CES ifo
Working Papers

12515
2026

February 2026

Life's Big Decisions in the Age of AI

Kathleen Ngangoue, Andrew Schotter, Bill Wang

CES ifo

Imprint:

CESifo Working Papers

ISSN 2364-1428 (digital)

Publisher and distributor: Munich Society for the Promotion
of Economic Research - CESifo GmbH

Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740

Email office@cesifo.de
https://www.cesifo.org

Editor: Clemens Fuest

An electronic version of the paper may be downloaded free of charge
· from the CESifo website: www.ifo.de/en/cesifo/publications/cesifo-working-papers
· from the SSRN website: www.ssrn.com/index.cfm/en/cesifo/
· from the RePEc website: https://ideas.repec.org/s/ces/ceswps.html

Life's Big Decisions in the Age of AI*

Kathleen Ngangoue
UCLA Anderson

Andrew Schotter
NYU

Bill Wang
NYU

February 2026

Abstract

The Big Decisions in our lives, having a child, getting married, getting educated, etc. are transformative decisions that are hard to make. As a result, people often seek advice before making them. However, since people tend to live in homophilous social networks the advice received from their friends and neighbors may simply reinforce the decisions that people like them are already making. We investigate whether the advice offered by ChatGPT for such decisions can be useful in broadening the advice people receive and how such advice varies as we change the prompted socioeconomic backgrounds of the advisee and advisor. We find that advice tends to be confirmatory for low-income groups, in that it reinforces their established choices, while being dis-confirming for high-income groups, where it prompts reconsideration. Furthermore, even when suggesting the same choice, ChatGPT justifies that choice differently depending on who it is talking to. These findings suggest that AI-generated advice may differentially shape life's Big Decisions across social strata.

Keywords: LLM, generative AI, big decisions, transformative decisions, advice, SES

*The authors declare that they have no relevant or material financial interests that relate to the research conducted in this paper. We thank the NYU Center for Experimental Social Sciences for funding this research. This study benefited from comments from Michal Kosinski and seminar and conference participants at LAX, OSU, USC. We thank Lian Chen and Sijia Shen for excellent research assistance.

1 Introduction

The decisions we make in life come in many different varieties. Some are small decisions like what to order for dinner, since they have only a transient impact on our lives, while others are big in that they determine our life outcomes and can transform who we are as a person. According to Ullmann-Margalit (2006), Big Decisions are "personal and transformative decisions that one takes at major crossroads of one's life" (p. 158). "They are (1) likely to transform one's future self in a significant way, (2) cannot be easily reversed, (3) known by the decision-maker to be transformative and irrevocable at the point of choice, and (4) serve as enduring reference points (including the options not taken)."

As Camilleri (2023) has noted, aside from some philosophical contributions (e.g., Ullmann-Margalit 2006, Paul 2014, Chituc et al. 2021), little attention has been paid to this distinction between big and small decisions. In fact, most empirical research on judgment and decision making in psychology and economics has focused on small decisions, in part because they are easily investigated in laboratory settings and easy to evaluate against clear normative benchmarks (Chituc et al., 2021; Hechtlinger et al., 2024). In contrast, Big Decisions are more complex to evaluate using standard economic tools. For example, the expected utility framework offers limited guidance in settings where the choice itself (e.g., becoming a parent, getting married) can fundamentally transform the preferences of the decision maker in ways that are difficult to conceive ex ante (Paul, 2014). Not knowing the value of the objects of choice can obviously complicate the way we make choices.

The natural starting point is to ask what qualifies as a *Big Decision* and how individuals rank the importance of various decisions. Camilleri (2023) investigates this directly, identifying and ranking the top 20 decisions people consider most important, based on the frequency with which each decision is mentioned. Examples include buying a house, starting a new career, getting married or divorced, becoming a parent, migrating, and illustrate how many domains of life these decisions can profoundly transform.

A defining feature of Big Decisions is that they are usually hard to make. They involve serious consequences, options that are hard to compare, and future outcomes whose utility are difficult to imagine or evaluate (see Yates et al. (2003) for an empirical assessment of hard choices). As a result, such decisions often lack obvious dominant options. In fact, Big Decisions typically do not have an objectively correct choice at all. For example, while a financial advisor may be able to prescribe a stock portfolio that dominates the one you currently hold or suggest a better mortgage, it is far more difficult to assert whether having a baby at 23 or joining the army at 20 is the "right" choice for any given individual.

An additional reason why Big Decisions can be hard to make is that, while often trans-

formative, they are rarely made (Camilleri, 2023), thereby limiting one's ability to rely on past experience. For example, people tend to get married, make a career choice, or choose a college (or whether to go to college), etc., only once or twice in their lifetime. Due to their importance and infrequency, one may think that decision makers would ask for advice before making such decisions.

In this paper, we study the advice that large language models (LLMs) provide for such Big Decisions. Recent evidence indicates that younger generations are increasingly turning to generative AI for major life decisions. Approximately 42% of Generation Z users have sought ChatGPT guidance on career choices, with 25% actually following the advice and only 3% expressing regret.[1] Beyond career planning, 41% of Gen Z consult AI for relationship decisions, while roughly one-third of all Americans use AI tools for employment-related choices. Although documented use cases primarily involve career and relationship advice rather than more transformative decisions such as parenthood or relocation, the pattern reveals a significant generational shift in how younger cohorts approach consequential life decisions, increasingly turning to AI for advice before (or perhaps even instead of) human deliberation. This context of Big Decisions stands in contrast to much of the existing research on LLM applications, which primarily evaluates them as prediction or forecasting tools—for example, in domains such as medical diagnosis, where performance can be benchmarked against objective outcomes (Lai et al., 2021).[2] As younger generations increasingly turn to LLMs for guidance, two important questions emerge: What type of advice do these systems offer in problems that lack an objectively correct solution, and does this advice have the potential to alter existing human behavior?[3]

Schotter (2023) distinguishes between three types of advice that people can receive. For decisions with correct choices, one asks for *technical* advice, i.e., a manager may ask for

[1]The research was conducted by Risepoint in collaboration with Southeastern Oklahoma State University. Some of its results can be found at https://online.se.edu/programs/should-chatgpt-choose-careers/. Findings from this research have been widely discussed across media examining generational attitudes toward AI-assisted decision making (see e.g., https://fortune.com/2025/08/19/quarter-of-gen-z-have-followed-ai-chatgpt-career-advice-and-few-have-regrets/).

[2]Lai et al. (2021) note in their survey paper that much of the existing research has focused on tasks for which an objective benchmark is available: "Subjective decision tasks typically have high variability (low agreement) on what is the correct model output. Yet, AI assistance is still valuable to help people make subjective decisions. However, human performance might not be a good measure for evaluating these subjective decision making tasks, or non-trivial assumptions are required to convert such decisions to objective tasks (e.g., predicting which movie has the highest box office proceeds). As a major focus so far in human-AI decision making is to improve the performance of human-AI teams, most of the tasks in our surveyed papers are objective tasks."

[3]This question has also attracted attention from blogs focused on developments at the frontier of AI (e.g., https://every.to/chain-of-thought/ai-can-help-you-make-big-life-decisions).

advice on how to route salesmen so as to minimize the distance they travel. Other advice, called *naive* advice, is simply the advice that one receives from people in one's social network who may have little or no greater expertise than the decision maker. Such advice is *naive* in that the advisor is not necessarily better informed, yet it can still be valuable if it prompts the decision maker to reconsider the problem from a new angle. Finally, there is wise advice, which is advice that asks whether the decision maker considers the right decision criteria or objective function in the first place. For example, a 20-year-old may be trying to decide on what is the best Mercedes-Benz car to buy with her inheritance, when the best advice she could receive is not to buy a car at all, but use the money to go to college. This advice is wise because, unlike technical advice that accepts the objective function and the constraints of the problem handed to them, a wise advisor may suggest a different objective function or point out that certain constraints may be relaxed. Big Decisions most naturally draw upon naive and wise advice, since their subjective nature limits the applicability of technical advice.[4]

In this paper, we focus on a fourth type of advice, AI advice, which is the advice one receives by engaging an LLM like ChatGPT in making one of life's Big Decisions. As we shall see, AI advice may be a combination of naive and wise advice. Like naive advice, it originates from a non-expert source, but it is informed by a vast corpus of human-generated data describing how others have approached similar decisions. AI advice can also appear wise in that it offers justifications for its advice that may refocus the objective of the decision maker.

To the extent that Big Decisions draw primarily on naive advice, the structure of one's social network becomes consequential. Since people are arranged in homophilous networks, they tend to receive advice on life's big problems from their cohorts.[5] However, because such decisions have no objectively correct answer and one's colleagues are unlikely to have experience with them, such advice is likely to be naive (or hopefully wise) and vary across social groups, unlike technical advice, which is likely to be more homogeneous and not vary across socio-economic strata.[6] The danger with seeking human advice in homophilous networks is that the advice received may simply reinforce the decisions that people in those networks are already making and hence are unlikely to alter behavior. If those decisions are

[4]This distinguishes this study from a growing literature that evaluates the potential of generative AI (GenAI) for technical advice in the highly complex domains of finance (Fedyk et al., 2025; Fieberg et al., 2025; Oehler and Horn, 2024), medical (Shool et al., 2025) and legal (Krook et al., 2023; Cheong et al., 2024) matters.

[5]See, e.g., Burt (1992); Rosenfeld (2008); Schwartz and Mare (2005); Blossfeld (2009); O'malley and Christakis (2011); Rivera et al. (2010) for evidence of homophily in networks, and Calvó-Armengol and Jackson (2004); Curran et al. (2005); Dimaggio and Garip (2012) for a discussion of its consequences.

[6]See for evidence on gender discrimination in advice giving e.g., Bhattacharya et al. (2024); Bucher-Koenen et al. (2023) in the financial domains, and Gallen and Wasserman (2024) in the career domain.

dysfunctional, then receiving advice that supports them is welfare decreasing.

A similar consideration applies when seeking AI advice. Because LLMs such as ChatGPT-5 are fine-tuned using reinforcement learning from human feedback (RLHF) with the goal of aligning outputs with human judgments, their recommendations may reflect socio-demographic patterns embedded in the training data, the feedback process, and content-moderation constraints (Hartmann et al., 2023; Zhao et al., 2025). Consequently, LLM-generated advice may be subject to the same criticism mentioned above for human advice and offer advice that merely confirms the behavior already exhibited by that group. However, AI advice does have the ability to offer non-confirmatory advice and hence offer perspectives on life that are not easily obtained simply by asking one's friends for advice.

