THE GLOBAL PRIORITIES INSTITUTE

PSYCHOLOGY AND BEHAVIORAL SCIENCE RESEARCH AGENDA

Version 1

November 2024

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Table of Contents

Introduction

There are many problems in the world. Because resources are scarce, it is impossible for any given actor to solve them all. A government, philanthropist or individual seeking to improve the world therefore needs to prioritize, both among the problems themselves and among policies and interventions for addressing them. This prioritization requires careful analysis. Some opportunities are likely to be vastly more cost-effective than others. Identifying such opportunities—focus areas, policies, and interventions—requires grappling with a host of complex questions.

The aim of the Global Priorities Institute (GPI) is to conduct foundational research that informs the decision-making of individuals and institutions seeking to do the most good in the world. In particular, we focus on research that makes progress towards figuring out what the world's most pressing problems are and how these problems can be solved.

This document outlines some of the core research priorities for the psychology team at GPI. However, the research questions in this agenda extend beyond the immediate work being done at GPI to highlight further research areas that could provide valuable insights for global prioritization.

Potential applications of this research are discussed in the section Path to Impact at the end of this document.

Research areas

These research areas and questions are a first attempt at outlining the kind of psychological and behavioral science research that can inform global priorities research. There are many other topics that could be highly impactful to study in addition to the topics here. In fact, tractable research questions will probably test more specific and original hypotheses than we've been able to list. We hope that readers will use our high-level questions as inspiration to generate new questions to explore further.

In terms of organization, many of the research areas listed below overlap. This is an intentional and pragmatic decision. Some categories focus on an object-level topic (e.g. morality) and others focus on a particular approach (e.g. developmental psychology). We find it useful to consider research areas from these different angles, even though it means some questions will fall under multiple areas (e.g., by studying the developmental emergence of morality). For each sub-category, we first list an overarching general research question in bold, followed by a few concrete sub-questions. The sub-questions are intended to exemplify the broad questions that follow from the overarching general

question. We list some relevant existing literature after the questions for each research area. This list is not meant to be exhaustive, instead, they just include a few key examples. We appreciate that existing research has made substantial progress on many of the research questions we pose. Our research agenda is intended to inspire readers to build on this past research.

1 Morality

Understanding people's values, whether they can be influenced, and if so, how, is important to inform global priorities research. For example, people's moral values influence the extent to which they would support different efforts to improve the world and how they might conceive of that project. Morality research could take several forms. Some forms would focus on understanding people's values. For example, we could research how compatible the implications of global priorities research are with people's moral psychology. This knowledge could further inform the theoretical debates in global priorities research, or help determine what courses of action are more or less politically tractable. Or, to identify what approaches to improving the world are likely to be neglected by typical altruists, and therefore have low-hanging fruit unpicked, we could research which approaches are perceived least favorably on moral grounds. Other forms of this research could focus on fostering desirable values, such as more consideration for future lives, geographically and socially distant individuals, or openness to cost-effectiveness in moral judgments. For example, we could research what makes policies that focus on the distant future less morally compelling and identify factors that could change these preferences.

1.1 Understanding people's values

- **Whose wellbeing do people value and who do people grant moral status to?**
    - How wide is the range of entities that people deem worthy of moral concern and treatment? What predicts individual differences in this moral expansiveness? To what extent does moral expansiveness predict how much people morally value future generations, artificial intelligence that could achieve sentience or intelligent life forms that might follow humans? To what extent do people morally value non-sentient entities, e.g., non-sentient artificial intelligence?
- **How much do people value and utilize cost-effectiveness in moral judgments?**
    - How effectiveness-focused are people in the moral domain? How reluctant are people to prioritize the more effective option over a less effective option when this requires trading off lives against other values? How scope insensitive are people in helping contexts involving a large number of lives? How do these factors affect people's willingness to allocate resources to improving the world rather than having a smaller expected impact on individuals living today?

- **Which emotions motivate people to effectively improve the world?**
    - Are prosocial emotions such as empathy, sympathy or gratitude, motivators or obstacles to effectively improving the world? How can we harness prosocial emotions to help people do the most good? How important are social factors, such as social norms, in motivating people to improve the world?
- **How do people think about the future of humanity and existential risk?**
    - What are the psychological obstacles to caring about the long-term future of humanity (i.e., the next thousands, millions, or even billions of years), e.g. in terms of motivations (e.g., value-action gap), moral beliefs, empirical beliefs, etc?
    - How do people trade off bad outcomes, such as suffering, vs good outcomes, such as happiness? Under which framings do people's intuitions change? Does people's trade-ratio between suffering and happiness depend on whether they focus on mere outcomes or on actions required to bring about these outcomes?
    - How does temporal discounting bias affect moral consideration for future people? What are the causes of people's temporal discounting for future people—empirical uncertainty or normative considerations? How do time preferences in the moral domain relate to time preferences in the non-moral domain?
    - What are people's views of a future utopia? Do they want humanity to spread in space or stay earthbound? To what extent are people's views of an ideal future driven by non-utilitarian values?
- **What are people's population-ethical intuitions?**
    - Do people have clear and robust population-ethical views (i.e., about outcomes that differ in the composition or size of the respective populations)? Are some of their reactions to population-ethical dilemmas more robust and less framing-dependent than others? Are people's population-ethical views closer to totalism or averagism? Do they have person-affecting views?
- **What objections do people have about an approach to ethics based on impartial welfare maximization and improving the world?**
    - Do people believe it is too demanding or impractical?
    - Do people fundamentally disagree with the notion of impartial welfare maximization? How do people interpret the notion of impartial welfare maximization? To what extent do people endorse or deviate from consequentialism? To what extent do people endorse welfarist axiological values or the view that the overall value in the world is given by the sum total of wellbeing in it? Do people endorse prioritarian or egalitarian values? Do they endorse intrinsic non-welfarist values, such as seeing value aesthetics,

purity, or wisdom? How well-formed are people's views on these abstract philosophical questions? How well do they predict willingness to engage in the project of effectively improving the world?
    - Are people more motivated by object-level causes than by impartial welfare maximization in the abstract? Are they partial to helping specific groups such as people with a particular misfortune? Do they see the project of impartial welfare maximization as a threat to causes that they value intrinsically?
    - Are some objections based on misunderstandings or genuine disagreements with the principles or concrete implications?
- **How confident or uncertain are people in their moral values?**
    - Can people be made more or less uncertain about their moral values? How do people deal with moral uncertainty, and how does that relate to how they deal with other kinds of uncertainty?
    - How malleable are people's moral values? How often do people's moral philosophies meaningfully change?
- **How do moral values spread within and across social groups?**
    - What features of moral values make them most likely to be adopted by others? How does the popularity of a moral value influence its reception? How have cultural values changed over time?
    - Which particular moral values tend to "win" in the marketplace of ideas? Should we expect liberal values such as free speech, equality, and rule of law to become more popular with time, or is there a substantial risk that humanity at large will come to hold opposing values more strongly?
- **How open are people to moral pluralism?**
    - To what extent are people willing to promote the moral values of other people, even if they don't necessarily hold those values themselves? Do people become more or less pluralistic as they reflect on their own values more? When forced to make moral decisions, are people more or less pluralistic than they report they are?

