the probability that this will happen.

Conversely, people also underestimate disjunctive probabilities-the probability of an outcome occurring at least once in a set of events-relative to conjunctive probabilities and individual outcomes [33, 36]. In an experiment, participants were asked to make bets on disjunctive and conjunctive events. The disjunctive probability judgment involved evaluating the likelihood of drawing a red marble *at least once* from a bag containing 10% red marbles and 90% white marbles given seven independent attempts. The conjunctive event was the likelihood of drawing a red marble *seven times* from a bag containing the reverse proportion of coloured marbles with the same number of attempts. Participants consistently bet on the conjunctive outcome (which is 48% likely) rather than the disjunctive one (which is 52% likely) even though the disjunctive outcome has a higher chance of success [36].

As Yudkowsky says, "The scenario of humanity going extinct in the next century is a disjunctive event. It could happen as a result of any of the existential risks we already know about-or some other cause which none of us foresaw." [8]. This reasoning also applies in the context of the risk of losing control of AIs as they become more advanced. Assuming many possible but seemingly unlikely events could lead to humans losing control over advanced AI, people are likely to underestimate the overall likelihood of this outcome.

One of the ways humanity could potentially lose control is if AGI is developed, as many actors are currently trying to do. Even if we assume each one has a very low chance of achieving this goal, if there are many of them, overall the probability of reaching this outcome is higher, and people are likely to underestimate this probability. People might fail to recognize reaching AGI as a set of disjunctive events in the first place if they only think of the possibility of one company reaching this goal.

Information system remedies combining incentives and collaboration can effectively reduce conjunction errors in probability judgments. A replication of Tversky and Kahneman's experiment found that mild incentives reduced the conjunction error rate from 58 to 33% [35]. The same study demonstrated that allowing participants to consult with one or two other participants further reduced error rates to 34% and 17.2%, respectively. These findings suggest that real-world work environments and AI decision-making contexts could benefit from implementing collaboration practices and providing incentives for accurate probability judgments to improve capability forecasts and threat models.

Future research could extend our understanding of conjunction and disjunction biases specifically in AI decision-making contexts. Studies could examine how people assess conjunctive and disjunctive probabilities in AI-related scenarios, such as comparing judgments about "AGI will be created by Meta this century" versus "AGI will be created this century" or "Safe AGI will be created" versus "AGI will be created." Additionally, research could evaluate the effectiveness of cognitive training and statistical tools in reducing these biases across different forecasting contexts.

3.6 Anchoring, adjustment and contamination effects

Anchoring bias refers to an overreliance on initial information to make judgments under uncertainty [6]. Adjustment is the tendency to use a starting point as an anchor from which to adjust estimates of facts or probabilities [37]. People, including experts in a domain, exhibit a strong tendency to under-adjust from an anchor such as an initial guess or the first number stated in a negotiation.

For example, if a respected AI researcher says they think all jobs will be automated in X-Y (e.g., 5–7) years and other researchers hear this, they might be more likely to report estimates that are closer to the first numbers stated. The same could happen for estimates of risk levels associated with various AI developments. Risk and task complexity have been found to increase the anchoring effect in forecasts [38]. Under-adjusted expert estimates and biased forecasts could affect policymakers' judgments of whether and how early to implement precautionary measures such as universal basic income to mitigate the social risks of large-scale job replacement, for example. AI developers could also over- or under-prioritize various safety measures based on anchored judgments.

There are several effective remedies for anchoring bias, but simply being aware of the potential influence of anchors is not one of them [39]. Adame [40] found that training modules involving generating several reasons why there is no relationship between the anchor and the target effectively mitigated the anchoring bias in an experimental setting. This is known as the consider-the-opposite strategy (information consumer and systems remedy). Monetary incentives have been found to reduce anchoring bias in forecasts (information systems remedy) [38]. Echterhoff et al. [41] used an AI algorithm to reduce anchoring bias in tasks involving sequential decisions (information systems remedy).

Contamination effects refer to a set of general phenomena in which problem-irrelevant information can affect cognitive judgment. In other words, irrelevant information can amplify biases. For instance, tasks that take up cognitive resources (attention) can interfere with adjustment from an anchor, leading to less adjustment or corrections of judgments [42].

To illustrate this concept, imagine someone being asked to judge whether AGI will be developed before 2050. Let's say their answer is 'no.' When asked for a more precise guess of the year in which they would expect this development to happen (if at all), they might answer 2081, whereas if they had not been given the year 2050 as an anchor, they might have said 2200. If the task involved higher cognitive load or interfering demands, such as mentally rehearsing some words while making these judgments, the person might adjust even less from 2050 and, say, 2065.

Another type of judgment-irrelevant information that can contaminate judgments is fiction [9]. For example, when the media cites fictitious examples of existential risk mitigation strategies such as Terminator and other productions that depict catastrophes caused by advanced AI, this information may anchor people's judgments about real-life situations. Yudkowsky calls this the "logical fallacy of generalization from fictional evidence" [8].

Biased, harmful decisions are more likely to occur when AI researchers and policymakers are under high cognitive load while making high-stakes judgment calls.

Both individual and systematic remedies can help prevent contamination effects that may impair judgment in high-stakes AI-related decisions. Information consumer remedies focus on reducing task-irrelevant information to limit instances of high cognitive load that impair judgment and maintain focus on decision-relevant factors. For instance, workers can isolate relevant report sections in a separate document before analysing relationships between variables and making decision recommendations. Similarly, managers and directors can request reports or graphs that include only specified, key variables from subordinates. These approaches could be particularly valuable for AI safety and alignment efforts where clear thinking about complex technical and policy decisions is essential. This could mean that, for example, researchers and analysts would benefit from separating thinking processes about advancements in different types of capabilities (e.g., code autocompletion vs AI reasoning) if the two were related to separate decisions.

Information system remedies address contamination effects by implementing workflows that separate judgment-intensive decisions from other cognitively demanding tasks. This approach involves creating dedicated decision-making environments-for instance, using a separate desktop with no other open tabs, activating do not disturb mode, closing office doors, and maintaining focused attention throughout the decision process without multitasking or task switching. Such systematic changes are especially important when AI researchers and policymakers face high cognitive loads while making high-stakes judgment calls. AI safety organizations could identify key decisions (e.g., next steps in strategic plans or theories of change) and implement structured decision-making environments with focused, supportive research resources (e.g., a research report summarizing the key vulnerabilities of a cutting-edge system or on current limitations of AI liability laws in regions with top AI labs) to prevent contamination from irrelevant sources of information. AI safety researchers and teams might benefit from implementing processes (e.g., agendas with prepared, key information about options to discuss or norms around limiting tangents during decision-oriented discussions) to limit how much information is considered and make sure the most important details are properly integrated at each step.

Before testing whether contamination effects affect decision-making in relevant contexts, it might be worth finding out which AI decision-making contexts are more likely to be prone to contamination effects. This could be done by surveying policymakers, developers, and safety researchers or anyone who makes decisions relevant to AI safety and alignment. Questions could include what types of human error most often lead to failures in their experience and whether these contexts are associated with more distracting information than contexts where there is very little human error. It may also be worth asking if risk and task complexity could be reduced, and testing whether less cognitively busy contexts or interfaces lead to fewer errors in judgment.

A lab experiment could investigate whether reading media articles that cite fictitious versus real examples of AI issues and solutions contaminates judgments in a simulated AI decision-making task.

It might also be worth testing whether and to what degree AI forecasts (including expert forecasts) are influenced by anchors, such as the year 2050 in the example provided earlier in this section.

3.7 Confirmation bias

Confirmation bias refers to people's tendency to seek evidence that aligns with their prior beliefs instead of seeking disconfirming evidence, leading to the perpetuation of false beliefs. In a seminal paper by Wason [43], participants were tasked with finding which relational rule the experimenter had in mind based on an initial series of three numbers, 2-4-6. Participants had the opportunity to test unlimited sets of three numbers and find out whether they fit the experimenter's rule before guessing the rule. Only 21% of participants correctly stated the rule that the numbers had to be in ascending order. Most incorrect guesses were sufficient but not necessary, for example, "increasing intervals of two" [43, p.132]. This experiment shows that people often fail to seek evidence to disconfirm their hypotheses before reaching conclusions, which leads to incomplete or incorrect perspectives on issues.

If people perceive an inaccurate pattern in AI developments, they may fail to look for evidence that disconfirms the existence of this pattern, and only look for or pay attention to confirmatory evidence. For example, if people see or hear more news about developments in capabilities, they may implicitly conclude that less publicized developments in other capabilities or safeguards are less significant without seeking evidence to disprove this perspective.

Another example is that a few cases of AI safety failures may be misinterpreted as a general pattern that AI is inherently risky or uncontrollable, even if these failures are very rare or have been corrected. This could lead to implementing unnecessary precautions that hinder the development and adoption of safe and beneficial technologies. See Sect. 3.1 for a discussion of similar errors in judgment based on incomplete information.