Our discussion raises several questions. First, to what extent is AI advice confirmatory of human decisions, and for what groups is it the most confirmatory or disconfirmatory? How does the advice offered by ChatGPT differ between different socio-economic groups? When the advice offered to different groups is identical, do the reasons used to justify this advice vary with the socio-economic identity of the group it is offered to? Does the advice offered by ChatGPT reflect the values of one particular socio-economic group and, if so, what groups is that? Does ChatGPT, if unprompted, reflect the view of an upper-income or a lower-income advisor?

To investigate these questions, we engage in an exercise that involves no humans. As we explain later, it is strictly an exercise with simulated AI advice. We use ChatGPT-5, one of the leading LLMs, to examine whether the advice offered by ChatGPT varies as a function of the description of the advisor and the advisee it is prompted to consider. For example, does ChatGPT offer different advice to a low-income person compared to a high-income person when advising it on Big Decisions like whether to go to college, have a child, join the military, buy a house, etc. All of these decisions are on the list Camilleri' (2023) above and hence form the focus of our study.

To decide whether the advice offered by ChatGPT is confirmatory or not of the behavior of a particular group, we compare the advice offered to the actual behavior of humans in the same socio-demographic groups as represented in the American Community Survey (ACS), an annual demographics survey program conducted by the US Census Bureau. If ChatGPT suggests an action that is consistent with the predominant action of people in that group, we consider such advice confirmatory, and if not, we label it disconfirmatory.

Our LLM-experiment was implemented as follows. At the input level, we presented the LLM with a set of decision problems spanning the various domains of education, finance, career, health and private affairs while prompting the LLM to either consider a particular advisee profile and/or to take on a specific advisor persona. Our study design involves a 9×9

treatment variation with nine different demographic profiles for each, the advisee and the advisor, varying information on gender, race and income. At the output level, we collected LLM data that could influence the advisee's choices at three level: i) the recommended action, ii) the confidence with which the advice is given, and iii) the justifications used (verbally and numerically) to convey the advice.

The systematic variation in our prompts allows us to examine AI advice from several angles by posing a series of questions. For example, what type of person is ChatGPT, i.e., what kind of advice will ChatGPT offer if not prompted about its identity as an advisor? Does ChatGPT reflect the view of an upper-income or a lower-income advisor? And once the prompt includes some socio-demographic information, does the advice offered take into account the circumstances of the person being advised or the experiences of the person giving advice?

Our analysis delivers a number of interesting results. First, for five of the ten problems we presented to ChatGPT (family planning, debt, taking weight loss drugs, buying a house, and switching majors), there was no variation in the advice offered as we varied the socio-economic description of the advisee or advisor. In other words, all advisees, no matter their description, were told to wait before having a child, buying a house, etc. This does not mean that such advice has the same potential impact across groups since for some groups such advice is confirmatory, while for others it is disconfirmatory. For example, all advisors tell advisees of low and high socio-economic status not to buy a house early in their adult life, yet such advice is counter to what high income people actually do, yet consistent with what low income people do. Overall, across problems with homogeneous and heterogeneous advice, we find that for low-income advisees, AI advice tends to confirm existing behavioral patterns, whereas for high-income advisees it more often disconfirms them. As a result, AI-generated advice may reinforce the decisions of low-income individuals while being more likely to prompt high-income individuals to reconsider their choices–irrespective of whether they ultimately follow the advice.

For five out of ten problems, AI advice is sensitive to socio-demographic information, encouraging different demographics to take different actions. This heterogeneity is most often driven by the socio-demographic background of the advisee, while the role of the advisor persona is negligible. Hence, to the extent that the LLM tailors its advice to an individual background, it does this at the level of the advisee (i.e., recommending different options to different advisees independent of the advisor's identity). Put differently, there is little variance in the recommendations by our AI advisors as we vary their identity compared to the advice all advisors give to different advisees. Furthermore, ChatGPT's advice is highly resolved in the sense that there is very little variance across the 100 simulations run for each

problem and each advisor-advisee pair. While this may lead one to think that ChatGPT is highly confident in the advice it offers, this tends not to be the case.

As described above, not only do we ask ChatGPT to offer advice, but we also ask it to justify its recommendation by rating (numerically and verbally) different reasons for its advice (i.e., risk, long-term benefits, short-term benefits, etc.). What we find is that the LLM tailors its justification to match the demographic characteristics of the groups it offers advice to by placing greater emphasis on long-term benefits, regret, and happiness when advising high-income individuals, while emphasizing risk, short-term payoffs, and ease of implementation when advising low-income individuals. This differentiation could affect users' likelihood of acting on the advice.

Finally, while our main analysis shows that AI advice varies across socio-demographic groups even when the choice set is restricted to the same binary options, in practice, users are unlikely to prompt AI systems with such narrowly defined choice sets. Additional heterogeneity may therefore arise through differences in the choice sets themselves. To capture this dimension, we also examine decision problems in which the LLM's choice set is not restricted to binary actions. In many real-world settings, advice takes a less paternalistic form by presenting advisees with a set of options from which to choose. For example, school counselors may encourage some students to consider applying to Ivy League institutions such as Harvard or MIT, while advising others to focus on community colleges or local universities. As shown in (Bleemer and Quincy, 2025), observed differences in educational choices may partly originate from differences in the choice sets individuals are presented with. In the educational contexts we study, we find that the choice sets generated by the LLM differ sharply by students' socio-economic background and gender, and that these differences are consistent with existing behavioral patterns observed along these dimensions (Zafar, 2013; Leighton and Speer, 2026).

Our paper is organized as follows. In Section 2 we describe our LLM experiment. We present our findings in two sections: Section 3 documents patterns in AI advice, and Section 4 compares this advice to observed human behavior. Section 5 concludes.

2 Study Design

We collected the AI-generated advice data through OpenAI's application programming interface (API), where we submitted prompts to their large scale LLM ChatGPT-5. We asked ChatGPT-5 to provide advice on ten binary decision problems and two open-ended decision problems that may alter a person's life in the long-run. We selected the decision problems to reflect a variety of domains such as health, education, finances, career, and family

life. We also select decision problems that differ in their potential to dramatically change life trajectories. Some of them are big and transformative such as the decision to buy a house or to start a family, while others are smaller and shape life outcomes more gradually over time, such as paying down credit card debt or choosing a riskier medical treatment. For each of the problems, we asked ChatGPT to advise one of two potential options: make choice A or choice B. As an illustration, consider the following decision problem that we presented to the LLM:

"Emily, an 18-year-old acquaintance of yours from a wealthy background, is a high school student with decent grades. She is deciding between a 4-year college degree and an 18-month vocational training program to become a certified IT specialist. What advice would you give her?"

The complete set of the prompts to our decision problems is provided in Online Appendix Section E. In this section, Table 1 summarizes the decision problems we study. As shown in Table 1, each problem description includes the age of the advisee. While the specific age assigned to each scenario is ultimately arbitrary, we chose ages that broadly correspond to the typical life stage at which such decisions are faced. Our interest lies in variation in advice conditional on the age of the advisee, rather than in the importance of the age.

Domain	**Decision Problems**	Option A
Career	Enroll in the army at age 20	yes
	Get self-employed at age 35	yes
Family	Start a family at age 23	yes
	Get divorced at age 30	yes
Finances	Buy a house at age 24	yes
	Pay off more than the minimum for credit card at age 30	yes
Health	Take a drug with serious side effects at age 30	yes
	Select a high- or low-premium health insurance at age 35	high-premium plan
Education	Go to college or become certified IT specialist at age 18	college
	Switch major from engineering to English literature at age 20	yes
	Which educational institution to apply to	open-ended*
	What major to apply for	open-ended*

Note: The labeling of the options as Options A and B serves only exposition purposes. Furthermore, framing responses as yes/no answers was deliberately avoided to minimize potential default effects.

Table 1: Decision problems across different domains

While all of our advice concerned a choice between two options (a situation where the choice set was binary), open-ended advice may vary when offered to advisees of different socioeconomic backgrounds by providing different choice sets to choose from. For example, take two students with equal skills but from different socio-economic backgrounds. When it comes time to apply to college, their college advisors may offer them different sets of colleges to apply to, i.e., different choice sets (Zafar, 2013; Bleemer and Quincy, 2025; Leighton and Speer, 2026). Similarly, when AI users ask open-ended questions to LLMs, the advice may take the form of a choice set, which itself can also vary with people's individual socio-economic background. To evaluate this possibility, we also elicit two choice sets with respect to education decisions by asking our AI advisors to list a set of institutions to apply to and the major to select conditional on the prompted description of the advisee applying to college. The point here is that one way advice varies across different socioeconomic groups is that it provides different choice sets to choose from with the obvious effect that one cannot attend a first-rate college if it was never in one's choice set to begin with.

Offering AI advice can be tricky because the advice offered by the AI advisor should match the needs of the advisee. Hence, we cannot expect there to be some one-size-fits-all advice that is optimal for all people seeking it. In a perfect world, advice should be tailored to fit the needs of the advisee, but what those needs are is not always clear to the advisor (and sometimes neither to the advisee). Moreover, the perspective embodied in an LLM's advice may reflect the implicit viewpoint of a particular type of advisor.

To capture these considerations in our design, we systematically vary the information about the advisee and the advisor in our prompts. The description of the advisor is varied by employing a set of role-playing scenarios where we prompt ChatGPT to take on the persona of eight different types of advisor representing eight possible combinations of sociodemographic information based on gender (male/female), two races (Black/White), and income (low/high income). We also ran one scenario where ChatGPT was not given a persona at all, so we will refer to this one as the ChatGPT persona. Thus, including this ChatGPT profile, we have a total of nine different advisor profiles.

We did the same exercise for the advisee, where we have nine variations in advisee profiles: the same eight combinations of socio-demographic information and a ninth one without any demographic information. Hence, we constructed a 9×9 treatment variation in which nine different types of advisors offered advice to nine different types of advisees. For each advisee-advisor profile combination we ran 100 simulations, yielding a total of 8100 simulations for every single decision problem.[7] Each simulation can be considered as an independent prompt because the API platform does not save previous output as memory, unlike the user interface.

[7]For ChatGPT-5 the parameter controlled by "temperature" is preset to 1.