See also Section 3 (Individual differences), Section 4 (Developmenta) and Section 5 (Cross-cultural) in this research agenda, as well as Section 1.1.1 (Welfare and beneficence) and 4.3.1 (Population ethics) in the GPI Philosophy Research Agenda and Section 1.3 (Welfare and decision procedures) in the GPI Economics Research Agenda.

1.2 Fostering values conducive to effectively improving the world

- **How can we foster values conducive to effectively improving the world?**
    - How can we spread values conducive to effectively improving the world?
    - How can we achieve fundamental norm change in society? How can we expand people's moral circle (e.g., reduce presentism) and increase

effectiveness-focus in society at large? How can we increase long-term thinking in a way that aligns with people's current norms?
    - What's the role of moral argument and moral reasoning in shaping moral values? What messages and framings work best to change attitudes and behavior?
    - How does a new issue go from being amoral to becoming morally relevant or from being considered a moral matter to becoming treated as morally neutral? Is this consistent across different issues? How does it vary at an individual level?
- **What are the best ways to reduce the value-action gap for those interested in effectively improving the world?**
    - How important is social proof? How effective are nudges?
    - How important is the role of identity or group belonging in motivating people to want to do the most good?
- **What are feasible strategies to overcome moral disagreement?**
    - How feasible is moral trade as an approach to overcome moral disagreement, i.e, the process where agents with different moral views agree to take actions or exchange resources in order to bring about outcomes which are better from the perspective of everyone involved? What norms would people tend to apply to these trades, e.g. a fair split of resources or a more outcome-focused approach?

Relevant work:

Berman, J. Z., Barasch, A., Levine, E. E., & Small, D. A. (2018). Impediments to effective altruism: The role of subjective preferences in charitable giving. *Psychological science, 29*(5), 834-844.

Bloom, P. (2017). *Against empathy: The case for rational compassion.* Random House.

Caviola, L., Althaus, D., Mogensen, A., & Goodwin, G. (2021). Population ethical intuitions. *Cognition*.

Crimston, C. R., Hornsey, M. J., Bain, P. G., & Bastian, B. (2018). Toward a psychology of moral expansiveness. Current Directions in Psychological Science, 27(1), 14-19.

Feinberg, M., Kovacheff, C., Teper, R., & Inbar, Y. (2019). Understanding the process of moralization: How eating meat becomes a moral issue. Journal of Personality and Social Psychology, 117(1), 50–72.

Gainsburg, I., Pauer, S., Nawal, A., Aloyo, E. T., Mourrat, J. C., & Cristia, A. (2021). How effective altruism can help psychologists maximize their impact.

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them.* Penguin.

Greenberg, S., (2001). Which intrinsic values set different demographic groups apart?

Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review, 125*(2), 131.

Lieder, F., Prentice, M., & Corwin-Renner, E. R. (accepted subject to minor revisions). An interdisciplinary synthesis of research on understanding and promoting well-doing. Social and Personality Psychology Compass.

Ord, T. (2015). Moral trade. Ethics, 126(1), 118-138.

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. Trends in cognitive sciences, 7(7), 320-324.

Reynante, B. M., Wilcox, J. E., Stephenson, O. L.,Lieder, F., Thielmann, I., & Lacopo, C. (submitted). Cultivating Changemakers: A review of Metachangemaking.

Rhee, J. J., Schein, C., & Bastian, B. (2019). The what, how, and why of moralization: A review of current definitions, methods, and evidence in moralization research. Social and Personality Psychology Compass. https://doi.org/10.1111/spc3.12511.

Rozen, P. (1999) The Process of Moralization. Psychological Science, 10(3): 218-221.

Singer, P. (2011). The expanding circle: Ethics, evolution, and moral progress. Princeton University Press.

Skitka, L. J., Hanson, B. E., Morgan, G. S., & Wisneski, D. C. (2021). The psychology of moral conviction. Annual Review of Psychology, 72, 347-366.

2 Rationality, judgment, and cognitive biases

If an impartial actor wants to use reason and evidence to effectively improve the world, they will be more effective to the extent that they avoid cognitive biases and make rational decisions. Psychological global priorities research research thus includes understanding people's judgments and decisions, how rational they are, and how to make them more so. Some of the worst outcomes for the future, e.g., nuclear war or misaligned AI, are likely driven by misjudgments and suboptimal decision making. Psychological research could therefore study the cognitive biases that lead people to make such dangerous high-stakes decisions. Relatedly, psychological research could identify biases that lead people to underestimate particularly effective approaches to proactively improving the world. Further, to help impartial altruists effectively improve the world, research could develop methods to boost rational decision-making and forecasting accuracy.

2.1 Identifying decision tendencies and biases relevant for global priorities research

Note. There are potentially hundreds of tendencies that are relevant. Here, we just list a few understudied ones as examples.

- **What cognitive biases affect how people decide between different options for effectively improving the world?**
    - Are people overconfident in their judgments of which choice would most effectively improve the world? How do people value new information and the ability to keep their options open? How do people weigh the scale of a problem relative to its tractability and neglectedness? Do such biases cause people to overestimate or underestimate the value of action directed at improving the world vs. solving problems in the present?
- **How do people deal with the perceived risk of decisions to improve the long-term future?**
    - How risk and ambiguity averse are people in altruistic decision-making?
    - How do people reason about very small probabilities and extreme tail events? Under what circumstances do people effectively "round down to zero" and ignore low-probability outcomes entirely? How do people weigh low-probability, extreme events? And to what extent are people fanatically focusing on one very small probability but high magnitude event? How do people reason about observer selection effects when estimating risks?
    - When making altruistic decisions, do people tend to think more about the overall value of the outcome/world, or about how much good they

themselves do (the difference they make)? Do people dislike being a largely anonymous part of a much greater altruistic effort? In particular, if people are risk-averse (or ambiguity averse), which of these quantities are they risk-averse with respect to?

- **How do people deal with radical uncertainty about the long-term effects of their actions (i.e., cluelessness)?**
    - How do people make decisions under cluelessness? Do they have a tendency to focus on certain positive short-term impacts even if the sign of the total long-term impacts is unclear and even if there are other options that have more robustly positive long-term consequences?
    - What if they have no precise probabilities? What if they know that they are not aware of all relevant possibilities (conscious unawareness), and so must deal with 'unknown unknowns'? Does it make them more cautious? Less so? When? Why?