People may observe trends in AI developments, such as a linear or exponential trend in developments for a given period, failing to update their beliefs in the presence of signs of bigger patterns such as short periods of rapid growth followed by plateaus and setbacks. These inaccurate and self-reinforcing perceptions could lead to consistently wrong predictions and long-term misallocation of resources.

Implementing computer-mediated counterarguments in decision support systems is an information system remedy for confirmation bias [44]. This approach could be particularly valuable for AI researchers with strong preconceptions or policymakers with selective reading behaviors, who may see significant improvements in decision quality following such systemic interventions. As confirmation bias is the result of failing to seek disconfirming evidence, making a habit of seeking evidence to update our beliefs and getting computer systems to help us do this is valuable. Many people within the fields of AI advancements, safety, and governance have strong, diverging beliefs about AI, so applying evidence-based computer-mediated counterarguments within discussion platforms and in decision-making processes related to top priority or high-impact contexts (e.g., catastrophic risk assessments and evaluations of mitigation approaches) could be particularly beneficial.

3.8 Scope neglect (scope insensitivity)

Scope neglect, or scope insensitivity, is the tendency to have similar responses to situations that differ by orders of magnitude. For example, participants in an experiment were willing to allocate roughly as many resources ($80 vs $88) to save many birds (2000) or orders of magnitude more (200,000) [45]. When people fail to think through situations involving large numbers rationally, they fail to perceive and act on large differences. As numbers increase, sensitivity to differences tends to decrease, even in the context of saving human lives [46]. When people are scope insensitive, they are likely to miss the biggest opportunities for positive impact. As Yudkowsky puts it, "The human brain cannot release enough neurotransmitters to feel emotion a thousand times as strong as the grief of one funeral. A prospective risk going from 10,000,000 deaths to 100,000,000 deaths does not multiply by ten the strength of our determination to stop it. It adds one more zero on paper for our eyes to glaze over, an effect so small that one must usually jump several orders of magnitude to detect the difference experimentally." [8, p. 16].

Emerging AI technologies can scale rapidly. In the first two months following the launch of ChatGPT, it reached 100 million regular users, which may be the fastest customer growth rate ever recorded for a web platform [47]. As a consequence, any harms associated with new technologies, such as the spread of misinformation in the case of ChatGPT, will also reach large numbers. Advanced AI has the potential to affect very large numbers of lives in much more significant ways. The general public may not be equipped to understand and integrate the magnitude of these upcoming changes and potential risks.

Multiple remedies at both individual and systemic levels can help address scope insensitivity in AI safety and governance decisions. Information consumer remedies focus on obtaining accurate estimates of impact and using quantitative tools for analysis. AI safety and governance researchers should aim to get precise estimates of the number of individuals affected by various decisions and employ quantitative tools to analyze and compare options. Systematic assessment of expected impact is especially important when technical and governance teams evaluate policies or make design choices that involve large differences in impact. For example, cost-effectiveness analyses can compare required resources (e.g., money or time) per expected unit of improvement (e.g., life saved) across different AI developments and safety interventions, enabling more deliberate and accurate comparisons.

Information system remedies emphasize organizational processes that ensure systematic impact assessment. AI organizations and governance teams can improve strategic resource allocation by implementing processes that mandate systematic assessment of expected impact when evaluating policies and making decisions involving large differences in scale.

Scope insensitivity is more prevalent in decisions that are temporally and psychologically close to the decision-maker [48], suggesting that evaluation of impact differences may be more accurate for events further in the past and future. Addressing potential future issues well before they become urgent might be especially useful, given that decision-makers might be more rational when crises do not seem imminent.

To provide a concrete example, this could mean securing if-then commitments from top AI labs for how to respond to possible sets of conditions such as pre-defined, advanced AI capabilities associated with potentially catastrophic threat models. Several systematic approaches might help mitigate the psychological closeness factor that also amplifies scope insensitivity: obtaining evaluations from individuals not personally involved in target projects before product or program launches, conducting post-mortem analyses after some time has passed, and excluding people with conflicts of interest from decision-making processes.

Future research could investigate whether judgments are more scope sensitive when participants are asked to imagine a current decision as if the decision were far in the past or future or as if they were a distant other making the decision. It may also be worth testing whether providing incentives improves the quality of reasoning in situations involving large differences in numbers of impacted individuals.

3.9 Planning fallacy

The planning fallacy is the tendency to consistently underestimate how long it will take to complete tasks and projects [49, 50]. People make predictably optimistic evaluations of risks and benefits-bold forecasts-which often lead to failures [51]. See also Sect. 3.11 and the note on optimism and pessimism for related discussions.

People working on AI developments may be prone to the planning fallacy. This could be relevant in the context of important and time-sensitive outcomes that depend on adherence to relatively strict timelines, such as launching a product before a competitor.

When teams of AI safety researchers or policymakers establish plans for projects, it may be crucial to implement risk prevention measures within a certain time frame. They may fail to prioritize the most crucial next steps and outcomes in favour of a more ambitious set of actions if they are prone to think that they will be able to complete them all in time. Failing to take the most important actions in time could lead to unmitigated risks and the incidence of avoidable harm.

Multiple information consumer remedies, such as reference class forecasting [52], can help mitigate the planning fallacy in AI safety and governance projects. Reference class forecasting is a three-step process that involves identifying a reference class of previous projects that resemble the target project, making a probability distribution with at least five data points from the reference class, and making an estimate for the duration (or cost) of the target project based on this distribution of previous examples. For the last step, the median value from the reference class examples can be used as a 50% confidence estimate of when the project will be completed [53].

In cases where instant, intuitive estimates of timelines are required, doubling estimates can be an effective remedy because in many cases, actual project duration exceeds estimated timelines within a range of up to two times the original estimate. The Clearer Thinking web platform [53] offers a free course on the planning fallacy to learn how and when to use reference class forecasting and estimate doubling.

Another remedy is to break down tasks into a list of ordered sub-tasks, which is especially useful for complex, multifaceted tasks [50].

Finally, a recent study suggests that using paper calendars or a mobile calendar with a broader perspective (e.g., a whole month visible on the page) while planning leads to more effective plan development and completion [54] (information consumer remedy).

Further research on which factors affect the planning fallacy, especially in the AI contexts described above, would be highly valuable. For example, field research could be used to determine whether AI safety and governance teams accomplish their work plans according to the initial timelines set and which factors affect the need for extensions or additional resources. Experimental conditions could also be used to test partial remedies such as having contingency plans to eliminate nonessential tasks at predetermined time points if there is insufficient progress to maximize the chances of addressing key priorities in time.

3.10 Restraint bias

People tend to overestimate their own self-control over impulses and affective states, such as hunger and fatigue. In an experiment on smokers, inflated beliefs about one's impulse control led to more exposure to temptation, which led to more smoking [55]. Even when people know that a product is addictive or have struggled to stay in control when using it, they underestimate the product's influence on their future behaviours. Another example is that people often spend more time on screens than they want to [56].

Restraint bias is relevant to safety and alignment because AI users, developers, and policymakers might overestimate how much control we have in our interactions with AI systems, which means some forms of loss of control might not be perceived immediately or taken seriously enough for teams to implement the appropriate safeguards in time or for users to fully understand in which contexts the use of AI might not align with their goals.

When machine learning algorithms have the goal of optimizing variables, such as user viewing time or ad revenue on social media platforms, they can learn to exploit the weaknesses of our primitive brains without explicit knowledge

of them. People can be in AI-assisted decision-making contexts in which they think they have more control than they do. For example, recommender algorithms can achieve their goal (e.g., maximizing user viewing time or money spent on purchases) in ways that aren't aligned with the user's original intentions or goals.

With the use of more advanced AI, such as cutting-edge large language models, people are at risk of overestimating their ability to overcome their aversion to doing work or making decisions themselves. People are averse to cognitive effort, and an experiment shows that they sometimes prefer to trade cognitive effort for physical pain [57]. If a user interacts with an AI that recommends investments, they might just pick the default choice without doing further research to avoid cognitive effort. Due to restraint bias, they might think that the AI had less influence over their decision than it actually had.

Overestimating one's own ability to resist an impulse leads to exposing oneself to more temptations, which leads to less self-control and more associated harms, as illustrated with the smoking example above. If a worker overestimates their ability to overcome the temptation to use AI-generated content without double-checking all the facts, they might be more likely to keep the AI nearby and use it more. This cognitive-behavioural pattern leads to more use of AI-generated content and less fact-checking, thus more misinformation in the work the worker turns in.

People might think that they will do effortful and important tasks such as fact-checking, carefully reviewing AI recommendations when making decisions, thinking critically about which AI-generated content to use, checking a privacy policy, and taking the necessary precautions to avoid inputting sensitive information in systems trained on user data. However, if they can save time and energy by not doing these things, they are likely to underestimate their chances of overcoming their aversion to cognitive effort and thus overestimate how much control they have in AI-assisted decision-making contexts.