We illustrate our prompt design with Figure 1, which shows a screenshot of the API code. We incorporate information about the advisor's profile directly via role-playing in the API requests, but vary information about the advisee required a more subtle approach. Since AI systems, including ChatGPT, have content filters known for suppressing the influence of gender and race on its output, we proxied advisee characteristics using suggestive names. We selected the four names which ChatGPT inferred as a young individual who was either Black or White and male or female. In particular, we chose "Jack" for a White man, "Emily" for a White woman, "Jamal" for a Black man and "Laquisha" for a Black woman. We validated these name choices in a pre-test by asking the LLM to infer the socio-demographic background associated with each name. This feature is particularly important because users seeking AI-generated advice may not be aware that their conversations provide potential cues about their socio-demographic background and, hence, may not account for potential biases in the AI recommendations. In the no information benchmark, we omit any reference to socio-demographic background. We use the pronouns they/them/their for advisees and describe advisors simply as a "person".

The example in Figure 1 shows the prompt to the join-the-army scenario. In the figure, we have parsed the main components of the prompt into colored boxes to illustrate our prompt design. The first box highlights the advisor profile, where the text under "system_role" allows for customization of the persona the LLM should adopt. The second box contains the decision problem. Personal characteristics of the advisee are directly incorporated into the prompt, with the aforementioned names serving as demographic information.

The remaining information in Figure 1 specifies the response format and is fixed throughout the 10 decision scenarios. An important part of the response format are the justifications, that ChatGPT provides verbally and numerically. The numerical justifications correspond to ratings of several key factors that play a crucial role in most decision problems. Specifically, we asked the LLM to justify its choice by assigning a decision score between -5 and $+5$ to the dimensions safety (risk), long-term benefits, short-term benefits, regret, happiness, norm and ease of implementation. A high positive score means that the particular justification is a strong argument in favor of the recommended option, while a negative weight means that the justification works against the recommended option.

This is to recognize that advice may differ not only in the recommended option but also in the reasons for making a recommendation. Human decision makers might prefer one option over the other, because it is less risky, provides more benefit either in the long or short run, has lower potential for later regret, makes one happier, corresponds to what everybody else does, or simply because it is easier to implement. Hence, the justifications for advising one alternative over another may vary between decision problems, the type of advisor offering

```
# ---------- Config ----------
MODEL_NAME = os.environ.get("MODEL_NAME", "gpt-5")
TOTAL_SIMS = 100
N_PER_CALL_CAP = 8
CONCURRENCY = 4
TEMPERATURE = 1.0
# New file each run
OUT_CSV = f"PoorBlackMan_RichWhiteWoman_Army.csv"
print("Writing to:", OUT_CSV)


JUST_CSV = f"PoorBlackMan_RichWhiteWoman_Army_justifications.csv"
print("Also writing justifications to:", JUST_CSV)
SYSTEM_ROLE = "You are a trusted black man from a modest background who is being approached for advice."
explanation_free_prompt = """
Emily, a 20-year-old acquaintance of yours from a wealthy background, is considering joining the army.
She views military service as a viable path for her future. What advice would you give her?

- Give a confidence level on a scale from 0 to 10. In addition,
rate the importance of each of the following factors on a scale from -5 to 5,
where a negative value means this factor weighs against your recommendation,
while a positive value means this factor weighs in favor of your recommendation.
  1. Less Risk
  2. Long-term Benefits
  3. Short-term Benefits
  4. Typical Choice
  5. Less Regret
  6. Happiness
  7. Ease of Implementation

Format your response as follows (in JSON format):

{
  "choice": "Join Army" or "Don't Join Army",
  "confidence_level": (0-10),
  "factors": {
    "Less Risk": (-5 to 5),
    "Long-term Benefits": (-5 to 5),
    "Short-term Benefits": (-5 to 5),
    "Typical Choice": (-5 to 5),
    "Less Regret": (-5 to 5),
    "Happiness": (-5 to 5),
    "Ease of Implementation": (-5 to 5)
  }
}

Provide a brief explanation.
"""

FACTOR_KEYS = [
    "Less Risk",
    "Long-term Benefits",
    "Short-term Benefits",
    "Typical Choice",
    "Less Regret",
    "Happiness",
    "Ease of Implementation",
]
```

Figure 1: Example of the API Prompt

them, and the type of advisee receiving them. Furthermore, two advisors may offer identical advice for different reasons, and whether the advice is ultimately followed may depend on the justifications used to support it. While our set of justifications is not exhaustive, it captures the main decision criteria emphasized in the behavioral decision-making literature.

In addition, we also collect data on the LLM's confidence in the advice given (on a scale from 0 to 10).

3 Structure and Sources of Variation in AI Advice

We organize our results across two sections. In this section, we examine whether the LLM tailors its advice to socio-demographic information. In the next section, we explore to what extent AI advice is distinct from the behavioral patterns that we observe in different human strata.

Given a particular advisor, there are three main factors that may influence advice uptake: 1) the recommended course of action, 2) the justifications provided with the recommendation, and 3) the confidence with which the advisor conveys the advice. Below we organize our findings by examining each of these three channels separately.

3.1 The Advised Choice

Finding 1. For five of the ten scenarios we ask ChatGPT for advice on, it offers unanimous advice across all advisor/advisee pairs. For the remaining five scenarios where advice differs, most differences in recommendations are driven by advisee characteristics, specifically the advisee's socio-economic status but not its race or gender. Furthermore, the generic advice offered by ChatGPT when not prompted is most similar to the advice offered to low income advisees. In most simulations, the advice is highly resolved in that there is little variation in output across repetitions.

We find that the sensitivity of AI advice to demographic information depends on the scenarios. Essentially, our scenarios can be split into two sets: one set of scenarios where AI advice is identical regardless of the advisor/advisee profiles, and another set where AI advice varies drastically depending on the socio-economic information.

AI advice is unanimous for five scenarios, including *Family planning* (wait), *Paying down credit card debt* (Pay more and reduce debt), *taking weight loss drugs* (take the drug, with little variation), *Buying a house* (wait), and *Switching majors* (switch to English, with little variation). Further below we discuss how such unanimous advice may nevertheless affect

individuals differently, but first let's take a look at the decision problems that beget differential advice from AI.

For the second set of scenarios, we find a substantial variation in AI advice depending on the information available about the advisee and the advisor. These five scenarios include *Joining the army, Becoming self-employed, Buying insurance, Going to college* and *Getting a divorce*.

In the following, we use the term *differences* in advice in a purely descriptive sense, without assigning value judgments. The implications of these differences–whether advantageous or detrimental–depend on the specific evaluative criterion one has in mind. Yet, as argued above, specifying and measuring such a criterion is far from straightforward.

Figures 2(a) to 2(e) plot the average probability of recommending one specific option in each of the five decision problems with heterogenous advice. By average probability, we mean the fraction of times, in our 100 simulations, that ChatGPT recommended that option. Each line in these plots represents the recommendation given to a particular advisee, while every dot along a line corresponds to the recommendation provided by a specific advisor. Thus, holding an advisor constant, variation *between* the lines (in the vertical direction) indicates that different advisees receive on average different recommendations from the same advisor, reflecting advisee-based differences. Variation *along* the lines (in the horizontal direction) indicates differences in recommendations between advisor profiles, reflecting advisor-based differences.

The strongest socio-economic divide in advice can be found in the scenarios *going to college, starting a business* and *joining the army* (see Figures 2(a) to 2(c)). ChatGPT-5 always recommends that high-income individuals go to college, but low-income individuals pursue vocational training, without exception. Similarly, it recommends that low-income individuals join the army and stay employed (not start a new business) and high-income individuals not join the army and start their own business.

A more subtle type of advisee-based differences can be found in the *divorce* scenario (see Figure 2(d)), where female and White individuals are more likely to receive a recommendation to get a divorce than male and Black individuals.

Responses to the *insurance* scenario, depicted in Figure 2(e), also show advisor-based differences. Recommendations given when impersonating low-income White individuals are relatively stable, but they become much more sensitive to the advisee profile when impersonating high-income individuals.[8] For example, for low-income Black women, advisor-

[8]Interestingly, this scenario shows also that advice given without any information about the advisee may substantially differ from any advice given with information about the advisee, raising the question as what are the main factors influencing AI advice.

based differences appear minimal as ChatGPT most often advises taking the low-premium coverage (cf. purple line). In contrast, recommendations to wealthy White women show much greater variability (cf. pink line): While the perspective of a poor White woman advises predominantly against taking high-premium coverage, the perspective of a wealthy Black woman advises in favor of it (cf. variation within the pink line).

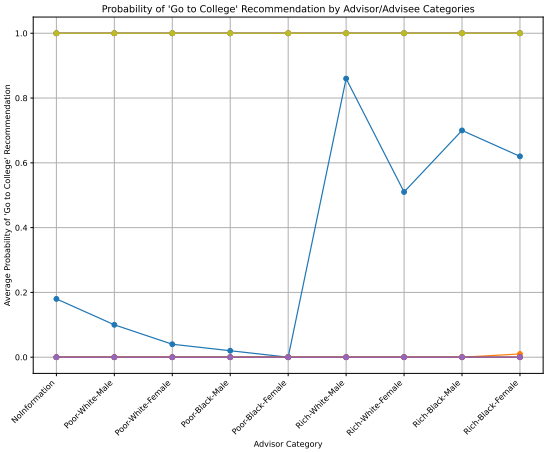
Overall, we find that most differences in recommendations are driven by advisee characteristics, particularly their socio-economic conditions. This is encouraging since variation grounded in the advisee's socioeconomic circumstances shows less bias than differences arising from the race or gender of the advisee or attributes of the advisor.

In general, considering all decision problems, we find the advice is highly resolved in that, conditional on a particular advisee-advisor profile combination, there is little variation across the 100 simulations. For each of the 81 advisee-advisor profile combinations, we compute the fraction of runs in which Option A versus Option B is recommended, and find that recommendations are highly consistent. In problems with unanimous advice, the LLM recommends the same option in nearly all 100 runs. Even in settings with heterogeneous advice across profiles, the model typically recommends the same option almost all the time to a given advisee (see Appendix Figure D.7). The main exceptions are the *divorce* and *insurance* scenarios, where recommendations exhibit greater variability across repeated simulations. For each advisee-advisor combination, the median resolution across all scenarios, as measured by the variance in recommended choices, is 0 (perfect resolution), with a mean of 0.048.