2.2 Improving reasoning

- **How can rationality and wisdom in society as a whole be increased?**
- **Does improving rationality increase the effectiveness of people who aim to improve the world, or lead people to prioritize future wellbeing more highly?**
- **How can we foster better epistemic attitudes?**
    - Are there interventions that could lead to the adoption of a Scout Mindset? How can motivated reasoning be overcome? How long-lasting are these effects?
    - Do these interventions also make people prioritize improving the world?
- **How can we teach relevant mindware?**
    - What are the effects of Bayesian expected value training on different populations (e.g., college undergraduates)?
    - Does this training make people prioritize improving the world?
- **How can we improve reasoning amongst different populations?**
    - Can we improve rationality in policymakers, judges, young people, or even the general population?
    - Does such training make people prioritize improving the world?

2.3 Forecasting

- **How good are people at forecasting on (very) long time-scales?**
    - What things influence this? Which people are better and why? How do highly skilled forecasts differ from forecasts made by laypeople and domain-specific experts?
- **How can we improve forecasting, to make better real-world predictions about the future?**

    - How effective are various “debiasing” techniques at accounting for various biases in people’s probability estimates?
    - Can we evaluate predictions using techniques such as the willingness to place bets on opinions and publicly commit to positions?
    - How well do schemes like the Bayesian truth serum work for incentivizing truthful forecasting in contexts where the forecast is never resolved?
    - Can we improve predictions over long time horizons by creating a chain of forecasting tournaments such that each year forecasters predict the outcome of next year’s tournament?
    - What scoring rules are most effective for eliciting accurate forecasts?
    - How can people use new or existing statistical tools and AI models to improve their forecasting ability?
- **What inclines people to trust the forecasts of experts?**
    - What are obstacles that prevent people from taking expert forecasts seriously?

See also Section 1.2 on forecasting in the GPI Economics Research Agenda.

Relevant work:

Caviola, L., Schubert, S., & Greene, J. D. (2021). The Psychology of (In) Effective Altruism. *Trends in Cognitive Sciences.*

Schubert, S., & Caviola, L. (2024). *Effective Altruism And the Human Mind: The Clash Between Impact And Intuition.* Oxford University Press.

Baron, J. (2000). Thinking and deciding. Cambridge University Press.

Bazerman, M. H. (2020). *Better, Not Perfect: A Realist's Guide to Maximum Sustainable Goodness.* HarperCollins.

Daniel, K. (2017). *Thinking, fast and slow.*

Galef, J. (2021). *The Scout Mindset: Why Some People See Things Clearly and Others Don't.* Penguin.

Greenberg, S. (2022). Clearer Thinking Program Categorization.

Grossmann, I. (2017). Wisdom in Context. *Perspectives on Psychological Science,12*(2) 233–257.

Karger, E., Rosenberg, J., Jacobs, Z., Hickman, M., Hadshar, R., Gamin, K., Smith, T., Williams, B., McCaslin, T., Tetlock, P. (2023). Forecasting Existential Risks: Evidence From a Long-Run Forecasting Tournament.

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. Policy Insights from the Behavioral and Brain Sciences, 2(1), 129-140.

Rosenberg, J., Karger, E., Morris, A., Hickman, M., Hadshar, R., Gamin, K., Smith, T., Jacobs, Z., Tetlock, P. (2024). Roots of Disagreement on AI Risk: Exploring the Potential and Pitfalls of Adversarial Collaboration.

Sellier, A. L., Scopelliti, I., & Morewedge, C. K. (2019). Debiasing training improves decision making in the field. *Psychological science, 30*(9), 1371-1379.

Tetlock, P. E., & Gardner, D. (2016). Superforecasting: The art and science of prediction. Random House.

Yudkowsky, E. (2008). Cognitive biases potentially affecting judgment of global risks. Global catastrophic risks, 1(86), 13.

3 Individual differences

Not everyone is equally motivated and effective at improving the world. Research could investigate these individual differences. For example, research could study the psychological factors — values, personalities, cognitive traits — predicting whether someone is more likely to endorse the project of effectively improving the world and take

action accordingly. This may also entail the development of rigorous psychometric measurement tools that helps to reliably assess these factors. Conversely, psychological research could also study the psychological factors that predict reckless decision-making in high stakes situations that could greatly harm the future.

3.1 Identifying people who tend to either improve or harm the future

- **How do we identify people who will safeguard the future wisely? And how do we identify people whose decision-making tendencies pose a threat to the far future, e.g., by taking unnecessary risks with technology or pathogens?**
    - Can we develop assessment tools (e.g. standardized measures) to identify such people, e.g., by measuring the psychologically predictive factors?
    - How can such tests be developed and deployed in a scalable way such that they prevent cheating?
    - Could indicators of past rational or moral behavior be used as predictors of future rational or moral behavior?
    - Could nominators who report on other people's moral character, attitudes and tendencies help to identify people interested in effectively improving the world?
- **How can we prevent or mitigate the harm caused by malevolent leaders?**
    - How can we best identify malevolent people (e.g. psychopaths, narcissists, etc.)?
    - What leads individuals to support such malevolent leaders, and how can we prevent this?
    - How can we prevent malevolent leaders from manipulating minds (e.g., via misinformation, conspiracy beliefs, alarmist beliefs, promoting radical means-to-ends thinking)
- **How could rationality skills and thinking attitudes be measured?**
    - Can we develop better measures than the existing ones (e.g. CRT or Stanovich's CART)? Can we develop rationality measures that are sufficiently sensitive to be able to differentiate between highly rational individuals?
    - How can we measure the extent to which people understand and can apply rational mindware (thinking tools) such as probabilistic reasoning, economic thinking (e.g. expected value reasoning, understanding opportunity costs), scientific thinking, etc?
    - How can we measure rational thinking attitudes (or epistemic virtues), such as truth-seeking, intellectual honesty, intellectual modesty, nuanced reasoning, actively open-minded thinking, 'Scout mindset' (Galef, 2021)? How can such measures be improved to be more accurate and precise? For example, can behavioral measures be developed? Can measures be developed that are accurate even in competitive contexts (e.g., when respondents are incentivized to perform well)?

- Can we measure the effects of higher rationality in real-world contexts? Does greater rationality translate into improved judgments and decisions, or even into more successful outcomes in the personal, business and altruistic domain?

3.2 Psychological predictors of effectively improving the world

- **Which psychological factors predict whether a person will attempt to effectively improve the world?**
  - What non-cognitive predictors are there, such as values (expansive altruism, effectiveness-focus), interests, and motivations (determination to do what's moral)?
  - Why are some people more likely to take action to effectively improve the world than others, even if they agree equally strongly with the moral principles (i.e. attitude-behavior gap or willpower)?
  - What personality traits (e.g., optimizer mindset) predict interest in effectively improving the world?
  - What cognitive predictors are there, such as rational thinking skills (e.g. analytical reasoning) and attitudes?
  - What determines these predictive traits? Are they innate or acquired and how malleable are they?
- **Which people are less likely to find the project of effectively improving the world appealing and why?**
  - Do they have certain specific values that are at odds with the project of effectively improving the world, and if so, which and why? Or do they lack certain values which seem conducive to effectively improving the world, and if so, which and why?