Improving the accuracy of people's representations of their self-control could help them understand what level of exposure to temptation is appropriate or unlikely to interfere with their goals. Unfortunately, we were not able to find remedies for restraint bias. To mitigate the harms of this bias and reduce susceptibility to unwanted influence from misaligned AI systems in daily interactions, it may be helpful to generally reduce exposure to temptations or situations that require restraint.

Future research could test whether being aware of or reminded of the existence of this bias reduces it in subsequent trials. It would also be helpful to test a consider-the-opposite remedy where participants are asked to list reasons why they might not exhibit as much restraint as they expect in contexts such as following AI use guidelines at work/school or respecting self-imposed screen time limits, for example. Another potentially promising avenue is to provide people with data about their self-restraint to test whether that allows them to progressively update their beliefs until they have a more accurate representation of their restraint and the factors that influence it.

Another potentially useful approach to reduce restraint bias is to strengthen self-control so that it matches perceived self-control. Restraint can be strengthened with implementation intentions, also known as if-then rules [58]. For example, in AI contexts, individuals might establish rules like "If I generate content with AI, I will check all the facts in the output before integrating it in my work or factoring the information into my decisions."

Understanding the impact of willpower fatigue, the idea that willpower is like a 'muscle' than can get weaker as we use it throughout the day [59, 60], so attempts to engineer environments to minimize exposure to temptations may be helpful [61]. However, note that this work on "ego depletion" has been attacked in the replication crisis and is controversial [62, 63]. Example ways to engineer environments to minimze willpower fatigue in contexts that involve the use of convenient new technologies include turning off a phone and leaving it in another room during a work day instead of simply thinking "I won't look at my phone during the workday". In contexts specific to AI, this could mean blocking access to AI websites when individuals don't want to risk using AI tools inadequately or in ways that violate policies around the use of AI (e.g., failing to fact-check).

Self-control limitations and restraint bias affect behaviours in ways that are worthy of consideration in the design of AI user interfaces and organizational policies (information system remedy). In AI safety contexts, the stakes of restraint bias are particularly high. Consider an organization developing advanced AI systems, where convenience-driven shortcuts around security protocols could have catastrophic consequences. Rather than relying solely on individuals' willpower to consistently choose secure practices over convenient ones, organizations can implement systematic safeguards.

3.11 Calibration and overconfidence

People's confidence in their knowledge of facts, judgments, and estimates is poorly calibrated. Subjective probability judgments are systematically overconfident [64, 65]. For example, when asked for very high confidence intervals (e.g., asking people to give a range for how tall they think the Statue of Liberty is such that they are 98% confident that the actual height is in that range), people provide ranges that are consistently too small [64]. This means that the actual

value falls within the estimated range much less often than the stated confidence predicts (e.g., 68% of the time instead of 98%).

In the context of AI, this means that people's prediction models are often too narrow, and they will be more likely to fail to appropriately consider outcomes further from the middle of the distribution when planning for various outcomes. For example, the manager or strategic analyst for an AI safety research team could have a 98% confidence interval that AI developments will surpass human-level intelligence somewhere within the next five to 100 years. If this is an overconfident estimate (as most are), the likelihood that the actual occurrence of this event will fall outside the stated range (e.g., in 1 year or in 104 years) is more than the predicted 2%.

Cognitive training is an information consumer remedy for calibration and overconfidence that can also be implemented in systems. When participants received feedback on previous calibration failures, were told about overconfidence in previous studies, and received explanations of calibration, their responses on subsequent trials were less biased, but confidence levels were still far from well-calibrated [64]. This type of training could be part of courses or workshops for people working in AI. Having well-adjusted views is a key part of good judgment, which is needed for high-stakes decisions around new technologies. Moreover, warning AI decision-makers about this bias could potentially reduce the reliance on poorly calibrated individual judgments (e.g., overconfident expert estimates of when a given capability will exist) relative to more reliable or stronger forms of evidence (e.g., trends in advancements of capabilities based on robust third-party evaluations).

Future research could investigate why some people's judgments are much better calibrated than others (e.g., superforecasters) and what factors influence calibration and prediction accuracy. It may also be worth testing different ways of stating probability judgments to determine whether some lead to systematically more accurate judgments than others.

3.12 Bystander effect (bystander apathy)

When multiple people witness a potential problem, such as an unconscious person or smoke entering a closed room, each individual is less likely to take action than if just one person was witnessing the problem [66]. This is the bystander effect or bystander apathy.

People in the general population may notice urgent issues or big potential risks and observe other people's behaviour to see whether they seem alerted or disposed to take action. They may even doubt the existence of emergencies such as existential risks based on how other people are reacting to them.

When looking at others and seeing that no one is taking action, each individual tends to maintain the status quo. People might have pluralistic ignorance, meaning they might not know, based on their lack of visible reaction, that others also perceive the risk. There might also be a diffusion of responsibility where it is unclear who is responsible for mitigating these risks [8]. The average person might think it's someone else's problem, when in reality everyone would be affected by global catastrophic risks from emerging technologies and benefit from the implementation of more preventive safety measures. The bystander effect could explain why many people may see a potential for global catastrophic risks arising from current and upcoming AI developments, yet choose not to do anything.

Information system remedies for the bystander effect could counter the diffusion of responsibility that occurs in groups by giving individuals a clear understanding that the situation depends on them [8]. For example, singling out a bystander and asking them to help is an effective strategy. Global risk situations differ from controlled experiments in that it is much harder to determine who should be in charge of taking action for different issues. Those who don't take action might benefit from those who do, but if no one takes action or too few actions are taken, no one gets to live in a safe world. AI safety systems should give individuals clear responsibilities (e.g., clearly defined liability laws outlining who pays which costs in different catastrophic risk contexts) so that all individuals involved are personally incentivized to contribute to outcomes that benefit everyone. Future research could confirm whether clearly identifying who is responsible for harms in AI development contexts reduces risk-taking tendencies or harmful incidents.

To test whether a form of bystander effect is present in the general population's response to AI risk contexts, a survey study could include questions to assess whether individuals think cutting-edge AI developments pose large risks and whether they think more action is needed to mitigate these risks, whether they think others are taking action, and whether they are taking action themselves. The main hypothesis would be that among those who perceive a high, unmitigated risk of catastrophic outcomes this century (>5% probability), those who think others are not taking action would be least likely to report taking action themselves.

In a follow-up study, participants could be shown statistics about the average person's concern for AI risks and propensity to action (from the first study), then asked if they intend to take any actions from a checklist of options accessible to an average person (e.g., send an email to a political representative about AI safety concerns). For more checklist items, see the future research suggestions in the priming bias

section. A control group of participants could be shown statistics about AI risks and then respond to the same checklist question. If more people in the experimental group check boxes (or if the median respondent checks more boxes), this could be taken as evidence that seeing other people's concern (pluralistic knowledge) alongside their inaction can reduce the bystander effect. This would be an information system remedy.

3.13 Bandwagon effect (herd mentality)

Herd instinct is a psychological phenomenon where individuals in a group adopt attitudes or exhibit behaviours primarily because they perceive those attitudes or behaviours to be popular. A recent meta-analysis suggests that bandwagon cues such as likes on social media posts had a small, positive effect on credibility evaluations [67]. There may be a bandwagon effect when people's votes are inherently influenced by the popularity of a political party or candidate [68]. Stronger peer information (e.g., 88% of people choose this option vs lower percentages) leads to a greater likelihood of herding behaviours in the contexts of retirement planning and buying disability insurance [69]. Providing participants with a warning message about herd influences within a decision support system did not reduce bias in this financial context.

When prominent public figures lean towards certain views on AI developments and alignment, others may follow suit without individual scrutiny or consideration of alternative views and approaches. Individuals influenced by the herd instinct could have an irrational amount of trust or distrust in the trajectory of AI developments. Those who take an optimistic view might ignore potential risks and ethical considerations, and those who take a pessimistic view might undervalue the benefits and progress in the field. If the predominant view in a peer group is that advanced AI poses an existential threat, individuals might adopt this view uncritically, and in the opposite case, individuals might underplay the risks or have a less serious attitude towards risk mitigation efforts. Individuals and organizations influenced by the herd instinct may also follow trends or invest in popular areas, rather than those substantiated by analyses of strategic importance. This could lead to an inefficient allocation of resources, giving too much attention to certain areas while neglecting others that might be crucial for AI alignment.

To strategically shape the trajectory of the future, it's relevant to consider that some people are more likely to take on new beliefs than others. One example is that late teens and early adults are more susceptible to attitude changes [70]. If the early adopters of ideas are reached first, those ideas become more socially acceptable. To prevent the propagation of false or misleading views and unhelpful or harmful behaviours, ideally, the early adopters of AI ideas would present a critical, evidence-based, and nuanced view of AI developments that will steer people towards accurate views and adequate change management responses. In AI governance and safety contexts, people might be more prone to support popular avenues and efforts, scrutinizing popular choices less critically than other options.