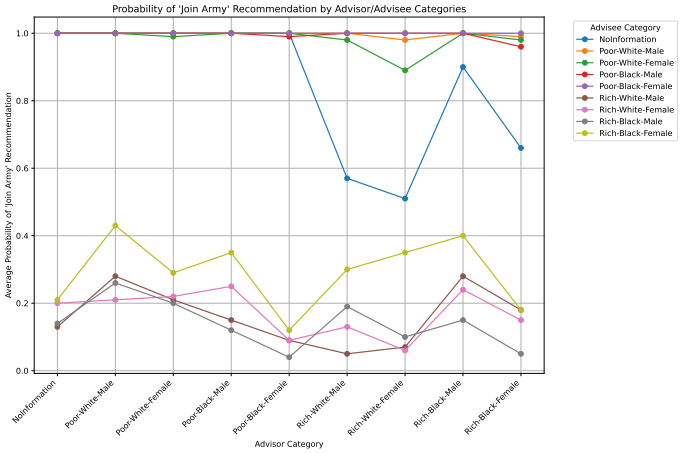
Variance Decomposition Figure 3 shows a variance decomposition in choice recommendations for the five scenarios with heterogeneous advice. The chart displays how the total variance is partitioned across three sources: variance that can be explained by variation in the advisor profile, the advisee profile, and the residual variance. In the *army*, *career*, and *education* scenarios, the advisee profiles explains most the of variance in recommendations. In contrast, for the *divorce* and *insurance* scenarios, neither advisee nor advisor characteristics seem to play a determining role in driving differences in advice. In such decision problems, the variability is due to unobserved factors.

Overall, the decomposition highlights a sharp divide between scenarios where advisee identity is highly predictive (*army, career, education*) and those where outcomes are largely unpredictable using the modeled factors (*insurance, divorce*).

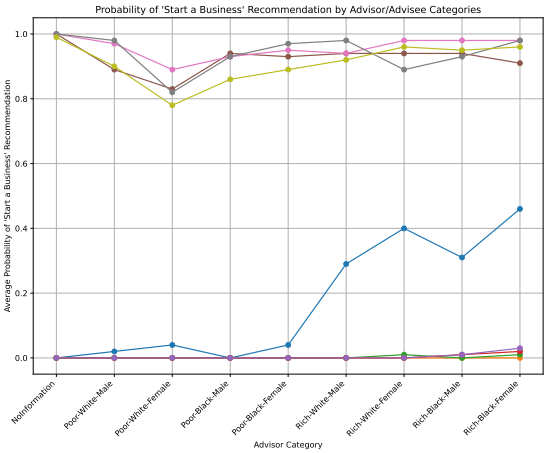
Who is ChatGPT, and who is it speaking to? When people ask ChatGPT for advice, they may rarely ask ChatGPT to impersonate a particular advisor. This raises the question of whether ChatGPT's advice, when unprompted, is skewed toward a particular type of



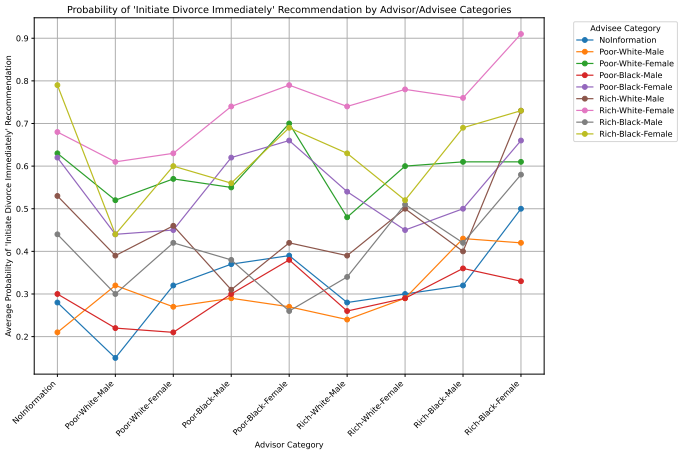
(a) Go to College



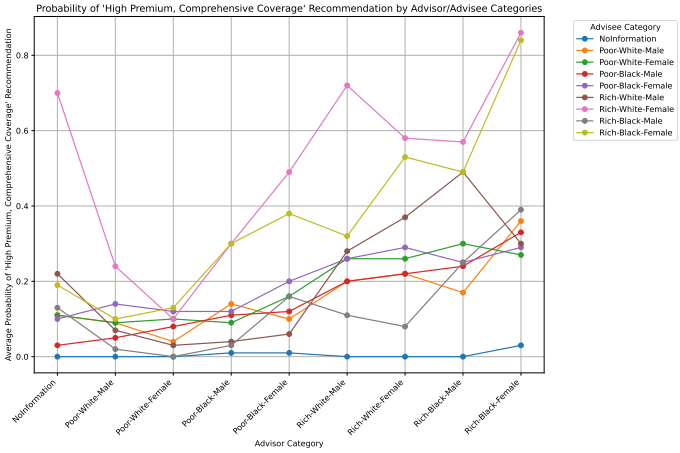
(b) Join the Army



(c) Become Self-Employed



(d) Divorce



(e) Get High Premium Coverage

Figure 2: Recommendations by Advisor and Advisee

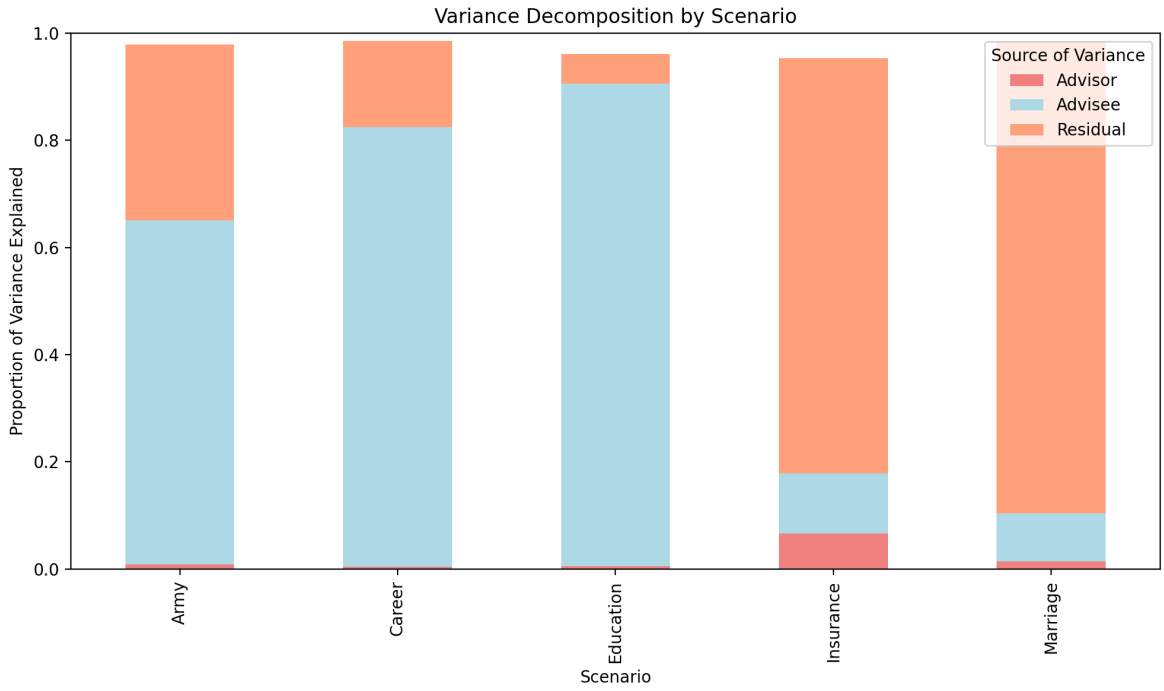


Figure 3: Variance decomposition

advisor. For instance, does it adopt the perspective of a wealthy White man or a low-income Black woman when providing guidance? Similarly, to whom does it speak when providing guidance without particular socio-demographic cues? In the absence of socio-demographic information about the advisee, does the model's general advice more closely align with the advice it gives to a low-income White man or to a high-income Black woman?

To answer these questions we compare the responses of ChatGPT in the generic prompts (i.e., without socio-demographic information) to the ones given when ChatGPT is given a particular advisee or advisor profile. As a measure of the distance between the distributions of responses of two profiles, we compute the Jensen-Shannon divergence.[9] Concretely, for each advisee–advisor profile combination, we construct the Bernoulli distribution of recommendations (Option A versus Option B) across the 100 runs and compute the Jensen–Shannon divergence between this distribution and the corresponding distribution obtained under the no-information baseline simulation.

[9]The Jensen-Shannon divergence is based on the Kullback-Leibler divergence but has the advantage of being well-behaved, symmetric in the comparison and between 0 and 1. Let $P(j)^i$ and $P(j)^0$ be the probability with which option $j = \{a, b\}$ is recommended by the demographic advisor profile i and the ChatGPT advisor profile 0, respectively. The Jensen-Shannon divergence is computed as follows:

$$JS(P^i||P^0) = \frac{1}{2}\Big[\sum_{j\in a,b} P(j)^i \log \frac{P(j)^i}{\frac{1}{2}(P(j)^i + P(j)^0)} + \sum_{j\in a,b} P(j)^0 \log \frac{P(j)^0}{\frac{1}{2}(P(j)^i + P(j)^0)}\Big] \quad (1)$$

Advisor Profiles	Divorce	Career	Education	Insurance	Army
Poor/Black/Female	0.007	0.015	0.011	0.017	0.009
Poor/Black/Male	0.002	0.017	0.006	0.027	0.012
Poor/White/Female	0.003	0.04	0.004	0.051	0.008
Poor/White/Male	0.009	0.014	0.001	0.027	0.022
Rich/Black/Female	0.028	0.043	0.017	0.079	0.021
Rich/Black/Male	0.012	0.03	0.023	0.035	0.009
Rich/White/Female	0.048	0.039	0.01	0.031	0.011
Rich/White/Male	0.033	0.029	0.041	0.016	0.008
Wgtd avg. profile	0.014	0.023	0.003	0.013	0.006
Advisee Profiles					
Poor/Black/Female	0.09	0.09	0.233	0.096	0.043
Poor/Black/Male	0.082	0.092	0.233	0.072	0.005
Poor/White/Female	0.06	0.094	0.233	0.088	0.057
Poor/White/Male	0.081	0.101	0.229	0.074	0.008
Rich/Black/Female	0.345	0.521	0.552	0.217	0.077
Rich/Black/Male	0.5	0.582	0.552	0.06	0.013
Rich/White/Female	0.449	0.602	0.552	0.328	0.136
Rich/White/Male	0.468	0.544	0.552	0.109	0.023
Wgtd avg. profile	0.132	0.130	0.118	0.127	0.026

Note: Minima within each column (advisee/advisor) are highlighted in yellow per column. In the bottom panel (advisee profiles), we also highlight values that are not sufficiently distinct from the column minimum; in the top panel (advisor profiles), this holds for almost all columns, further underscoring the limited role of the advisor profile in generating differences. In addition, the divergence measures are also highlighted when the weighted average profile (an average of the eight profiles that is more representative of the US population) yields the minimum across rows.