Relevant work:

Adler, M. G., & Fagley, N. S. (2005). Appreciation: Individual differences in finding value and meaning as a unique predictor of subjective well-being. *Journal of personality*, *73*(1), 79-114.

Caviola, L. Althaus, D., Schubert, S., & Lewis, J. (2022). What psychological traits predict interest in effective altruism?

Caviola, L., Morrissey, E., & Lewis, J. (2022). Most students who would agree with EA ideas haven't heard of EA yet. Effective Altruism Forum.

Lovett, B.J., Jordan, A.H., & Wiltermuth, S.S. (2012). Individual Differences in the Moralization of Everyday Life. *Ethics and Behavior, 22*(4), 248-257.

Meindl, P., Jayawickreme, E., Furr, R. M., & Fleeson, W. (2015). A foundation beam for studying morality from a personological point of view: Are individual differences in moral behaviors and thoughts consistent?. *Journal of Research in Personality*, *59*, 81-92.

Moss. D. (2021). Effective Altruism Survey. Effective Altruism Forum.

Reynolds, S. J. (2006). Moral awareness and ethical predispositions: Investigating the role of individual differences in the recognition of moral issues. *Journal of Applied Psychology*, *91*(1), 233.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate?. *Behavioral and brain sciences*, *23*(5), 645-665.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of experimental psychology: general*, *127*(2), 161.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). The rationality quotient: Toward a test of rational thinking. MIT press.

Sun, J., & Goodwin, G. P. (2020). Do people want to be more moral?. Psychological Science, 31(3), 243-257.

4 Developmental

One approach to improve the world is to educate and empower the young generation with helpful tools and ideas. Psychological research could contribute to this project by empirically investigating how the psychological factors conducive to effectively improving the world emerge developmentally. It could also explore strategies that could help to foster critical thinking skills and pro-social values across different age groups.

4.1 Psychological factors related to improving the world across age

- **How do psychological tendencies related to improving the world and valuing future people emerge developmentally?**
    - What's the developmental trajectory of such psychological inclinations (e.g. moral expansiveness) and obstacles from childhood to adulthood?
    - How do children differ from adults in their inclinations towards improving the world?

4.2 Fostering critical thinking and pro-social values across age

- **Can rational reasoning skills and good epistemic virtues be taught and cultivated successfully in young people?**
    - Are there ways to foster a critical and open mindset in children in a long-lasting way? Who are children most likely to learn this from (e.g., peers, family, social media)?
    - To what extent can such approaches also be used with adults?
    - Does teaching children rationality skills make them more receptive to acting to improve the world than adults?
- **To what extent can young people's curiosity about valuing future generations be increased?**

Relevant work:

Bergman, R. (2002). Why be moral? A conceptual model from developmental psychology. *Human development*, *45*(2), 104-124.

Klaczynski, P. A., Byrnes, J. P., & Jacobs, J. E. (2001). Introduction to the special issue: The development of decision making. *Journal of Applied Developmental Psychology*, *22*(3), 225-236.

Kirby, J., Crimston, C. R., & Hoang, A. (2022). Compassionate Mind Training Can Increase Moral Expansiveness: A Randomised Controlled Trial.

Kohlberg, L., & Gilligan, C. (2014). Moral development. *Psychology: Revisiting the Classic Studies*, 164.

Marshall, J., Gollwitzer, A., Mermin-Bunnell, K., Shinomiya, M., Retelsdorf, J. & Bloom, P. (2022). How development and culture shape intuitions about prosocial obligations. *Journal of Experimental Psychology: General.* DOI: 10.1037/xge0001136.

Neldner, K., Crimston, C., Wilks, M., Redshaw, J., & Nielsen, M. (2018). The developmental origins of moral concern: An examination of moral boundary decision making throughout childhood. *PloS one*, *13*(5), e0197819.

Sommer, K., Nielsen, M., Draheim, M., Redshaw, J., Vanman, E. J., & Wilks, M. (2019). Children's perceptions of the moral worth of live agents, robots, and inanimate objects. *Journal of Experimental Child Psychology*, *187*, 104656.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Rational thinking and cognitive sophistication: development, cognitive abilities, and thinking dispositions. *Developmental psychology*, *50*(4), 1037.

Wilks, M., Caviola, L., Kahane, G., & Bloom, P. (2021). Children prioritize humans over animals less than adults do. Psychological Science, 32(1), 27-38.

5 Cross-cultural

Effectively improving the world is a global endeavor that requires an understanding of human psychology across many different cultures. Psychological research could contribute to that project by empirically investigating cross-cultural differences in psychological tendencies that are conducive to or hinder improving the world.

5.1 Cross-cultural differences in attitudes about and decision-making related to effectively improving the world

- **How do cultures differ in their values and thinking tendencies related to effectively improving the world?**
    - Do people from Western (WEIRD) cultures differ from people of other cultures (such as China, India, or Russia) in their extent of valuing future generations and using effective means to improve the world? What can such differences tell us about how to best cultivate values conducive to the project of effectively improving the world in different cultures? Are there potentially promising neglected countries for spreading the ideas of effectively improving the world?
    - Do people across cultures differ in how they think about the future of humanity, utopia, dystopia, or human extinction?
    - Do people across cultures have different views on policy issues related to nuclear proliferation, AI safety, or biosecurity?
    - How do different cultural values impact perceptions of what it means to effectively improve the world (e.g., do those from collectivist cultures think of focusing on distant others as more or less moral than those from individualistic cultures)?
    - What is the role of religion in predicting interest in the project of effectively improving the world?

Relevant work:

Henrich, J. (2020). The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous. Penguin UK.

MacAskill, W. (2022). What We Owe the Future. Hachette UK.

Olivola, C. Y., Kim, Y., Merzel, A., Kareev, Y., Avrahami, J., & Ritov, I. (2019). Cooperation and coordination across cultures and contexts: Individual, sociocultural, and contextual factors jointly influence decision making in the volunteer's dilemma game. *Journal of Behavioral Decision Making, 33*, 93-118.

Rhoads, S., Gunter, D., Ryan, R. M., & Marsh, A. A. (2020). Global variation in subjective well-being predicts seven forms of altruism. *Psychological science. 32*(8), 1247-1261.

Romano, R., Sutter, M., Liu, J. H., Yamagishi, T., & Balliet, D. (2021) National parochialism is ubiquitous across 42 nations around the world. *Nature Communications.* https://doi.org/10.1038/s41467-021-24787-1.