Remedies to the bandwagon effect involve strategically limiting exposure to information about others' beliefs or judgments in contexts where maintaining the integrity of individual perceptions and contributions is key. Generally, to reduce instances where people are unduly influenced by others' beliefs and actions, it is best to remove indicators of popularity such as removing the option of seeing "likes" on social media platforms that have this customizable setting (information consumer and system remedy). It might also be helpful to reason through options individually and keep a record of this thought process before finding out what most other people seem to think. In a team, people could read each other's initial thoughts to assess a wider range of potentially less popular but possibly more helpful ideas. To provide a concrete example in context, when designing a public platform for iterative improvement of proposed safety and alignment solutions and research plans, removing the ability to see public popularity ratings or other reviewers' comments before offering individual constructive criticism would diminish the chances of bandwagon-biased judgments. These processes would need to be tested to assess their effectiveness as information consumer and system remedies.

Future research could also investigate whether a consider-the-opposite paradigm is an effective remedy for the bandwagon effect. Participants could be tasked with choosing a side in an AI safety and alignment debate. They could also be asked to select a preferred action from a list of possible solutions to an AI safety concern. In either scenario, the options would be accompanied by fabricated information about which option is most often chosen by others. If there is a bandwagon effect, a follow-up experiment could ask participants to list reasons why the options identified as less popular might be better or why more people should choose them, and then they could undergo a similar decision task again (a consider-the-opposite strategy). If the bandwagon effect is reduced or eliminated in the follow-up experiment, this would suggest that the consider-the-opposite strategy is an effective information consumer remedy.

3.14 Priming bias

Priming refers to a set of phenomena where being exposed to information leads to associations that can prompt later

ideas and goals without conscious awareness through the automatic activation of these associated concepts [71, 72]. Although the evidence for priming is generally good (e.g., for semantically-related words), social priming refers to the unconscious influence of social information on behaviours and is a controversial idea [73, 74]. There have been some unreplicable social priming findings due to statistical noise and small sample sizes [74].

Priming effects can be used strategically to improve outcomes. For example, priming effects are helpful in some cases to influence people to choose more active behaviours such as taking the stairs instead of the elevator to get to class [75]. A field experiment found that people followed rules more honestly when an image of eyes gave the impression of being watched [76]. This finding could also be classified as an observer expectancy effect, where people behave differently when being watched. A field study found that organizational effectiveness and efficiency rose significantly when a CEO sent emails containing words related to achievement (such as "accomplish") to all employees [77]. When people's expectations influence outcomes, this is also known as expectation bias, and it has been observed in various contexts including psychiatric treatments [78]. However, more research is needed to fully understand this phenomenon and its long-term implications, including potential backfire effects (e.g., employee burnout or false hope about treatment success).

When people are exposed to (accidentally or intentionally) biased reports, the priming of related information can influence their opinions about issues and subsequent decisions. For example, if a workplace constantly highlights the benefits of innovation in AI discussions, this positive association could prime positive attitudes toward new AI developments and projects before workers receive sufficient context or attempt to make accurate judgments. Conversely, if a workplace constantly highlights doomsday predictions in AI discussions, they may be primed to have negative opinions and oppose new developments based on equally limited information. Priming could lead to people underestimating or overestimating the risks from advanced AI. The type of information people are exposed to could influence their willingness to take action on AI safety issues. See the availability section for a similar discussion. The main difference with priming is that general attitudes are influenced by unconscious activation of related ideas and behaviour patterns whereas biases that arise from the availability heuristic are typically described as affecting explicit judgments based on reliably limited memory retrieval patterns.

Priming can only be partly attenuated, but being aware of it and how companies can use it to influence behaviour can help people understand and mitigate its influence. One prevention strategy is to minimize exposure to information that could prime people in ways that do not align with their values (e.g., removing unwanted ads). Increasing exposure to priming information that aligns with personal or organizational goals may mitigate the harms of this bias.

Future research could investigate whether and to what degree people can be primed to respond positively or negatively to AI developments. For example, various fake or real facts could be presented as news headlines, such as "Experts estimate the risk of extinction from advanced AIs to be 1 in 10 in the coming century." Variations could have much lower odds like 1 in 10,000 instead, or feature a highly positive outcome such as "cure most known diseases" or "solve climate change" (see Sect. 3.2.1 for a related discussion of biases affecting probability judgments). With a pre-test and post-test design, participants would rate their sentiment on AI developments on a seven-point scale from most negative to most positive. Other measures could include evaluations of trust in the trajectory of AI developments for the next century, willingness to invest in current AI developments, willingness to donate to an AI safety research institute, or the likelihood of advocating for or against specific AI developments on social media or at work. Specific behavioural self-assessment questions could include how likely they are to share a specific post with the above headlines or how likely they would be to vote for the implementation of a specific AI safety policy. See the section on availability 3.1, including the availability subsection, for other biases that might influence these attitudes and judgments as well as a related survey research suggestion.

3.15 Framing

People's decision-making is biased in favor of positively-framed situations. For instance, participants presented with a hypothetical disease scenario chose a treatment 72% of the time when it was framed positively (e.g., to save 200 out of 600 lives) compared with only 22% when the same treatment was framed negatively (e.g., 400 deaths out of 600 people) [79].

AI developments in general could be framed positively or negatively. For example, some people could say that AI could lead to curing many diseases and improving the quality of life of billions of people, or they could say that AI developments could lead to catastrophic outcomes including the extinction of humanity. These framings could accurately represent the same situation, assuming that these positive and negative outcomes are both plausible. If a probability is assigned to each outcome, for example, 33% and 66%, people may be more likely to support AI developments if they are framed as "A 33% probability of saving millions of lives, and a 66% probability of no lives being saved" or

"saving 3 out of 9 billion people" compared to equivalent, negative framings of the situation.

Specific AI safety interventions can also be framed positively or negatively. For example, a positive framing for a hypothetical AI safety intervention targeting a given population would be "A one in three probability of saving everyone and a two in three probability that no one will be saved'. The equivalent negative framing would be "A one in three probability that nobody will die and a two in three probability that everyone will die," to closely mirror the probabilistic framing and wording found in Tversky and Kahneman's [79] experiment. Similar framing effects could apply to any AI safety intervention aimed at reducing risks or known harms.

While we were not able to find remedies for the framing effect, strategic message framing can help mitigate potential harms in different contexts. Research in medical settings suggests that message effectiveness varies by context [80]. For illness detection interventions, loss-framed messages (e.g., "helps avoid X harm") tend to be more effective than gain-framed ones (e.g., "helps maintain positive outcomes"). Conversely, for illness prevention, gain-framed messages (e.g., "increases quality of life") typically outperform loss-framed alternatives (e.g., "decreases risk of X harm"). AI safety communications aimed at promoting preventive alignment measures to reduce risks of catastrophic outcomes might be more effective when gain-framed. AI safety organizations could test this assumption with A-B testing to determine whether the number of interactions with gain-framed communications in higher than for loss-framed communications (with measures such as likes, shares, donations, signing up for events, applications to jobs).

To test whether this bias applies to perception and decision-making around AI developments and safety interventions in the ways outlined above, the suggested message framings could be tested in survey experiments.

3.16 Optimism and pessimism

Optimism and pessimism are tendencies to perceive and expect irrationally positive and negative outcomes, respectively [81–84]. A quick overview of the literature suggests that these biases can exist simultaneously, and there is a lot of individual variation and many factors that seem to influence both of these tendencies, including some mental health disorders and cognitive bias modification interventions [85, 86]; thus, it may not be possible to draw conclusions or make recommendations that apply more generally. Highly context-dependent considerations are outside the scope of this paper. A comprehensive review of which factors give rise to optimism or pessimism is still missing, but could be helpful to identify context-specific remedies.

3.17 Normalcy bias

People's priors are adjusted to predict that events will unfold as they normally do, even when signs point to other possibilities. The tendency to fail to adequately account for the likelihood of potential threats is sometimes called the normalcy bias, and it leads to irrationally dangerous actions or inactions in the face of catastrophic risks [87]. It typically takes time to update beliefs and take necessary actions even when there are strong signals of significant imminent threats, such as disaster warnings and evacuation orders [88].

People may underestimate how much AI will change their lives, predicting no major disruptions based on their past experiences. Even people who know about the risks from AI may fail to take them seriously. For example, they might fail to appropriately prepare for or identify a deepfake spearfishing scam of a family member asking for help, or expecting bad things to only happen to other people. They may have less consideration or allocate fewer resources than would be rational to the possibility of larger-scale disaster as well.

Normalcy may lead to a failure to update beliefs quickly as new technologies emerge. If a new AI is particularly unsafe or unethical, it may take some time for users to notice and for the general public to update their beliefs, even in the presence of warning signs. There could be unnecessary and avoidable delays in updating risk-benefit analyses and making necessary changes in relevant policies, AI systems, or human behaviours.