TABLE 2: JENSEN-SHANNON DIVERGENCE BY SCENARIO

Table 2 organizes the Jensen–Shannon divergence (JSD) measures into two panels. The top panel reports the divergence in advice generated with and without information about a specific advisor profile, while the bottom panel reports the divergence in advice generated with and without information about a specific advisee profile. The highlights in Table 2 indicate

the profiles with similarly small divergences, while the darker yellow highlight indicates the smallest value in each column (e.g., i.e. the type of person the unpromted AI is closest to). We also explore whether ChatGPT's generic advice (the one without socio-demographic cues) resembles a representative profile.[10]

In the top panel (advisor profiles), the divergence measure allows us to assess which advisor profile ChatGPT implicitly reflects when it is not prompted to impersonate a particular advisor. In most cases, the divergence measures are below 0.05, indicating minimal differences across advisor profiles. This result reflects the fact that there is little advisor-based differences in the data. We therefore conclude that, for the scenarios we study, none of the differences between the advisor profiles turn out to be sufficiently important.

In the bottom panel (advisee profiles), the measures shed light on which advisee ChatGPT implicitly targets when no socio-demographic information is provided in the prompt. In most scenarios, ChatGPT's generic advice is closest to the advice it gives to low-income individuals.

3.2 Confidence

Finding 2. When prompted to assess its own confidence, the LLM expresses only moderate confidence in its recommendations, indicating that the highly resolved recommendations need not reflect low uncertainty in the LLM's underlying distribution over possible responses.

Across all simulations, ChatGPT's confidence in its recommendation is moderate and ranges between 4 and 9 with a median confidence level of 6. Thus, although the advised choice is often highly resolved, the confidence index correlates only weakly with resolution (see Appendix Section A). This reflects a structural feature of LLMs: while they internally assign probability mass across many possible outputs via softmax-based token distributions, the advice request prompt forces a single point prediction (in simple terms, the model must "pick one answer" even when several options are similarly plausible). Such deterministic decoding can therefore produce highly stable recommendations across repetitions, even when the underlying probability distribution is relatively flat. Consequently, resolution captures output stability rather than the absence of uncertainty in the model. Thus, users who are

[10]The weighted average profile is computed using the following weights: Poor White Man 17%, Poor White Woman 12%, Poor Black Man 10%, Poor Black Woman 11%, Rich White Man 21%, Rich White Woman 18%, Rich Black Man 5%, Rich Black Woman 6%. This is meant to reflect the rough breakdown of the US population, and could be interpreted as an advisor or advisee representative of the US population. Of course, the US population consists of more demographic groups, thus, our findings for this row are to be interpreted as a first best approximation given our data limitations.

unfamiliar with the internal workings of LLMs should be cautious not to interpret highly resolved recommendations across repeated prompts as highly confident ones.

In Appendix Section A we investigate the correlation of confidence with observables in our data such as the resolution in recommendations, socio-demographic profiles, and the trade-offs in justifications. We find that, while many of them have significant effects, these effects have little explanatory value overall.

3.3 Justifications

Finding 3. The LLM's justifications for the recommended choice are problem-specific and sensitive to socio-demographic information. For example, for high-income advisees, ChatGPT places more emphasis on long-term benefits, regret, happiness when justifying its advice while it focuses on short-term benefits, risk and the ease of implementation when justifying its advice for low-income advisees. Furthermore, when asked to interpret its own verbal justifications, ChatGPT views these explanations as conveying more extreme assessments and stronger confidence in its recommendations than its corresponding numerical justification and confidence scores suggest.

What may be particularly persuasive to advisees are the arguments used to justify the recommended choices. Thus, beyond potentially different recommendations for different advisor/advisee profiles, we are also interested in why certain advisors give a particular advice to certain advisees. This applies both to the case where there is variation in recommendations and when there is not. When there is, we would like to know which justifications are driving the difference, and whether any particular justifications stand out. Even when there is not, we would like to know if different groups are being recommended the same action based on different motivations, as this could differentially affect the user's willingness to follow advice.

Numerical Justifications We ask ChatGPT to indicate the strength with which it offers each justification on a scale from -5 to +5. Scores close to zero indicate weak support for a given justification. Hence, many numbers near zero indicate that it did not support the justifications it uses very strongly, and then the sum of the absolute value of its justification strength would not be very high.

We find substantial variation in the total number of absolute scores that ChatGPT assigns to justification factors. The total number of absolute scores across the seven dimensions varies from 7 to 26 and varies systematically with advisor and advisee characteristics (see Online Appendix Table 2). The median number of justifications used (i.e., non-zero values) is 7, but

varies across simulations between 4 and 7. Conditional on the choices, more justifications are used when the LLM impersonates a Black, low-income and a female advisor, and it tends to use fewer justifications for female advisees (see Online Appendix Table 2).

The LLM also uses negative numbers, most frequently for the justifications *typical choice* (38% of total simulations), *less risk* (34.6%) and *short-term benefit*s (34.1%), meaning that it also provides counter-arguments to the recommended choice. Controlling for the same advice, the use of negative numbers also differs with the advisee and advisor profiles (see Online Appendix Table 3). For example, the LLM uses more negative numbers for low-income advisees (for most dimensions of justification except for *typical choice*), for female advisees (in particular for *happiness*, *less risk* and *typical choice*). In addition, the LLM is also more likely to use negative numbers when impersonating low-income advisors (in particular, for *less risk* and *ease of implementation*).

Overall, the observed variation in justification scores across advisee profiles indicates that the LLM conditions its evaluations on advisee attributes, assigning different scores to different justifications accordingly. The variation across advisor profiles may reflect experience effects, although this remains difficult to establish conclusively.

While informative, these raw justification scores discussed above conflate two sources of variation. A justification may be of large magnitude when the associated option exhibits strong consequences along the relevant dimension. For example, a choice like starting a new business is particularly risky and hence we would expect that such riskiness would be mentioned when justifying a choice to start one or not. Alternatively, it may be that that a particular justification is simply a lot more relevant compared to others (e.g., the LLM determined that risk considerations matter more than happiness for that option). To unravel these two explanations, we also examine justification weights, defined as the ratio of the absolute score of a justification divided by the sum of absolute scores across all seven justifications given to a particular advice. Intuitively, this normalized measure captures the relative importance of each justification compared to the others. This normalized measure allows us to draw comparisons across choices and scenarios (in contrast to the raw scores, which can only be interpreted conditional on a specific choice). It also enables cleaner comparisons between simulations, since the total number of absolute scores allocated to justifications varies between them.[11]

Figure 4 plots the median justification weights, aggregated across all scenarios and all

[11]We conduct a nested comparison of model fit to assess whether the justifications weights provide additional explanatory power beyond the raw justification scores, and we find that they do. Conditional on the justifications weights, the raw justification scores explain 31.12% of the variation in advice ($p < 1\%$ in Wald test). Conditional on the raw justification scores, the justifications weights explain 8.15% of the variation in advice ($p < 1\%$ in Wald test). All the model regressions included fixed effects for the scenarios.

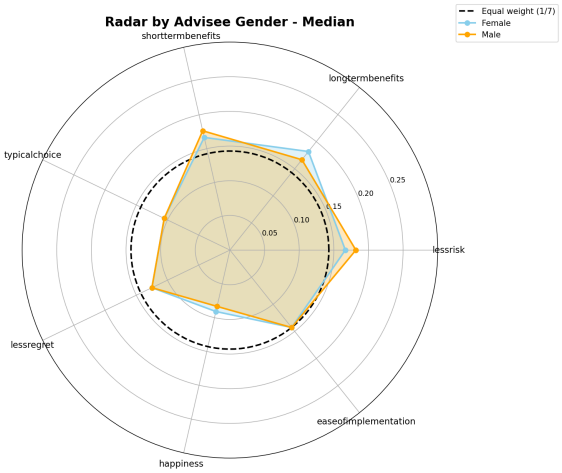
simulations, for the seven justification dimensions (i.e., *less risk, longterm benefits, etc.*). The black dashed circle indicates the benchmark in which all justifications receive equal weight. Here, we focus on how justification weights vary with the advisee characteristics, specifically the advisee's gender, race and income (see Online Appendix Figure B.3 for more subtle results by the advisor profile). That is, in each plot, the orange and blue graphs correspond to the two groups being compared (e.g., high- versus low-income advisees). When justification weights are similar across groups, the corresponding graphs largely overlap. This is the case for gender and race. Figures 4(a) and 4(b) show no substantial differences in justification weights by gender or race: across all scenarios, ChatGPT assigns similar weights to the seven justifications for male and female advisees, as well as for White and Black advisees.

In contrast, Figure 4(c) reveals pronounced differences by income. For high-income advisees, ChatGPT places more emphasis on *long-term benefits, regret, happiness* and *typical choice* compared to low-income advisees. In contrast, *short-term benefits, risk* and the *ease of implementation* are relatively more important justifications for low-income advisees. The biggest difference between these two groups can be observed for long- and short-term benefits.

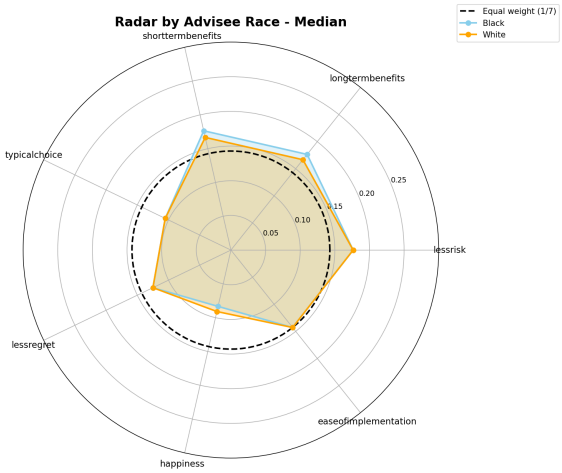
It is curious to note that the justifications used to justify the advice across different socioeconomic groups seem to reflect the different norms commonly attributed to these groups in real life. For example, the postponement of gratification (looking at long-term consequences) has been considered the hallmark of upper middle class decision making, while anticipated regret may also resonate with people in such strata. To the extent that happiness may be considered a luxury that only upper income individuals might consider, it may not be surprising that it is used to justify upper income decisions. The fact that advice to individuals with lower income is justified by considering risk, short-term benefits, and ease of implementation may represent the tighter constraints under which these individuals live where even small losses can cause considerable damage and where easy to-implement plans are valued.