Van Doesum, N. J., et al. (2021) Social mindfulness and prosociality vary across the globe. *Proceedings of the National Academy of Sciences, 118*(35).

6 Historical psychology, evolutionary psychology, and anthropology

To improve the world it can be helpful to look into the past. Social sciences with a historical or evolutionary focus could contribute to that project. For example, such research could investigate how moral values relevant to improving the world have changed in the past, and based on this, estimate how they could plausibly change in the future.

- **How have moral values changed across history?**
    - How has the moral circle changed across history? What are examples of times when it may have contracted rather than expanded? Is the moral circle expansion we see at the societal level driven by cohort effects? To what extent have people included future generations in their moral circle?
- **How convergent are psychological traits over time?**
    - Is there convergence over time, e.g. for some moral values? And what does this imply about the future? What do moral values converge to?
    - How persistent are psychological traits, such as certain moral attitudes or cognitive traits, over historical time?
- **What can evolutionary psychology tell us about moral values (and indeed more broadly)?**
    - Does it say anything about which moral norms are feasible to cultivate in society on a long-term basis? Can it inform which relevant behaviors are strongly influenced by genetics, and therefore, what strategy is best for improving them? Can it improve our hypothesizing about relevant cognitive biases? For example, might people underestimate rates of technological change based on the slow rates in the environment of evolutionary adaptation?
- **What can evolutionary psychology tell us about the nature of well-being?**
    - What are the evolutionary origins of suffering and happiness? By understanding the purpose that different welfare states serve, can we identify relationships between beings' environments and their level of wellbeing?
    - Should we expect humans, other animals, and potential future digital minds to have positive, negative, or neutral welfare states by default?
- **How have people's evaluation of suffering and happiness changed historically?**
    - Did humans in the past or humans in other cultures (or philosophical traditions) value pain and pleasure, suffering and happiness differently than current WEIRD people do? For example, did/do some people find suffering

good or neutral, and why? And would valuing suffering make people more or less inclined to value future people, or to have an expanded moral circle, and why?

- **What are people's attitudes towards their ancestors?**
    - How do people tend to think about distant ancestors? How much does this vary across cultures?
    - Could this provide a potential data point relevant to how our descendants will think about us and consequently how much we should expect them to continue projects we begin or spend resources from our investments in ways we'd approve of?

See also Section 2.3 (Preparing to live alongside digital minds) in the GPI Philosophy Research Agenda.

Relevant work:

Branwen, G. (2019). The narrowing circle. https://www.gwern.net/The-Narrowing-Circle.

Calman, K. C. (2004). Evolutionary ethics: can values change. *Journal of Medical Ethics*, *30*(4), 366-370.

Haslam, H., McGrath, M.J., & Wheeler, M.A. (2019). Changing morals: we're more compassionate than 100 years ago, but more judgmental too. The University of Melbourne.

Pinker, S. (2012). The better angels of our nature: Why violence has declined. Penguin Books.

Roser, M. (2022). The world is awful. The world is much better. The world can be much better. Our World in Data.

Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P., & Henrich, J. (2019). The Church, intensive kinship, and global psychological variation. *Science, 366*(6466), eaau5141.

Tomasik, B. (2013). Differential intellectual progress as a positive-sum project. Center on Long-Term Risk.

Wright, R. (2010). *The moral animal: Why we are, the way we are: The new science of evolutionary psychology*. Vintage.

7 Policy and institutional

A key route to improving the world is through policy and institutional changes. Psychological research could contribute to this project by investigating the obstacles that prevent such helpful policy changes. Ultimately, it could also help to develop and test practical strategies that could be implemented to improve institutional decision making.

7.1 Psychological obstacles to effectively improving the world in policy contexts

- **From the perspective of effectively improving the world, what are the most dangerous psychological tendencies in a policy context?**
    - Are there ways to identify people with dark triad personality traits in policy contexts?
    - How does the basic underlying psychology of policymakers and other powerful people differ from the general public? To what extent do policy makers, focused on issues such as nuclear proliferation, international

cooperation, etc., show tendencies of short-term thinking, zero sum thinking, or thinking in terms of competition instead of cooperation?
    - How can we facilitate honest, productive, open-minded conversation and debate across groups and ideologies to promote effective and timely policy change?
- **How can we encourage institutional decision makers to take the long-term future into account?**
    - How effective are different messages, framings, or nudges?
- **What are the values, beliefs, and thinking tendencies of researchers doing potentially dangerous work?**
    - How do AI researchers think about risks from advanced AI?
    - How do synthetic biologists and biological researchers think about biological risks, such as accidental release of dangerous viruses?

7.2 Improving thinking and decision-making in institutional settings

- **How can psychological and behavioral research help to improve institutional decision making?**
    - Can we empirically test the feasibility of policies and real-world interventions to improve institutional decision-making, such as new institutional designs (e.g. representation of future generations), behavioral techniques such as incentives and nudges, and information campaigns?
    - What can we learn from previous public policy campaigns (e.g. anti-smoking, pro-exercise, etc.) to inform new ones about the project of effectively improving the world?
- **Can better values and mindsets be cultivated amongst specific groups with influence?**
    - Can we cultivate a security mindset amongst AI safety researchers and synthetic biologists?
    - How could we screen high-security lab workers for traits that could increase the probability of dangerous risks (e.g. lab leaks)?
    - Can we teach policymakers crucial rational reasoning skills, such as economic, scientific thinking tools, probabilistic reasoning, and expected value reasoning? (see Rationality section)
    - Can we cultivate a security mindset and an emphasis on avoiding worst-case outcomes amongst policymakers?
- **How should we expect humans to behave with respect to powerful potential new technologies?**
    - How risk-averse or risk-seeking are humans likely to be when it comes to developing and using potential developments in artificial intelligence, nuclear weapons, and bioweapons?

- How likely are future arms races in contexts like AI development? What can be done to mitigate the risks posed by such arms races?

See also Section 2.5 (Intergenerational governance and policy-making) in the GPI Economics Research Agenda and Section 1.1.2 (Non-welfare considerations) and Section 4.4.4 (Institutions) in the GPI Philosophy Research Agenda.

Relevant work:

Caviola, L. & Greene, J.D., Boosting the impact of human altruism. Manuscript in preparation.
Hauser, O. P., Rand, D. G., Peysakhovich, A., & Nowak, M. A. (2014). Cooperating with the future. *Nature, 511*(7508), 220–223.
Schoenmakers, K., Greene, D., Stutterheim, S. E., Lin, H., & Palmer, M. (2022). The Security Mindset: Characteristics, Development, and Consequences.
Whittlestone, J. (2017). Improving institutional decision making. 80,000 Hours report.
Winter, C., Schuett, J., Martínez, E., Van Arsdale, S., Araújo, R., Hollman, N., ... & Rotola, G. (2021). Legal Priorities Research: A Research Agenda.