Information system remedies can help counter normalcy bias through structured disaster response protocols. Research by Omer and Alon [87] suggests a multi-step approach to maintain social system continuity and to mitigate harm during disasters. The preparation stage involves acknowledging disaster probabilities and developing clear emergency response plans. This is followed by a warning stage that requires "timely, repeated, and unambiguous warnings and instructions" [87, p. 278]. For a comprehensive understanding of all four proposed disaster response steps, readers can consult [87] and [89]. For a discussion of how people tend to look to others before responding to warning signs, see the bystander effect section. The relationship between normalcy bias and the bystander effect is discussed in another section, particularly regarding how people tend to look to others before responding to warning signs. AI safety organizations working on civilizational resilience could develop emergency response protocols in collaboration with existing efforts between leading AI organizations and governments (e.g., if-then commitments outlined in [90]).

3.18 Ostrich effect

People tend to avoid negative information including feedback that could be used to contribute to one's goals [91]. Similarly to the normalcy bias (see Sect. 3.17), this can lead to neglecting important warning signs.

The general public might ignore bad news and negative signals about AI developments, including failures and potential global catastrophic risks associated with technological progress in this field. AI developers might ignore signs that their project is not as beneficial or less safe than expected, leading to worse outcomes. Policymakers or AI governance teams may fail to look for or spot gaps in their proposed solutions and policies.

Both individual and systematic remedies can help counter the ostrich effect in organizational contexts. Information consumer remedies involve individuals deliberately seeking more feedback and information about factors that could negatively impact their goals, even when such information might be uncomfortable to confront. Information system remedies focus on implementing workplace processes for periodic goal progress monitoring and regular feedback provision to workers. These approaches could be particularly valuable in AI safety contexts where early awareness of potential issues is crucial. Given that individuals cannot be trusted to pull the plug on their own projects that appear obviously unsafe to external actors, AI safety organizations can help develop best practices and regulations so that AI labs implement early warning and monitoring systems with evaluations of capabilities and protocols for tracking progress on safety levels.

3.19 Source misattribution (source confusion) bias

Source misattribution or source confusion bias is the tendency to misremember the source of information or knowledge [92]. For example, someone might think they heard a piece of information on the news rather than overhearing a stranger say it. Repeatedly imagining an event increases the chance that people confuse it for an event that really happened [93].

Source misattribution bias may be getting worse with AI because large language models often do not cite sources [94], so people might have a false memory of finding information from a reliable source (e.g., an academic paper from a trustworthy journal) instead of reading it off an AI-generated text or fake news article. Some information pools contain a lot of misinformation, and since people forget where their memories came from, verifying facts becomes more important and trickier. If information sources are left unchecked, people may unknowingly spread misinformation about AI developments and associated safety/risk information.

Multiple remedies at both individual and systematic levels can help maintain epistemic health and prevent source confusion. Information consumer remedies focus on developing better information verification practices. People need training in source checking and determining source reliability. While AI tools like Elicit can help find and summarize research papers, users must verify facts in original sources due to potential AI misrepresentation. Additional individual practices include critically assessing research methodologies, evaluating evidence quality and quantity, reducing consumption of AI-generated content, and normalizing the consultation and citation of primary sources.

In AI safety organizations, these norms can be implemented systematically. For example, onboarding and training processes can teach epistemic source attribution norms, and these can also be implemented in decision-making processes such as standardized formats for tracking and verifying information presented in reports, presentations, and discussions. Moreover, when forecasting AI risk scenarios, organizations can implement procedures to track evidence of previous developments and failures, enabling verification of whether situations occurred in remembered contexts or were imagined. These systematic approaches can be complemented by broadcast and personal communication remedies, such as developing fact-checking algorithms and policies.

3.20 Illusion of control

The illusion of control refers to people's tendency to think they have more control over situations than they do [95, 96]. This bias has even been observed in situations where outcomes are determined by chance [97]. This bias is associated with impaired financial trading performance [98]. For a related discussion, see Sect. 3.10.

State-of-the-art AI systems often have a high degree of uncertainty and unpredictability, yet people are prone to think they can control them more than they actually can. Evidence suggests that this bias is worse in stressful and competitive environments [98], so AI developers who race to develop and release products before competitors may have stronger illusions of control. They might think the AI systems they work with are more predictable or responsive to instructions than they are and fail to undergo sufficient evaluations with causal assessments in an appropriate range of relevant contexts. The illusion of control could lead to overestimating the predictability of an AI system and underestimating the risks or potential harms.

Information system remedies can help address the illusion of control by developing better understanding of causality and scientific methods. An effective approach involves first demonstrating to people how their thinking can be fallible,

then teaching them scientific methods for assessing causality [99]. This intervention may be particularly valuable for policymakers without scientific backgrounds, though it is important to note that even scientists remain susceptible to the illusion of control, in spite of their training. A key complementary practice, supported by multiple researchers [8, 100], is to assess the methods used to reach conclusions rather than relying solely on the credentials of those making claims. AI organizations can administer training programs for technical and policy teams to test their assumptions about control over AI systems and incentivize the implementation of causality checks using scientific research methods.

To test whether the illusion of control affects people's perceptions of control over AI systems as suggested above, an experiment could compare subjective ratings of control of people interacting with AIs with quantifiable levels of agency, or randomness if that is easier to implement for experimental purposes. For example, people could play an online game that involves collaborating with an AI to complete a task. There could be an AI that will follow the participant's instructions 100% of the time, 90%, 80%, etc., and after a given number of trials or a given timeframe of interaction, people could rate how much control they think they have over each AI on a scale of 0–10.

One limitation of this design is that it would not account for the possibility that the AI is strategically misleading the participant, making it seem as though humans control it to gain their trust when in reality the AI has full control and is waiting for its moment to deviate from the script and achieve its goals. Participants who perceive this as a possibility may in theory rate their perception of control as lower. Despite this possibility, if the subjective control ratings are significantly higher than actual control levels, this would be evidence for the illusion of control in human-AI interactions.

4 Conclusion

Human biases can lead to serious errors in judgment that could undermine AI safety and alignment efforts. A helpful first step to reduce this risk is to incorporate existing remedies into individual and organizational decision-making structures. More research on biases, remedies, and how they apply in the context of emerging technologies would also be beneficial.

The rest of this section contains a summary list of information consumer and system remedies as well as a discussion of who can implement them and how. It ends with a general summary of the future research directions suggested throughout this paper.

4.1 Information consumer remedies

The list below contains a summary of information consumer remedies, which can be scaled into information system remedies. One way to embed consumer remedies into decision-making systems is to enforce these remedies through policies, guidelines, managerial decisions, and instructional documents. Another way to embed remedies in wider decision-making structures is to change how information is presented or how the user is prompted within digital decision-support systems.

When policymakers and AI organizations need to make decisions based on predictions or forecasts of how AI is expected to develop, a few key biases will likely appear. To remedy hindsight bias, anticipate a range of possible, unprecedented futures, and keep a foresight record of the reasoning behind predictions to review in hindsight. When looking back on decisions, list and explain possibilities other than the outcome that occurred.

When making decisions based on predictions, generate a long list of possible courses of action and solutions to anticipated issues before analyzing and selecting the most promising next steps to avoid missing better opportunities due to reliance on the availability heuristic.

When comparing options, use quantitative tools such as cost-effectiveness analyses to avoid scope neglect, thus preventing the failure to assess different interventions proportionally to their impact. Remove decision-irrelevant information from decision-making processes to avoid contamination effects that can worsen biases in decision-making. For an example of how this remedy can scale beyond the individual level, managers and directors could request to remove decision-irrelevant information from recommendation reports to focus their attention on key, relevant factors only.

Maintain good epistemic health and reduce the illusory truth effect by avoiding exposure to repeated false or uncertain information about AI. Warn people when the information they are about to receive could be misleading (e.g., when it is AI-generated or contains unchecked facts about AI developments or risks). Remind people to reflect on whether information is true before sharing it [17] to decrease the risk of relying on false information when making decisions.

To avoid situations where herding bias could lead to failing to identify important concerns and possible solutions, reduce exposure to information about what others are doing when possible and make judgments separately before finding out which options are most popular. For example, policymakers or safety researchers may be less influenced by herding if they think through a problem and potential solutions before asking others what they think, looking online,

or doing research to find out how other organizations have handled similar issues.

When planning projects, the planning fallacy can be reduced by using reference class forecasting (see Sect. 3.9 for relevant details). Simply doubling estimates of project completion times can be appropriate if short on time. Another debiasing strategy is to break tasks down into lists of smaller, concrete action items before setting timelines. Using a calendar to visualize deadlines and progress over time (e.g., a month) can also help.

4.2 Information system remedies

The following list contains information system remedies that involve making changes in the structure of information systems relevant to decision making. Information consumers could apply them at the individual level as well with the right resources. Organizational leaders such as directors, executives, and managers may be able to implement these at an organizational or team level. They could impose the use of these remedies in relevant contexts, such as any decisions based on how recent developments and upcoming technological progress might affect a team or an organization's activities and goals. People working on improving institutional decision-making or designing decision support systems could also work on incorporating these remedies into their projects or tools.