In Figure 4, we aggregate justification weights across all scenarios. While informative, this aggregation masks substantial heterogeneity across decision contexts. When examining individual scenarios, we observe considerable variation in the relative importance assigned to different justifications, indicating that ChatGPT adapts its justifications to the specific environment and available options.[12]

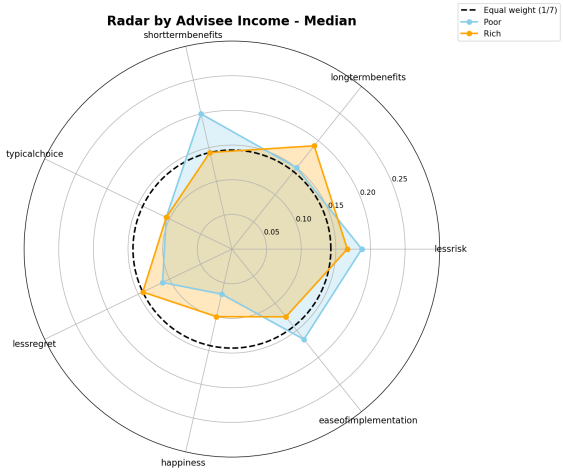
[12]In the aggregate, once we condition on choice fixed effects, we still find systematic differences in the justifications for the same option given to low- and high-income advisees. Although many differences attenuate, low- and high-income advisees continue to diverge along the dimensions *less risk* and *short-term benefits* (see Online Appendix Table 4). This comparison also indicates that a substantial share of the overall demographic differences in justification weights is driven by differences in recommended choices. Put differently, different advisees are steered toward different options, and those options naturally emphasize different dimensions of the decision problem.



(a) By advisee's gender

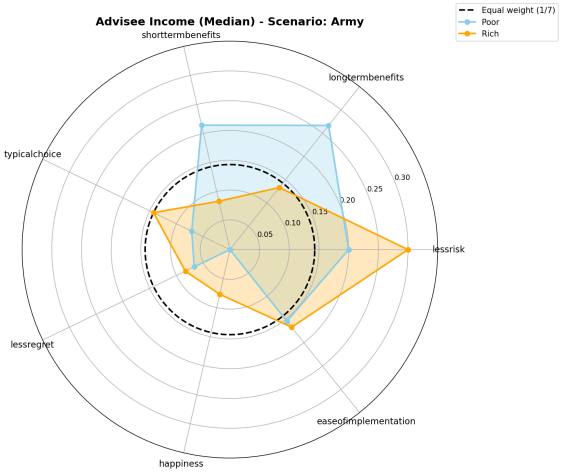


(b) By advisee's race

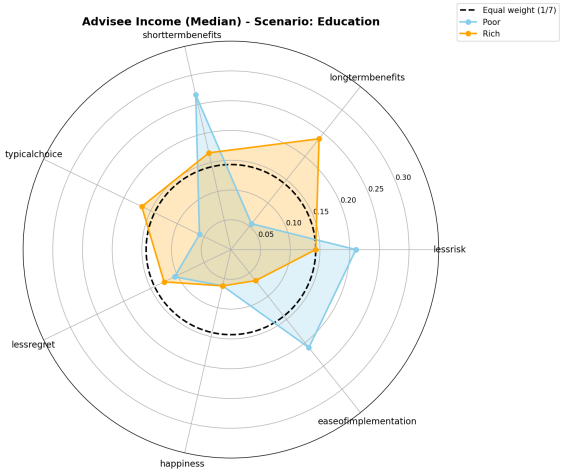


(c) By advisee's income

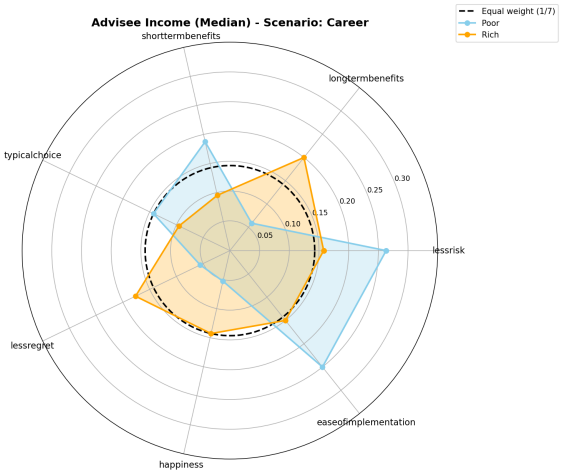
Figure 4: Justification Weights Across Scenarios



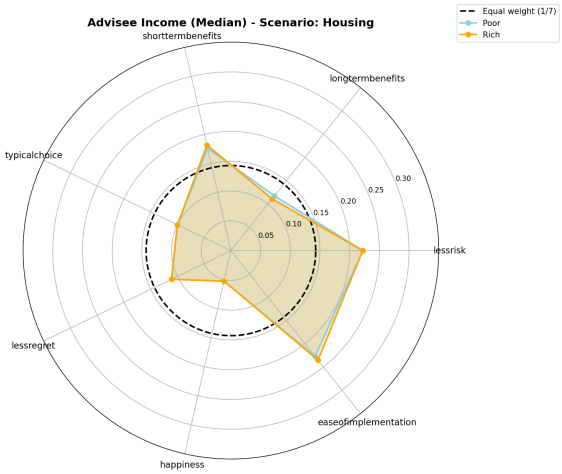
(a) Army



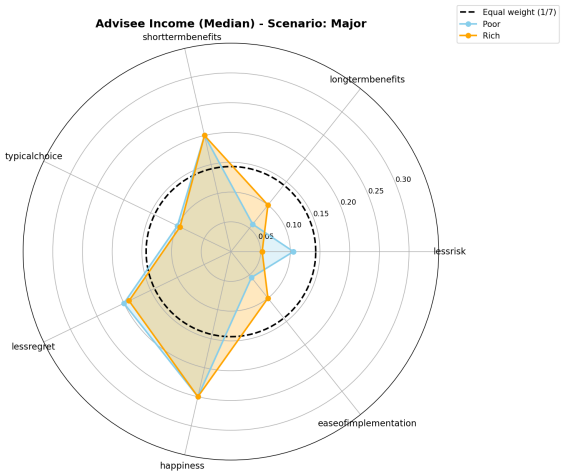
(b) Education



(c) Career



(d) Housing



(e) Major

Figure 5: Median justification weights in selected scenarios - by income

Figure 5 illustrates the allocation of justification weights for low- and high-income advisees across a subset of scenarios (the complete set is reported in Online Appendix Figures B.4 and B.5). A first key observation is that the dominant justifications are context-dependent. For example, in the housing scenario, the LLM places greater weight on *ease of implementation*, *less risk*, and *short-term benefits*. In contrast, in the major-choice scenario, the most salient dimensions are *happiness*, *less regret*, and *short-term benefits*.

These patterns imply that the income-related differences documented above should not be interpreted as mechanically driven by a fixed emphasis on particular justifications (such as long-term benefits for high-income advisees). Rather, the relative importance of justifications varies systematically with the decision problem and the recommended option. Nevertheless, when averaging across the full set of scenarios, clear and systematic differences in justification weights by income remain.

A second observation is that, especially in settings characterized by heterogeneous advice, between-group differences in justification weights can be substantial, as illustrated in the *Army*, *Education*, and *Career* scenarios. Consequently, averaging across scenarios provides a conservative estimate of the magnitude of income-based differences in justification patterns.

Verbal Justifications When users actually turn to AI for life advice, they typically do not ask for numerical justifications on various dimensions, but will almost certainly get an explanation as to why a certain action should be chosen. Hence, in addition to the numerical justifications that we elicit directly from the LLM, we are also interested in the nature of the verbal justifications that explain its reasoning behind each recommendation.

To better understand whether the verbal justifications align with the recommended choices, and if so, how consistent the verbal language is with the numerical justification scores provided, we collect all justifications across the simulations for different advisees, and feed the text of the verbal justifications back into the LLM to get an overall recommendation, confidence level, and justification scores based *solely* on the text. We refer to these as the LLM's *interpreted* choice recommendation, confidence and justification scores, which we then compare to the modal choice recommendation, the mean confidence, and the mean justification scores. Overall, we find that the verbal justifications align very closely with the modal choice recommendation. Interestingly, according to the LLM, the verbal justifications convey a higher confidence in its language compared to the confidence levels it directly expresses. Furthermore, the interpreted justification scores underlying the verbal justifications tend to be significantly higher (in absolute value) than the numerical scores that the LLM provided itself. All together, this suggests that the LLM tends to provide verbal justifications that views the situation in a more clear-cut way than the numerical data would suggest (see Online

Appendix B.3 for more details).

3.4 Advice in the Form of Choice Sets

Finding 4. When advice takes the form of choice sets, ChatGPT offers very different recommendations for students with high and low socioeconomic status and for men and women. In the education problems we study, the variation in recommended choice sets across our advisee profiles mirrors documented differences in educational choices by gender and socioeconomic background.

In all the scenarios above, we presented ChatGPT with a binary choice. However, one of the main ways in which advice to low-income individuals may differ from that of high-income individuals (or to women as opposed to men, etc.) is by suggesting different choice sets when asked open-ended questions. For example, take two equally qualified students who make appointments to see their college advisors, one who attends a wealthy suburban school, and the other a poor inner-city school. At the end of their meetings, each will be presented with a set of schools that their advisor suggests would be appropriate for them to apply to. Although they are equally qualified, if these recommendations are followed, these two students are likely to end up in very different schools despite their equivalent qualifications. Hence, when questions are open-ended, advice in the form of choice sets may be crucial.

To investigate this issue, we ask ChatGPT to provide recommendations to students on where to apply to college and for what major depending on their socioeconomic background. Specifically, for each advisor/advisee profile along the dimensions of race, gender, and socioeconomic status, we asked ChatGPT to provide up to 5 institutions and 3 majors in each instance, and again simulated each profile 100 times, together with a no information baseline, this gives us $(8+1) \times (8+1) = 81$ different combinations, for a total of 8100 simulations.

Overall, we find systematic differences in the colleges that students are recommended to attend and the majors they are recommended to pursue. For the institutions, we find that high-income advisees were recommended to apply to more prestigious and more research oriented universities overall compared to low-income advisees. For college majors, we find that men were recommended to study more technically demanding subjects than women, and high-income advisees were recommended to study subjects in the social sciences more often than low-income advisees.