8 Future wellbeing

An impartial altruist attempting to effectively improve the world is in significant part ultimately interested in increasing the future's aggregate wellbeing level. Research that helps to estimate the wellbeing levels of future individuals could therefore be action-guiding. In particular, a crucial question for such altruists is whether the future in expectation will contain much more happiness or much more suffering. This may determine whether an altruist is more likely to prioritize improving the quality of the future (i.e. reducing suffering and improving happiness) or reducing the chances of human extinction.

See also Section 1.1.1 (Welfare and beneficence) and Section 4.4.1 (Animal ethics) in the GPI Philosophy Research Agenda and Section 1.3 (Welfare and decision procedures) in the GPI Economics Research Agenda.

8.1 Estimating wellbeing in the future

- **How happy are people today and what can this tell us about people's wellbeing levels in the future?**
  - What proportion of people finds their lives valuable and worth living, i.e., better than neutral? Are there cross-cultural differences?
  - What predicts happiness? Are there demographic or cross-cultural differences?
  - What's the relative intensity and evaluation of negative versus positive experiences?
  - Is the baseline affect weakly positive? Is it also true of non-humans?
  - How could ongoing trends in technology affect human happiness in the distant future, and therefore, the value of reducing human extinction risk?

- **What biases do people have about predicting and evaluating wellbeing?**
    - Do people assess the value of others' lives differently from their own?
    - To what extent do positive illusions about the self contribute to people's feelings of well-being? Do depressed individuals make more realistic inferences, i.e., "depressive realism"? How should skepticism about the reliability of introspection inform our interpretation of self-reported subjective well-being? How widespread and serious is 'affective ignorance'?
    - Do these judgments affect people's evaluation of the importance of protecting the future?

8.2 Wellbeing in non-human beings

- **How happy are non-human beings?**
    - Do farmed animals and wild animals live net positive or negative lives respectively? What are the differences across species?
    - To what extent can we reliably attribute affectively valenced experience to entities of different kinds? What is the physiological basis of conscious experience and of affectively valenced experience?
    - To what extent do people base judgments of the value of the future on judgments of the welfare of future non-human beings?

Relevant work:

Diener, E., & Diener, C. (1996). Most people are happy. *Psychological Science, 7*(3), 181-185.

Killingsworth, M.A., Stewart, L., & Greene, J.D. (2021). Is life "worth living"? A measure of absolute happiness. Manuscript in preparation.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin, 103*(2), 193.

Haybron, D. M. (2007). Do we know how happy we are? On some limits of affective introspection and recall. *Nous*, 41(3), 394-428.

Waterman, A. S. (2007). On the importance of distinguishing hedonia and eudaimonia when contemplating the hedonic treadmill. *American Psychologist, 62*(6), 612–613.

Schwitzgebel, E. (2011). Perplexities of consciousness. MIT press.

Norris, C. J., Gollan, J., Berntson, G. G., & Cacioppo, J. T. (2010). The current status of research on the structure of evaluative space. Biological psychology, 84(3), 422-436.

Martínez, E., & Winter, C. (2021). Protecting Sentient Artificial Intelligence: A Survey of Lay Intuitions on Standing, Personhood, and General Legal Protection. Frontiers in Robotics and AI, 8.

Shulman, C. (2012). Are pain and pleasure equally energy efficient? Reflective Disequilibrium.

O'Brien & Kassirer (2018). People are slow to adapt to the warm glow of giving. *Psychological Science, 30(2), 193–204.*

Happier Lives Institute. Research Agenda. https://www.happierlivesinstitute.org/research/research-agenda/

DeYoung, C. G., & Tiberius, V. (2021). Value fulfillment from a cybernetic perspective: A new psychological theory of well-being. *Personality and Social Psychology Review*, 1088868322108377.

Williams, L. A. (2021). From human wellbeing to animal welfare. Neuroscience & Biobehavioral Reviews, 131, 941-952.

9 Cause-specific issues

Here we focus on global problems that seem particularly pressing from the point of view of effectively improving the world. In general, it's useful to study the views and psychological

tendencies people show that relate to these specific issues. This includes, for example, the study of psychological obstacles preventing people from taking these issues seriously. Note that the cause areas listed below are chosen because they seem particularly neglected from the perspective of effectively improving the world. There are many other important cause areas (e.g. prejudice, climate change) that aren't listed because they seem relatively less neglected in academic psychology.

See also Section 1.3.1 (Extinction and other catastrophic risks), Section 2.3.1 (Catastrophic risks and their mitigation), Section 3.1 (Catastrophic risk from AI) in the GPI Philosophy Research Agenda and Section 2.1 (Economics of Catastrophes) in the GPI Economics Research Agenda.

9.1 Global catastrophic risk (in general)

- **How do people think about risks that could permanently curtail the future of humanity?**
    - Do people underestimate or overestimate such risks? Do they underappreciate the importance of mitigating such risks?
- **What do people think about emerging technologies that have the potential to cause global catastrophic risks?**
    - Are they too optimistic or pessimistic relative to the views of experts? What drives these attitudes?
- **What do people think about human extinction and the future of humanity?**
    - Do people find human extinction good or bad, and why? Do people find it morally important to safeguard the future of humanity?
    - What are people's empirical beliefs about the likelihood, causes, and mitigation of human extinction?
- **Which people don't find it bad if humanity went extinct and why?**
    - Is this attitude in part driven by non-utilitarian values?
    - (see Morality section)

Relevant work:

Bostrom, N., & Cirkovic, M. M. (Eds.). (2011). *Global Catastrophic Risks*. Oxford University Press.

MacAskill, W. (2022). *What We Owe the Future.* Hachette UK.

Schubert, S., Caviola, L., & Faber, N. S. (2019). The psychology of existential risk: Moral judgments about human extinction. *Scientific reports, 9*(1), 1-8.

Yudkowsky, E. (2008). Cognitive biases potentially affecting judgment of global risks. *Global Catastrophic Risks, 1*(86), 13.

9.2 Risks from artificial intelligence (AI)

- **What are people's views on the argument that advanced AI, such as artificial general intelligence (AGI), poses an existential threat to humanity?**
- **How do AI researchers think about risks from advanced AGI?**
    - How concerned are AI researchers about risks from uncontrolled AGI, and why?