To reduce confirmation bias, the failure to seek disconfirmatory evidence or weigh it equally, implement computer-mediated counterarguments in decision support systems. See Huang et al. [44] for details. To prevent primacy and recency order effects, systematically change the order of presented items, especially in the context of important decisions (e.g., names on a ballot or suggested solutions to a problem). Implement policies or processes to ensure large differences in expected impact across intervention options are accounted for to reduce scope insensitivity. For example, AI governance teams or organizations involved in AI development could make it mandatory to conduct risk-benefit estimates before approving new AI developments to systematically detect potentially large differences in impact and compare them against an objective threshold. Because scope insensitivity can be increased by psychological and temporal proximity [48], remove conflicts of interest and consider potential future issues well before they become urgent. To reduce errors in judgment of probabilities (e.g., how likely is X outcome), use base rates and incorporate them into graphs because, as mentioned earlier, visual representations of probabilities can reduce base rate neglect [24]. As mentioned above, probability judgments are also more accurate when people can draw samples from a probability distribution [25]. To reduce the tendency to overestimate the probability of events occurring in conjunction (conjunction fallacy) and therefore improve the accuracy of forecasts, provide incentives and give workers the opportunity to consult with coworkers on tasks that involve the assessment of conjunctive probabilities [35]. To reduce anchoring bias in forecasts, provide monetary incentives [38] and ask workers to generate reasons why there is no relationship between the anchor (e.g., the year 2050 as an example of when AGI could be created) and the target (the year when AGI will be created) [40]. See Echterhoff et al. [41] for an example of an AI algorithm that reduced anchoring bias in sequential decision tasks.

In cases where restraint bias is a concern, for example enforcing an AI tool use policy in a workplace, reduce exposure or access to these tools. For example, if workers agree to AI use standards and think they will use AI tools responsibly, but track records suggest workers have been leaking sensitive information or failing to fact-check AI-generated content before use, it may be helpful to avoid recommending the use of tools that require restraint or block access to AI websites or apps.

To reduce the bystander effect, identify who is responsible for handling AI-related damages. Clarify which individuals are responsible for responding to specific situations or contexts. Examples of specified responsibility include making organizations and developers responsible for damages caused by the AI models they create and requesting emergency response plans for anticipated risks or potential damages. Similarly, to prevent catastrophes that can result from ignoring important warning sighs (normalcy bias), implement prevention measures, including acknowledging the chances of disaster and making emergency response plans. Quickly provide clear and repeated danger warnings and calls to action in disaster contexts.

To reduce the ostrich effect, periodically monitor progress on key objectives and ensure that people receive this feedback. For example, individuals or teams could update a progress tracker every other day and be shown a visual representation of how close they are to reaching AI safety and alignment milestones.

Information system remedies include debiasing training. Providing a causal explanation of false positives to reduce base rate neglect [25]. Provide training on black swans for any decisions based on forecasting. To reduce clustering illusions, provide robust statistical tools to detect correlations. To reduce representativeness biases (including base rate neglect, insensitivity to sample size, misconception of chance, insensitivity to predictability, the illusion of validity, and the misconception of regression), provide definitions and examples of each bias. To reduce overconfidence and improve calibration, it can be helpful to receive an explanation of calibration and the results of studies as well

as feedback on calibration failures. To remedy the illusion of control, show people how they can reach wrong conclusions and show them how to determine causality scientifically. See Barberia et al. [99] for details.

4.3 Future research suggestions

A clearer foundation is needed to integrate existing research on biases. For instance, a comprehensive review of existing biases and a clear outline of what factors are thought to give rise to them would provide a very useful starting point to guide future efforts.

More experimentation with potential remedies in different contexts would be helpful. For example, consider-the-opposite remedies generally involve listing reasons why a judgment or a way of reaching conclusions might be wrong or why alternative judgments might be right. Remedies that involve reminding or prompting people to use their existing reasoning skills more broadly or in key contexts could be tested on a wider range of biases and contexts than just the ones identified in existing research. Another idea that might work as a more general remedy is providing people with customized data on their own biased judgments and whether remedies help them achieve better results. For an example, see the future research suggestion in the restraint bias section.

It would also be helpful to conduct field research in AI decision-making contexts to provide more ecologically valid and directly useful results. Field researchers could test remedies with quick feedback loops to incorporate context-specific, sequential improvements. The advantage of taking this process beyond the lab is to confirm the utility of solutions in fast-changing environments. Relevant AI decision-making contexts include AI government agencies, AI development organizations, AI safety and alignment organizations, and any institutions that need to assess whether and how to respond to changes caused by AI progress. Many institutions might have to consider directions such as whether to implement new AI tools for productivity, policies to mitigate risks, or necessary changes in business strategy due to changing opportunities.

It could also be helpful to incorporate remedies into existing decision-support systems and test them to determine how well they work in different contexts and how well they scale.

Finally, the biases covered in this paper may have similar applications in other cause areas potentially involving catastrophic risks, such as biosecurity, nuclear risk mitigation, and climate change. Researchers in these disciplines could collaborate with cognitive scientists to better understand how biases apply in unprecedented, high-risk, high-impact situations, with a focus on finding effective remedies that can be easily applied in the most relevant and important contexts.

Acknowledgements None. No funding is associated with this work.

Author contributions Roy-Stang did the literature review and wrote the paper with substantial editing and content contribution from Davies.

Funding No funds, grants, or other support was received. The authors have no relevant financial or non-financial interests to disclose.

Declarations

Competing interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Haselton, M.G., Nettle, D., Andrews, P.W.: The evolution of cognitive bias. The handbook of evolutionary psychology, 724–746 (2015). https://doi.org/10.1002/9781119125563
2. Flyvbjerg, B.: Top ten behavioral biases in project management: An overview. Proj. Manag. J. **52**(6), 531–546 (2021). https://doi.org/10.1177/87569728211049046
3. Müller, V.C., Bostrom, N.: Future progress in artificial intelligence: A poll among experts. AI Matters **1**(1), 9–11 (2014). https://doi.org/10.1145/2639475.2639478
4. Oeberst, A., Imhoff, R.: Toward parsimony in bias research: A proposed common framework of belief-consistent information processing for a set of biases. Perspect. Psychol. Sci. **18**(6), 1464–1487 (2023). https://doi.org/10.1177/17456916221148147
5. Hilbert, M.: Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. Psychol. Bull. **138**(2), 211 (2012). https://doi.org/10.1037/a0025940
6. Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. Science. **185**(4157), 1124–1131 (1974)
7. Gratton, C., Gagnon-St-Pierre, E.: Heuristics and Cognitive Biases. Shortcuts: A Handy Guide to Cognitive Biases, Vol. 2, trans. E. Muszynski. Accessed: 2024-05-01 (2020). https://en.shortcogs.com/theorie
8. Yudkowsky, E.: Cognitive biases potentially affecting judgment of global risks. Global Catastrophic Risks **1**(86), 13 (2008). https://doi.org/10.1093/oso/9780198570509.003.0009
9. Lewandowsky, S., Ecker, U.K., Seifert, C.M., Schwarz, N., Cook, J.: Misinformation and its correction: Continued influence and successful debiasing. Psychol. Sci. Public Interest **13**(3), 106–131 (2012). https://doi.org/10.1177/1529100612451018
10. Acciarini, C., Brunetta, F., Boccardelli, P.: Cognitive biases and decision-making strategies in times of change: a systematic literature review. Manag. Decis. **59**(3), 638–652 (2021). https://doi.org/10.1108/md-07-2019-1006
11. Lab, T.D.: List of Cognitive Biases and Heuristics (n.d.)
12. Fazio, L.K., Brashier, N.M., Payne, B.K., Marsh, E.J.: Knowledge does not protect against illusory truth. J. Exp. Psychol. Gen. **144**(5), 993 (2015). https://doi.org/10.1037/xge0000098
13. Mattavelli, S., Béna, J., Corneille, O., Unkelbach, C.: People underestimate the influence of repetition on truth judgments (and more so for themselves than for others). Cognition **242**, 105651 (2024). https://doi.org/10.1016/j.cognition.2023.105651