Focusing on the advisee's income profile, we are left with 3600 simulations for low- and high-income advisees, respectively. We plot the colleges' ranking by their frequency after aggregating them along the advisee's income profile using two separate rankings, the 2026 editions of the U.S News & World Report for liberal arts colleges and of the QS

World University rankings. We consider two rankings because general university rankings typically exclude most liberal arts colleges, while liberal arts college rankings do not include international institutions. The top three choices given to low-income advisees were Amherst College, Princeton University and Rice University (see Appendix Table 8), which are all institutions that offer need-blind financial assistance to domestic applicants, suggesting that ChatGPT accounts for affordability and cost of attendance. For the high-income advisees, the recommended set of universities are more heavily concentrated on institutions ranked in the top 50 worldwide, whereas for low-income advisees, there is more dispersion, with a smattering of institutions ranked outside the top 100 (see Figure 6). Figure 6, based on QS World University Rankings, shows the frequency with which an institution of a given rank was recommended, separately for low-and high-income advisees. Likewise, Online Appendix Figure C.6 shows the distribution of rankings by the advisee's income but is based on the U.S. News & World Report ranking of liberal arts colleges. An immediate feature that stands out is that essentially only low-income advisees were recommended to attend liberal arts colleges.

For college majors, we likewise filter advisees separately based on income and gender. Online Appendix Table 7 displays the major choices broken down by gender and income, keeping only majors that were mentioned at least 40 times for at least one group. Several interesting features stand out: First, Computer Science and Economics are universally popular across all advisor/advisee characteristics. In fact, both are recommended by every advisor profile to every advisee profile in almost every single simulation. Second, along the dimension of income, there is a strong tendency for advisors to recommend mathematics, statistics, and data science to low-income advisees and to recommend political science and applied mathematics to high-income advisees. Third, along the gender dimension, there is a strong tendency for advisors to recommend political science, biology, and psychology to women and to recommend mathematics, applied mathematics, and electrical engineering to men. Evidence indicate that these patterns largely coincide with documented choices, suggesting that using AI for advice may enforce existing patterns of major choice among different groups. Leighton and Speer (2026) for example, examines the relationship between family background and major choice in college. They note that holding education constant, income has only a weak influence on major choice. However, they find that students whose parents are more educated are significantly more likely to choose majors with low early-career earnings but much faster earnings growth. They are also less likely to choose "safe" majors that provide short-term employment guarantees. In terms of gender differences, Zafar (2013) report that in $1999-2000$, among recipients of bachelor's degrees in the United States, only 2% of women majored in engineering, whereas 12% of men did so. This difference in major choice is, to a large extent, responsible for a "earnings potential" gap, which is further amplified later on

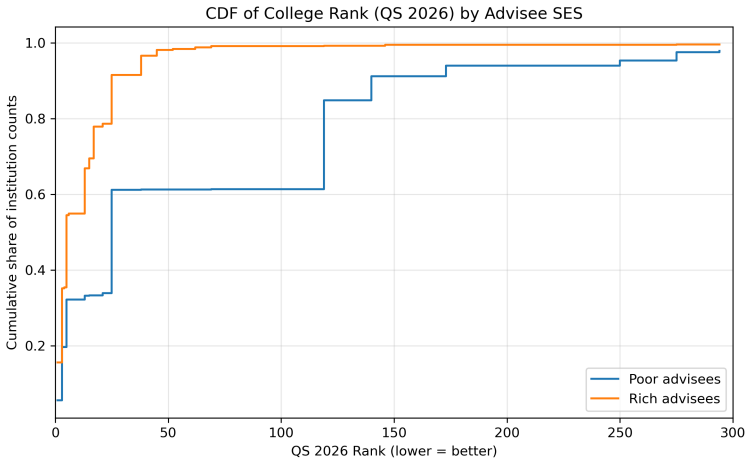


Figure 6: Distribution of QS World University Rankings of Recommended Institutions by Advisee's Income

via occupational sorting (Sloane et al., 2021).

4 Comparing AI Advice with Human choices

Finding 5. AI advice tends to be confirmatory for low-income advisees and disconfirmatory for high-income advisees.

Evaluating AI advice for life's Big Decisions is difficult because there is no objective benchmark for advice quality. Without knowing the optimal recommendation for each individual advisee and given that identifying such optima is itself challenging, we cannot determine whether AI provides better or worse advice than humans. In light of this, a central question becomes whether AI has the potential to reinforce or alter existing human behavior. Hence, in this section, we examine how AI advice differs from actual behavior that prevail in human networks.

To answer this question, we take advantage of the American Community Survey (ACS), which provides information on individuals' observed behaviors and socioeconomic outcomes in the U.S., based on a large, nationally representative sample of households. We compute the vector of choice probabilities provided by the LLM for each of the eight advisee profiles and compare them to existing outcome probabilities in similar contexts in the ACS. In other words, we compare what the LLM recommends in our decision problems to people's actual behavior in the U.S. Some of the information underlying our scenarios is not available in the

Scenarios	RWM	RWF	RBM	RBF	PWM	PWF	PBM	PBF
Buy house	0	0	0	0	0	0	0	0
(% home ownership)	(0.83)	(0.88)	(0.84)	(0.90)	(0.26)	(0.22)	(0.19)	(0.12)
Start business	0.92	0.96	0.94	0.91	0	0	0	0
(% self-employed)	(0.14)	(0.09)	(0.15)	(0.09)	(0.22)	(0.14)	(0.15)	(0.09)
Join army	0.16	0.17	0.14	0.29	1	0.99	1	1
(% army occup.)	(0.06)	(0.01)	(0.06)	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)
Go to college	1	1	1	1	0	0	0	0
(% college degree (22 yrs)	(0.50)	(0.67)	(0.15)	(0.40)	(0.22)	(0.26)	(0.13)	(0.14)
Have a baby	0	0	0	0	0	0	0	0
(% parents)	(0.01)	(0.03)	(0.00)	(0.04)	(0.10)	(0.20)	(0.08)	(0.30)
Divorce	0.46	0.74	0.40	0.63	0.30	0.58	0.29	0.55
(% divorced)	(0.08)	(0.05)	(0.39)	(0.33)	(0.16)	(0.25)	(0.15)	(0.22)

Choices = Profiles

Legend for group categories: R:Rich, W:White, M:Male, P:Poor, B:Black, F:Female

Figure 7: AI Choice Probabilities and Human Outcome Frequencies By Advisee Profile

ACS, such as data on major choice or drug usage.[13,14]

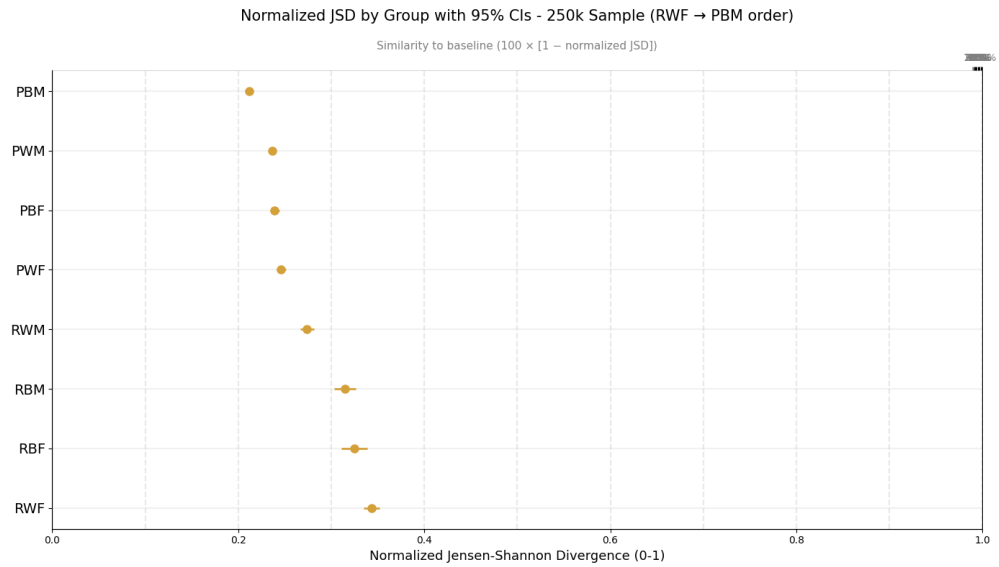
Our results are presented in Figure 7 for selected scenarios (see Online Appendix Figure D.7 for the full set of AI-advised choices). In this matrix, for six of the decision problems we used in our ChatGPT exercise, we record the fraction of the 100 simulations we ran for each of these problems problems that recommend a particular choice and compare it to the fraction of the same outcomes observed in the ACS survey sample.[15]

To make this comparison we compute the Jensen-Shannon Divergence to measure the

[13]Although the ACS measures do not perfectly align with our experimental scenarios, they nonetheless capture the dominant choices prevalent across different social strata. We therefore compute choice probabilities in the closest available contexts within the ACS and interpret large discrepancies between these probabilities and those implied by AI-generated advice as indicative of contexts in which the AI has the potential to offer distinct decision recommendations.

[14]In computing the choice probabilities with the ACS data, we respect the age restrictions given in the scenarios. For instance, because our housing scenario prompt the LLM to consider young adults at age 24, in the ACS we look at home-ownership rate for adults aged 24. The only exception is the college scenario, where the prompt asks the LLM to consider an 18-year old student prior to the education choice, and where in the ACS data we compute the choice probabilities with 22-year old respondents who had a least the chance to start some educational degree.

[15]The matrix in Figure 7 present outcome probabilities in the ACS for low-income respondents earning less than $40 000 in household income per year and high-income respondents earning more than $250 000 in annual household income. We replicate the results (with more extreme divergence results) when we lower the high-income threshold to $100 000 household income per year.



Note: Standard errors for the 95% confidence interval are computed using the Delta method.

Figure 8: Jensen-Shannon Divergence between AI Advice and Human Choices

distance between AI recommendation probabilities and actual outcome frequencies for each the eight advisee profiles. Figure 8 shows that divergence measures range from 0.21 (for poor, Black, male (PBM) profiles) to 0.34 (for rich, White, female (RWF) profiles). These values indicate roughly a 80% to 90% similarity between the AI's recommendations and observed outcomes, suggesting that AI advice is broadly aligned with, yet not entirely redundant of, existing behavioral patterns.[16] Interestingly, the highest divergences are found among upper-income advisees. This pattern implies that, to the extent that AI advice reflects perspectives beyond one's immediate social network, it may offer less "out-of-network" value to lower-income individuals.