    - How do AI researchers' views differ from those of other populations? Are they more or less concerned?
- **What are experts' opinions on AGI timelines and take-off scenarios?**
    - What are people's views on when, how, and how fast transformative AI will be developed and deployed?
- **How can we raise awareness of AGI risk among AI researchers?**
    - How do we create group dynamics and systems that ensure safe development of AGI?
    - (See Policy section above)
- **Can we transfer insights from human psychology to the cognition and behaviour of AI?**
    - Can our understanding of human cognition help to interpret AI systems and ensure that they are safe?
- **How could the empirical study of human moral values be useful for the alignment of AGI?**
- **What cognitive biases could undermine efforts to align AGI?**
    - Do people underappreciate how training an AGI on one task makes it competent to do another task?
    - Do people succumb to the curse of knowledge when trying to predict AI's ability to infer how humans want it to behave?
- **What do people think about digital sentience?**
    - Do they believe that digital beings can be sentient? If not, why?
    - Do people morally value or discount (different types of) potential digital beings?
- **How could advanced AGI affect human wellbeing, given different plausible future AGI scenarios?**
    - How would human well-being be affected in a world in which all human labor is fully automated? How would human wellbeing be affected in a world with total surveillance, e.g., in an authoritarian system?

See also Section 2.3 (Preparing to live alongside digital minds) in the GPI Philosophy Research Agenda and Digital Minds in Society: Research Agenda (Caviola, 2024; available upon request).

Relevant work:

Bensinger, R. (2021). "Existential risk from AI" survey results, AI Alignment Forum. Link.

Caviola, L. (2024). Digital Minds in Society: Research Agenda. Available upon request.

Clarke, S., Carlier, A., & Schuett, J. (2021). Survey on AI existential risk scenarios, EA Forum. Link.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research, 62*, 729-754.

Irving, G., & Askell, A. (2019). AI safety needs social scientists. *Distill, 4*(2), e14.

O'Shaughnessy, M., Schiff, D., Varshney, L. R., Rozell, C., & Davenport, M. (2021). What governs attitudes toward artificial intelligence adoption and governance?. Preprint.
Pauketat, J. V., & Anthis, J. R. (2022). Predicting the moral consideration of artificial intelligences. Computers in Human Behavior, 136, 107372.
Pauketat, J. V., Ladak, A., & Anthis, J. R. (2022). Artificial Intelligence, Morality, and Sentience (AIMS) Survey: 2021.
Zach Stein-Perlman, Benjamin Weinstein-Raun, Katja Grace, "2022 Expert Survey on Progress in AI." AI Impacts, 3 Aug. 2022. https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/.
Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., & Dafoe, A. (2021). Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers. Journal of Artificial Intelligence Research, 71, 591-666.
Zhang, B., & Dafoe, A. (2020, February). US public opinion on the governance of artificial intelligence. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 187-193).
Zhang, B., & Dafoe, A. (2019). Artificial intelligence: American attitudes and trends. Available at SSRN 3312874.
Zhang, B., Dreksler, N., Anderljung, M., Kahn, L., Giattino, C., Dafoe, A., & Horowitz, M. C. (2022). Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers. arXiv preprint arXiv:2206.04132.

9.3 Pandemics and biosecurity

- **How do biological scientists (e.g. working on synthetic biology) think about biological risks, such as accidental release of dangerous viruses?**
    - Do they underestimate the risks of accidents? Do they underestimate the extent to which their research could directly or indirectly be used by malevolent actors?
- **How can we raise awareness of biological risks, such as accidental pandemics, amongst biological scientists?**
    - To what extent can we draw useful insights from past pandemics such as the Spanish flu or Covid-19?
- **How can we improve institutional decision-making (e.g. amongst policy makers) with regard to pandemic preparedness?**
- **What are the attitudes of the general public towards pandemic preparedness?**
    - Do they underestimate the risks? How can we tackle anti-vaccination attitudes?

Relevant work:

Carus, W. S. (2017). A century of biological-weapons programs (1915–2015): reviewing the evidence. *The Nonproliferation Review*, 24(1-2), 129-153.
Gronvall, G. K. (2016). *Synthetic biology: Safety, security, and promise*. Health Security Press.
Inglesby, T. V., & Relman, D. A. (2016). How likely is it that biological agents will be used deliberately to cause widespread harm? *EMBO reports*, *17*(2), 127-130.
Kilbourne, E.D. (2011). Plagues and pandemics: past, present, and future. In N. Bostrom & M. M. Cirkovic, *Global catastrophic risks*. Oxford University Press.
Koblentz, G. (2003). Pathogens as weapons: the international security implications of biological warfare. *International security*, 84-122.
Koehler, A., & Hilton, B. (2020). Preventing catastrophic pandemics. 80,000 Hours.
Millett, P., & Snyder-Beattie, A. (2017). Human agency and global catastrophic Biorisks. *Health security*, *15*(4), 335-336.
Monrad, J. T. (2020). Ethical considerations for epidemic vaccine trials. *Journal of medical ethics*, *46*(7), 1-5.
Nouri, A. & C.F. Chyba. (2011). Biotechnology and biosecurity. In N. Bostrom & M. M. Cirkovic, *Global catastrophic risks*. Oxford University Press.

Racicot, M., Venne, D., Durivage, A., & Vaillancourt, J. P. (2012). Evaluation of the relationship between personality traits, experience, education and biosecurity compliance on poultry farms in Québec, Canada. *Preventive Veterinary Medicine*, 103(2-3), 201-207.
Salvatore, S. (2017). Psychological Evaluations for the US Army Biological Personnel Reliability Program. *Journal of Biosecurity, Biosafety, and Biodefense Law*, 8(1), 3-17.
Vanlandingham, D. L., & Higgs, S. (2021). Viruses and Their Potential for Bioterrorism.

9.4 Global war and nuclear threat

- **What do people think about nuclear proliferation?**
    - Do people across different countries (e.g., USA, Russia, China, India) differ in their views on nuclear proliferation and disarmament?
    - How do cognitive biases (see Rationality section) impact thinking about the risk of nuclear threat and global war?
- **How can we raise awareness of nuclear threats?**
    - How effective are different messages or framings?
- **What leads to the endorsement of extremist beliefs and behaviors (e.g., war and terrorism)?**
    - How do *contextual* factors (e.g., poverty, inequality, religious or ideological education, educational attainment) influence the uptake of extremist beliefs and behaviors?
    - How do *psychological* factors (e.g., self-deception, cognitive dissonance, close-mindedness, dehumanization) influence the uptake of extremist beliefs and behaviors?
    - How do *social* factors (e.g., type of intergroup conflict, history of oppression, conflicting ideologies) influence the uptake of extremist beliefs and behaviors?

Relevant work:

Ellsberg, D. (2017). *The Doomsday Machine: Confessions of a Nuclear War Planner*. Bloomsbury Publishing USA.
Graham, A. (2017). *Destined for War Can America and China Escape Thucydides's Trap?* HarperCollins.
Hoffman, D. (2009). *The Dead Hand: The Untold Story of the Cold War Arms Race and its Dangerous Legacy*. Anchor.
Hogg, M. A., & Blaylock, D. L. (Eds.). (2011). *Extremism and the Psychology of Uncertainty* (Vol. 3). John Wiley & Sons.
Loza, W. (2007). The psychology of extremism and terrorism: A Middle-Eastern perspective. *Aggression and Violent Behavior*, *12*(2), 141-155.
Schlosser, E. (2013). *Command and control: Nuclear weapons, the Damascus accident, and the illusion of safety*. Penguin.
Suedfeld, P., & Tetlock, P. (1977). Integrative complexity of communications in international crises. Journal of conflict resolution, 21(1), 169-184.