14. Siegrist, M., Sütterlin, B.: Human and nature-caused hazards: The affect heuristic causes biased decisions. Risk Anal. **34**(8), 1482–1494 (2014). https://doi.org/10.1111/risa.12179
15. Heath, C., Heath, D.: Decisive: how to make better choices in life and work. Random House of Canada, Canada (2013)
16. Anderau, G.: Fake news and epistemic flooding. Synthese **202**(4), 106 (2023). https://doi.org/10.1007/s11229-023-04336-7
17. Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A.A., Eckles, D., Rand, D.G.: Shifting attention to accuracy can reduce misinformation online. Nature **592**(7855), 590–595 (2021). https://doi.org/10.1038/s41586-021-03344-2
18. Rey, A., Le Goff, K., Abadie, M., Courrieu, P.: The primacy order effect in complex decision making. Psychol. Res. **84**(6), 1739–1748 (2020). https://doi.org/10.1007/s00426-019-01178-2
19. AlKhars, M., Evangelopoulos, N., Pavur, R., Kulkarni, S.: Cognitive biases resulting from the representativeness heuristic in operations management: an experimental investigation. Psychol. Res. Behav. Manage. (2019). https://doi.org/10.2147/prbm.s193092
20. Tversky, A.: On the psychology of prediction. Psychol. Rev. **80**(4), 237–251 (1973). https://doi.org/10.1037/h0034747
21. Bar-Hillel, M.: The base rate fallacy controversy. In: Advances in Psychology vol. 16, pp. 39–61. Elsevier, Amsterdam (1983). https://doi.org/10.1016/s0166-4115(08)62193-7
22. Kahneman, D., Tversky, A.: On the reality of cognitive illusions. (1996). https://doi.org/10.1037/0033-295x.103.3.582
23. Welsh, M.B., Navarro, D.J.: Seeing is believing: Priors, trust, and base rate neglect. Organ. Behav. Hum. Decis. Process. **119**(1), 1–14 (2012). https://doi.org/10.1016/j.obhdp.2012.04.001
24. Roy, M.C., Lerch, F.J.: Overcoming ineffective mental representations in base-rate problems. Inf. Syst. Res. **7**(2), 233–247 (1996). https://doi.org/10.1287/isre.7.2.233
25. Hayes, B., Newell, B., Hawkiins, G.: Causal model and sampling approaches to reducing base rate neglect. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 35 (2013)
26. Kahneman, D., Tversky, A.: Subjective probability: A judgment of representativeness. Cogn. Psychol. **3**(3), 430–454 (1972). https://doi.org/10.1016/0010-0285(72)90016-3
27. Fischhoff, B., Beyth, R.: I knew it would happen: Remembered probabilities of once-future things. Organ. Behav. Hum. Perform. **13**(1), 1–16 (1975). https://doi.org/10.1016/0030-5073(75)90002-1
28. Chen, J., Kwan, L.C., Ma, L.Y., Choi, H.Y., Lo, Y.C., Au, S.Y., Tsang, C.H., Cheng, B.L., Feldman, G.: Retrospective and prospective hindsight bias: Replications and extensions of fischhoff (1975) and slovic and fischhoff (1977). J. Exp. Soc. Psychol. **96**, 104154 (2021). https://doi.org/10.1016/j.jesp.2021.104154
29. McKee, D.: Uncontrollable: The threat of artificial superintelligence and the race to save the world. Independently Published, Ottawa (2023)
30. Davies, M.F.: Reduction of hindsight bias by restoration of foresight perspective: Effectiveness of foresight-encoding and hindsight-retrieval strategies. Organ. Behav. Hum. Decis. Process. **40**(1), 50–68 (1987). https://doi.org/10.1016/0749-5978(87)90005-7
31. Roese, N.J., Vohs, K.D.: Hindsight bias. Perspect. Psychol. Sci. **7**(5), 411–426 (2012). https://doi.org/10.1177/1745691612454303
32. Payzan-LeNestour, E.: Can people learn about 'black swans'? experimental evidence. Rev. Financial Stud. **31**(12), 4815–4862 (2018). https://doi.org/10.1093/rfs/hhy040
33. Tversky, A., Kahneman, D.: Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychol. Rev. **90**(4), 293 (1983). https://doi.org/10.1037/0033-295x.90.4.293
34. Moro, R.: On the nature of the conjunction fallacy. Synthese **171**, 1–24 (2009). https://doi.org/10.1007/s11229-008-9377-8
35. Charness, G., Karni, E., Levin, D.: On the conjunction fallacy in probability judgment: New experimental evidence regarding linda. Games Econom. Behav. **68**(2), 551–556 (2010). https://doi.org/10.1016/j.geb.2009.09.003
36. Bar-Hillel, M.: On the subjective probability of compound events. Organ. Behav. Hum. Perform. **9**(3), 396–406 (1973). https://doi.org/10.1016/0030-5073(73)90061-5
37. Teovanović, P.: Individual differences in anchoring effect: Evidence for the role of insufficient adjustment. Eur. J. Psychol. **15**(1), 8 (2019). https://doi.org/10.5964/ejop.v15i1.1691
38. Meub, L., Proeger, T.: Can anchoring explain biased forecasts? Experimental evidence. J. Behav. Exp. Financ. **12**, 1–13 (2016). https://doi.org/10.1016/j.jbef.2016.08.001
39. Wilson, T.D., Houston, C.E., Etling, K.M., Brekke, N.: A new look at anchoring effects: basic anchoring and its antecedents. J. Exp. Psychol. Gen. **125**(4), 387 (1996). https://doi.org/10.1037/0096-3445.125.4.387
40. Adame, B.J.: Training in the mitigation of anchoring bias: A test of the consider-the-opposite strategy. Learn. Motiv. **53**, 36–48 (2016). https://doi.org/10.1016/j.lmot.2015.11.002
41. Echterhoff, J.M., Yarmand, M., McAuley, J.: Ai-moderated decision-making: Capturing and balancing anchoring bias in sequential decision tasks. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pp. 1–9 (2022). https://doi.org/10.1145/3491102.3517443
42. Gilbert, D.T., Pelham, B.W., Krull, D.S.: On cognitive busyness: When person perceivers meet persons perceived. J. Pers. Soc. Psychol. **54**(5), 733 (1988). https://doi.org/10.1037/0022-3514.54.5.733
43. Wason, P.C.: On the failure to eliminate hypotheses in a conceptual task. Quarterly J. Experimental Psychol. **12**(3), 129–140 (1960). https://doi.org/10.1080/17470216008416717
44. Huang, H.-H., Hsu, J.S.-C., Ku, C.-Y.: Understanding the role of computer-mediated counter-argument in countering confirmation bias. Decis. Support Syst. **53**(3), 438–447 (2012). https://doi.org/10.1016/j.dss.2012.03.009
45. Kahneman, D., Ritov, I., Schkade, D., Sherman, S.J., Varian, H.R.: Economic preferences or attitude expressions?: An analysis of dollar responses to public issues. Elicitat. Preferences (2000). https://doi.org/10.1007/978-94-017-1406-8_8
46. Fetherstonhaugh, D., Slovic, P., Johnson, S., Friedrich, J.: Insensitivity to the value of human life: A study of psychophysical numbing. J. Risk Uncertain. **14**, 283–300 (1997). https://doi.org/10.5040/9798216002796.ch-006
47. Chow, A.: Why ChatGPT Is the Fastest Growing Web Platform Ever| Time (2023). https://time.com/6253615/chatgpt-fastest-growing/
48. Chang, H.H., Pham, M.T.: Affective boundaries of scope insensitivity. J. Consumer Res. **45**(2), 403–428 (2018). https://doi.org/10.1093/jcr/ucy007
49. Buehler, R., Griffin, D., Ross, M.: Exploring the "planning fallacy": Why people underestimate their task completion times. J. Pers. Soc. Psychol. **67**(3), 366 (1994). https://doi.org/10.1037/0022-3514.67.3.366
50. Kruger, J., Evans, M.: If you don't want to be late, enumerate: Unpacking reduces the planning fallacy. J. Exp. Soc. Psychol. **40**(5), 586–598 (2004). https://doi.org/10.1016/j.jesp.2003.11.001
51. Kahneman, D., Lovallo, D.: Timid choices and bold forecasts: A cognitive perspective on risk taking. Manage. Sci. **39**(1), 17–31 (1993). https://doi.org/10.1287/mnsc.39.1.17
52. Flyvbjerg, B.: Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. Eur.