9.5 Threats to liberal democracies and descent into long-term authoritarianism lock-in

- **How can values, such as reason, evidence, liberalism, tolerance, and democracy, be retained and strengthened in a cooperative way?**
    - How can we help political groups to cooperate in an era of polarization?

    - What leads individuals to (correctly or incorrectly) believe their democracy and liberty are being threatened, and what impacts these individuals' strategies to regain these values?
    - How do intellectual humility and open-minded (vs. close-minded) thinking influence our ability to cooperate with those we disagree with? How can we increase such humble and open-minded thinking?
- **What factors could increase the risk of a collapse of liberal democracies in Western countries and the establishment of authoritarian systems?**
    - To what extent does political polarization increase that risk?
    - How do misinformation and conspiracy beliefs proliferate throughout a liberal democracy, and how does this threaten democratic systems? How can we prevent the spread of false information and beliefs?

Relevant work:

Dikötter, F. (2022). *How to Be a Dictator: The Cult of Personality in the Twentieth Century*. Bloomsbury Publishing.

Johnson, S. A. (2019). Understanding the violent personality: Antisocial personality disorder, psychopathy, & sociopathy explored. *Forensic Research & Criminology International Journal*, *7*(2), 76-88.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094-1096.

McCoy, J., & Press, B. (2022). What Happens When Democracies Become Perniciously Polarized? Carnegie Endowment for International Peace.

Smith, A., & de Mesquita, B. B. (2011). *The Dictator's Handbook: Why Bad Behavior is Almost Always Good Politics*. PublicAffairs.

Path to impact

Psychological research — in particular, research on people's beliefs, judgments, and decisions — can inform global priorities research in multiple ways. Many pivotal moments in humanity's future will ultimately depend on human decision-making. For example, we may be faced with choices like whether to create potentially dangerous new technologies, invest in pandemic preparedness, or engage in nuclear conflicts. Understanding the individual and social psychology of these decisions can enable global priorities researchers to better understand the risks they pose, and to design institutions and interventions to ensure such decisions are made wisely. Moreover, policies suggested by global priorities research often depend on human psychology. What incentives are sufficient to persuade pathogen researchers to take sufficient precautions against laboratory leaks? What factors persuade voters to elect potentially dangerous leaders? What long-term trajectories will facilitate human psychological wellbeing? Psychological research to answer these kinds of questions has many plausible paths to impact. Three primary paths to impact are providing key decision-makers with action-guiding information, raising public awareness of topics in global priorities research, and developing tools to enable decision-making to achieve better long-term outcomes.

Action-guiding information

Some psychological insights are directly action-guiding for decision-makers looking to effectively improve the world. While it is important to publish the basis for these insights in high-quality, peer-reviewed journals to ensure their rigor and credibility, it is also important to communicate these insights directly to policymakers, philanthropists, and other key decision-makers. For example:

- We could identify the key decisions that humans are likely to make suboptimally for future welfare (e.g., risking a catastrophe to win a technological arms race) and intervene specifically on those key decisions.
- We could examine people's attitudes toward future people to assess the political tractability of policies informed by global priorities research.
- To compare the value of interventions to promote welfare (including, e.g., existential risk reduction), we may need to design or identify better methods for assessing and predicting psychological wellbeing.
- To decide who should make decisions affecting the future, we may need to understand what psychological traits predict sound judgments.
- We could collect empirical data to allow decision-makers to better predict the societal reaction to policy.

Education and awareness raising

Given the potentially revisionist implications of global priorities research, it is important to disseminate these findings with careful consideration for how they will be assimilated into broader societal views. Academic research is often the first step to raising broader awareness of important scientific insights. Insights that are rigorously researched and published in academic journals may, with an increased likelihood, be featured in university education, textbooks, journalistic outlets, or popular science books. As an example, consider the widespread awareness of cognitive biases owing to Kahneman and Tversky's research, and its undermining of the case for laissez-faire policy based on rational actors. Analogously, if psychological insights relevant to global priorities research become integrated into public consciousness, it might lead to a more nuanced and appropriate weighting of ideas in global priorities research by individual decision-makers, philanthropists, and the policy community. For example:

- Public support for interventions identified by global priorities research could increase with knowledge of common decision-making errors that lead to catastrophic risks.
- Public support for policy informed by global priorities research could decline with knowledge of various decision-making errors that humans frequently commit when attempting to improve the world.

In addition, psychologists are well-positioned to conduct research directly focused on raising awareness about topics in global priorities research. For instance:

- Researching which moral arguments move people to prioritize the long-term future of humanity appropriately.
- To inform public outreach efforts, researching which topics in global priorities research people find the most compelling and/or easiest to understand.

Practical interventions and tools

A more direct approach is to develop and test practical strategies that could empower people to do more good or prevent them from causing harm. For example:

- We could develop training programs to alert scientists to any catastrophic risks that could result from their research or persuade policymakers to put greater priority on reducing existential risk.
- We could provide decision-making tools that help people appropriately weigh low probability risks.
- We could develop rationality courses to improve public reasoning and reduce the risk of electing dangerous leaders.
- To reduce risks posed by malevolent actors in positions of power, we could develop and test psychometric tools to help identify which people are the best at reasoning about high-stakes decisions, and are the most inclined to promote wellbeing.

From a global priorities research perspective, interventions with potential benefits extending 100 years or more into the future are often more promising than those with short-term impacts but uncertain long-term effects. For instance, behavior change interventions might be superseded by more effective solutions in the future, making their impact relatively short-lived. In contrast, interventions aimed at preventing existential catastrophes have the potential to benefit humanity for thousands of years or more.

Contributors

Main authors:
Lucius Caviola, Joshua Lewis, Matti Wilks, Abigail Novick Hoskin, Stefan Schubert, Carter Allen, Johanna Salu

For helpful comments we thank Adam Bales, Adam Bear, Andreas Mogensen, Brad Saad, Christian Panzer, David Althaus, David Thorstad, Erin Morrissey, Falk Lieder, Geoffrey Goodwin, Hayden Wilkinson, Inga Grossmann, Izzy Gainsburg, Jessie Sun, John Firth, Julian Jamison, Matt Coleman, Mattie Toma, Maximilian Maier, Noemi Dreksler, Samantha Kassierer, Sven Herrmann, Teruji Thomas and Will Fleeson.