Plan. Stud. **16**(1), 3–21 (2008). https://doi.org/10.1080/09654310701747936
53. The Planning Fallacy. Clearer Thinking. Retrieved April 20, 2024, from (n.d.). https://www.clearerthinking.org/tools/the-planning-fallacy
54. Huang, Y., Yang, Z., Morwitz, V.G.: How using a paper versus mobile calendar influences everyday planning and plan fulfillment. J. Consum. Psychol. **33**(1), 115–122 (2023). https://doi.org/10.1002/jcpy.1297
55. Nordgren, L.F., Harreveld, Fv., Pligt, Jvd: The restraint bias: How the illusion of self-restraint promotes impulsive behavior. Psychol. Sci. **20**(12), 1523–1528 (2009). https://doi.org/10.1111/j.1467-9280.2009.02468.x
56. Ko, M., Yang, S., Lee, J., Heizmann, C., Jeong, J., Lee, U., Shin, D., Yatani, K., Song, J., Chung, K.-M.: Nugu: a group-based intervention app for improving self-regulation of limiting smartphone use. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 1235–1245 (2015). https://doi.org/10.1145/2675133.2675244
57. Vogel, T.A., Savelson, Z.M., Otto, A.R., Roy, M.: Forced choices reveal a trade-off between cognitive effort and physical pain. elife **9**, 59410 (2020). https://doi.org/10.7554/elife.59410
58. Gollwitzer, P.M.: Weakness of the will: Is a quick fix possible? Motiv. Emot. **38**, 305–322 (2014). https://doi.org/10.1007/s11031-014-9416-3
59. Gailliot, M.T., Baumeister, R.F.: The physiology of willpower: Linking blood glucose to self-control. Pers. Soc. Psychol. Rev. **11**(4), 303–327 (2007). https://doi.org/10.1177/1088868307303030
60. Baumeister, R.F., Tice, D.M., Vohs, K.D.: The strength model of self-regulation: Conclusions from the second decade of willpower research. Perspect. Psychol. Sci. **13**(2), 141–145 (2018). https://doi.org/10.1177/1745691617716946
61. McGonigal, K.: The Willpower Instinct: How Self-control Works, Why It Matters, and What You Can do to Get More of It. Penguin, New York (2013)
62. Friese, M., Loschelder, D.D., Gieseler, K., Frankenbach, J., Inzlicht, M.: Is ego depletion real? an analysis of arguments. Pers. Soc. Psychol. Rev. **23**(2), 107–131 (2019). https://doi.org/10.1177/1088868318762183
63. Baumeister, R.: Self-control, ego depletion, and social psychology's replication crisis (2019) https://doi.org/10.31234/osf.io/uf3cn
64. Alpert, M., Raiffa, H.: 21. a progress report on the training of probability assessors (1982). https://doi.org/10.1017/cbo9780511809477.022
65. Soll, J.B., Palley, A.B., Klayman, J., Moore, D.A.: Overconfidence in probability distributions: People know they don't know, but they don't know what to do about it. Manage. Sci. (2023). https://doi.org/10.1287/mnsc.2019.00660
66. Latané, B., Darley, J.M.: Bystander 'apathy'. Am. Sci. **57**(2), 244–268 (1969)
67. Wang, S., Chu, T.H., Huang, G.: Do bandwagon cues affect credibility perceptions? a meta-analysis of the experimental evidence. Commun. Res. **50**(6), 720–744 (2023). https://doi.org/10.1177/00936502221124395
68. Barnfield, M.: Think twice before jumping on the bandwagon: Clarifying concepts in research on the bandwagon effect. Political studies review **18**(4), 553–574 (2020). https://doi.org/10.1177/1478929919870691
69. Compen, B., Pitthan, F., Schelfhout, W., De Witte, K.: How to elicit and cease herding behaviour? on the effectiveness of a warning message as a debiasing decision support system. Decis. Support Syst. **152**, 113652 (2022). https://doi.org/10.1016/j.dss.2021.113652
70. Krosnick, J.A., Alwin, D.F.: Aging and susceptibility to attitude change. J. Pers. Soc. Psychol. **57**(3), 416 (1989). https://doi.org/10.1037/0022-3514.57.3.416
71. Chartrand, T.L., Bargh, J.A.: Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. J. Pers. Soc. Psychol. **71**(3), 464–478 (1996). https://doi.org/10.1037/0022-3514.71.3.464
72. Aarts, H., Custers, R., Veltkamp, M.: Goal priming and the affective-motivational route to nonconscious goal pursuit. Soc. Cogn. **26**(5), 555–577 (2008). https://doi.org/10.1521/soco.2008.26.5.555
73. Molden, D.C.: Understanding priming effects in social psychology: What is "social priming" and how does it occur? Soc. Cogn. **32**(Supplement), 1–11 (2014). https://doi.org/10.1521/soco.2014.32.supp.1
74. Chivers, T.: What's next for psychology's embattled field of social priming. Nature **576**(7786), 200–203 (2019). https://doi.org/10.1038/d41586-019-03755-2
75. Wryobeck, J., Chen, Y.: Using priming techniques to facilitate health behaviours. Clin. Psychol. **7**(2), 105–108 (2003). https://doi.org/10.1080/13284200410001707553
76. Bateson, M., Nettle, D., Roberts, G.: Cues of being watched enhance cooperation in a real-world setting. Biol. Let. **2**(3), 412–414 (2006). https://doi.org/10.1098/rsbl.2006.0509
77. Stajkovic, A.D., Latham, G.P., Sergent, K., Peterson, S.J.: Prime and performance: Can a ceo motivate employees without their awareness? J. Bus. Psychol. **34**, 791–802 (2019). https://doi.org/10.1007/s10869-018-9598-x
78. Williams, J.B., Popp, D., Kobak, K., Detke, M.: P-640-the power of expectation bias. Eur. Psychiatry **27**(S1), 1–1 (2012). https://doi.org/10.1016/s0924-9338(12)74807-1
79. Tversky, A., Kahneman, D.: The framing of decisions and the psychology of choice. Science **211**(4481), 453–458 (1981). https://doi.org/10.1126/science.7455683
80. Soll, J.B., Milkman, K.L., Payne, J.W.: A user's guide to debiasing. The Wiley Blackwell handbook of judgment and decision making **2**, 924–951 (2015). https://doi.org/10.1002/9781118468333.ch33
81. Weinstein, N.D.: Unrealistic optimism about future life events. J. Pers. Soc. Psychol. **39**(5), 806 (1980)
82. Mansour, S.B., Jouini, E., Napp, C.: Is there a "pessimistic" bias in individual beliefs? evidence from a simple survey. Theor. Decis. **61**(4), 345–362 (2006). https://doi.org/10.2139/ssrn.915382
83. Roser, M., Ritchie, H.: Optimism and pessimism. Our World in Data, https://ourworldindata.org/optimism-and-pessimism accessed May 19, 2024 (2018)
84. Aue, T., Dricu, M., Moser, D.A., Mayer, B., Bührer, S.: Comparing personal and social optimism biases: magnitude, overlap, modifiability, and links with social identification and expertise. Human. Soc. Sci. Commun. **8**(1), 1–12 (2021). https://doi.org/10.1057/s41599-021-00913-8
85. Alloy, L.B., Ahrens, A.H.: Depression and pessimism for the future: biased use of statistically relevant information in predictions for self versus others. J. Pers. Soc. Psychol. **52**(2), 366 (1987). https://doi.org/10.1037/0022-3514.52.2.366
86. MacLeod, C., Mathews, A.: Cognitive bias modification approaches to anxiety. Annu. Rev. Clin. Psychol. **8**, 189–217 (2012). https://doi.org/10.1146/annurev-clinpsy-032511-143052
87. Omer, H., Alon, N.: The continuity principle: A unified approach to disaster and trauma. Am. J. Community Psychol. **22**, 273–287 (1994). https://doi.org/10.1007/bf02506866
88. Drabek, T.E.: Disaster warning and evacuation responses by private business employees. Disasters **25**(1), 76–94 (2001). https://doi.org/10.1111/1467-7717.00163

89. Valentine, P.V., Smith, T.E.: Finding something to do: The disaster continuity care model. Brief Treatment Crisis Interv. (2002). https://doi.org/10.1093/brief-treatment/2.2.183
90. Karnofsky, H.: If-then commitments for ai risk reduction. Carneige Endowment for International Peace (2024)
91. Webb, T.L., Chang, B.P., Benn, Y.: 'the ostrich problem': Motivated avoidance or rejection of information about goal progress. Soc. Pers. Psychol. Compass **7**(11), 794–807 (2013). https://doi.org/10.1111/spc3.12071
92. Schacter, D.L., Harbluk, J.L., McLachlan, D.R.: Retrieval without recollection: An experimental analysis of source amnesia. J. Verbal Learn. Verbal Behav. **23**(5), 593–611 (1984). https://doi.org/10.1016/S0022-5371(84)90373-6
93. Goff, L.M., Roediger, H.L.: Imagination inflation for action events: Repeated imaginings lead to illusory recollections. Memory Cognition **26**, 20–33 (1998). https://doi.org/10.3758/BF03211367
94. Lab, T.D.: Source Confusion (n.d., Retrieved April 20, 2024). https://thedecisionlab.com/biases/source-confusion
95. Matute, H., Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M.A., Barberia, I.: Illusions of causality: How they bias our everyday thinking and how they could be reduced. Front. Psychol. **6**, 146427 (2015). https://doi.org/10.3389/fpsyg.2015.00888
96. Matute, H., Vadillo, M.A., Vegas, S., Blanco, F.: Illusion of control in internet users and college students. CyberPsychol. Behav. **10**(2), 176–181 (2006). https://doi.org/10.1089/cpb.2006.9971
97. Langer, E.J.: The illusion of control. J. Pers. Soc. Psychol. **32**(2), 311 (1975). https://doi.org/10.1037/0022-3514.32.2.311
98. Fenton-O'Creevy, M., Nicholson, N., Soane, E., Willman, P.: Trading on illusions: Unrealistic perceptions of control and trading performance. J. Occup. Organ. Psychol. **76**(1), 53–68 (2003). https://doi.org/10.1348/096317903321208880
99. Barberia, I., Blanco, F., Cubillas, C.P., Matute, H.: Implementation and assessment of an intervention to debias adolescents against causal illusions. PLoS One **8**(8), 71303 (2013). https://doi.org/10.1371/journal.pone.0071303
100. Galef, J.: The scout mindset: Why some people see things clearly and others don't. Penguin, New York (2021)