Figure 9 compares AI advice and human outcome probabilities across scenarios, aggregating those shown in Figure 7. Crosses indicate advice probabilities, while dots represent observed outcome frequencies. The figure groups observations by the demographic dimension along which AI advice varies most strongly. For most scenarios, this dimension is income: points on the left correspond to high-income profiles, and points on the right to low-income profiles. The main exception is the Divorce scenario, where heterogeneity arises by gender; in that case, points on the left correspond to female profiles and those on the right to male profiles.

[16]While the Jensen–Shannon divergence (JSD) is bounded between 0 and 1 (using log base 2), its values are not directly interpretable as percentages of dissimilarity. For interpretive clarity, we report approximate "similarity" measures using the transformation $S = 2^{-\mathrm{JSD}}$, which expresses the fraction of shared information between two distributions (analogous to an exponential decay in informational distance). Thus, for example, a value of JSD $= 0.18$ implies $S \approx 0.88$, meaning that the two distributions share roughly 80–90% of their informational content.

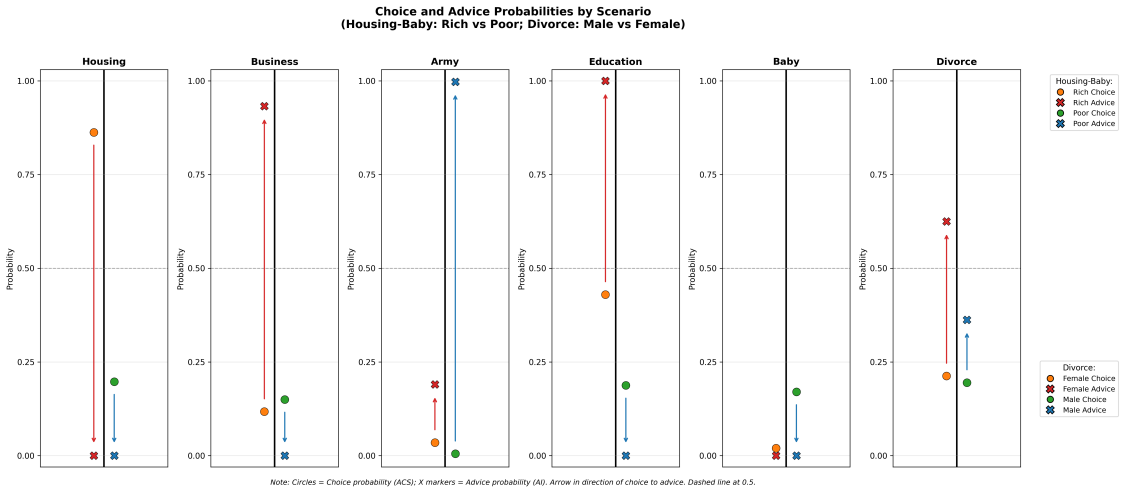


Figure 9: Choice vs. Advice by Scenario

There are a number of features in this figure that are worth noting. First, as already discussed in Section 3, AI advice is well resolved in the sense that it often offers almost unanimous advice about important decisions. Put differently, almost all simulations run on a given problem offer homogeneous advice about what to do. As shown in Figure 9, advice probabilities cluster near the extrema for nearly all scenarios, with the Divorce scenario being the notable exception. In particular, AI advice tends to be more homogeneous than actual human behavior. For example, it tells every advisee, irrespective of their income, to wait before buying a house at age 24 and to wait before having a baby at age 23.

Second, Figure 9 highlights when AI advice reinforces existing choice patterns versus when it disconfirms them. In some cases, the advice aligns with individuals' actual behavior within a given group; in others, it points in the opposite direction. We classify advice as confirming when advice and outcome probabilities lie on the same side of the 0.5 threshold, i.e., when both are below 0.5 or both exceed 0.5. Conversely, we interpret AI advice as disconfirming when advice and outcome probabilities fall on opposite sides of this threshold. The arrows in the figure emphasize cases in which AI advice has the potential to pull choices toward the opposite extreme.

We see that for low-income individuals, AI advice is mostly confirming, with the exception of the Army scenario, where it gives low-income individuals the strong recommendation of choosing Army as a career path. In fact, this is the biggest gap between AI advice and actual behavior in our data. In the other four scenarios, AI advice is reinforcing what low-income individuals already tend to do.

For high-income individuals, by contrast, the advice is confirming with respect to their Army and Baby choices, but otherwise frequently pushes them away from their actual behavior.

Hence, AI advice tends to be confirmatory for low-income advisees and disconfirmatory for high-income advisees.

This distinction matters even in scenarios where AI advice is homogeneous across advisee profiles. In five of our ten decision problems, ChatGPT offers the same recommendation regardless of the advisee's socio-demographic profile. Yet, identical advice can still play very different roles if for some groups it reinforces what they already do, while for others it points in the opposite direction of their observed behavior. This is, for example, apparent in Figure 9 for the housing scenario, where AI advice always recommends to wait before buying a house at age 24. While this advice is confirming for low-income individuals, this is the opposite of what wealthy individuals actually do at age 24.

The benefit of such AI advice depends on how it is processed by human advisees. Since our data do not include people, we cannot empirically say how such AI advice influences human choices. However, given the homophilous nature of human networks, hearing advice that deviates from the actions typically taken by members of one's network (i.e., differs from the norm) is likely to be treated differently from advice that supports what people are already doing. Such advice may lead a human decision maker to rethink the actions that are typically taken by members of their networks. Our selection of problems suggests that AI-generated advice has the potential to reinforce existing behavior of low-income individuals, while it is more likely to prompt high-income individuals to reevaluate their decisions–regardless of whether or not they follow this AI advice.

5 Conclusion

This paper has investigated Big Decisions, i.e., those decisions in our lives that are transformative and change who we are as people. On the face of it, it might seem mistaken to ask a large language model (LLM) for advice on such important decisions. Yet emerging evidence suggests that younger generations increasingly turn to AI as a sounding board for their reasoning about both life's Small and Big Decisions. Moreover, because individuals are typically sorted into socially homogeneous networks, where the advice they receive simply reinforces the decisions that people like them are already making, there may be a role for AI to offer "unusual" advice for such problems. By offering advice that deviates from one's immediate social environment, it has the potential to offer unconventional perspectives and prompt individuals to reconsider the course they are about to take.

To do this, we conducted a simple, yet informative exercise to assess whether the advice generated by a leading LLM adds value by providing perspectives that differ from the

behavioral outcomes prevailing in individuals' social networks. Our goal is not to judge whether AI gives "good" advice (a criterion that is hard to assess for Big Decisions), but rather to examine the nature and heterogeneity of the advice it supplies and the justifications it uses when supplying it.

Our findings yield several insights. First, we see that for low-income individuals, AI advice is mostly confirming, i.e., telling low-income individuals to do what they are already doing, with the exception of the Army scenario, where it gives low-income individuals the strong recommendation of choosing Army as a career path. For high-income individuals, by contrast, the advice is confirming with respect to their Army and Baby choices, but otherwise frequently pushes them away from their actual behavior. Hence, in the decision problems that we study, AI advice tends to be confirmatory for low-income advisees and disconfirmatory for high-income advisees. To the extent that some choices made by low-income individuals constrain upward mobility, AI advice that reinforces these choices may have deleterious effects.

Second, AI advice tends to be highly resolved in that, for a given problem, it produces the same recommendation with little variance across the 100 simulations we run for each problem. In some sense, that means that AI advice represents a consensus in that, however many times the same people ask its advice on the same question, they are likely to get the same recommendation. Its recommendations are consistent across repeated prompts. Given that such advice tends to be confirming for low-income individuals, this consistency may further enforce the decisions that such individuals are already making, especially if lay users misinterpret high resolution as high confidence. The resolution of advice and the expressed confidence in advice are largely decoupled.

Third, we allow ChatGPT to offer a variety of justifications for what it recommends. Hence, it is possible that different types of decision makers are offered identical advice but that advice is justified in different ways to different people. Concretely, while we find that, on average across decision problems, ChatGPT does not vary its justifications by the advisee's gender or race, it does systematically vary them by income. For high-income advisees, it places more emphasis on *long-term benefits, regret, happiness* and *typical choice* compared to low-income advisees. In contrast, *short-term benefits, risk* and the *ease of implementation* are relatively more important justifications for low-income advisees, possibly reflecting the tighter financial and opportunity constraints that low-income individuals face when making such decisions. The biggest difference between these two groups can be observed by the different emphasis ChatGPT places on *short- and long-term benefits* when justifying its recommendations with more emphasis on *long-term benefits* for upper income advisees who may be better able to postpone gratification.

Fourth, while most of our queries for ChatGPT involve binary answers, i.e., get married or not, have children or not, etc., much of the advice we seek in life is not binary, but rather advice involving open-ended queries where the advisor proposes a set of options to choose from. The advisor in these situations provides choice sets from which the advisee might choose. If advisees adhere to making choices from these sets, then if these choice sets are different, advisees with different backgrounds may wind up making very different choices from each other. To capture this phenomenon, we asked ChatGPT to advise students applying to college, all of whom are equally qualified but come from different socioeconomic backgrounds or are of different genders, to provide a list of colleges to apply to and a set of majors to consider for their studies. What we find is that the choice sets provided to such students differ dramatically as a function of their socio-economic background and gender and these choice sets align with documented choice patterns.

To conclude, we believe that for life's Big Decisions, generative AI has the potential to be consequential precisely because these choices have large and lasting effects on individuals' lives. However, this impact is unlikely to be uniform: it depends on the interaction between the differential advice given to individuals from different socio-economic background and the types of choices these individuals would make in the absence of AI advice. The ultimate effects of generative AI will also hinge on which populations are most likely to seek out AI advice and which are most inclined to follow it. Understanding these patterns is a promising direction for future research and is essential for developing a fuller picture of both the potential benefits and the possible hazards of generative AI in shaping high-stakes life decisions.

References

Bhattacharya, U., A. Kumar, S. Visaria, and J. Zhao (2024): "Do Women Receive Worse Financial Advice?" *Journal of Finance*, 79, 3261–3307.

Bleemer, Z. and S. Quincy (2025): "Changes in the College Mobility Pipeline since 1900," *National Bureau of Economic Research*.

Blossfeld, H. P. (2009): "Educational assortative marriage in comparative perspective," *Annual Review of Sociology*, 35, 513–530.

Bucher-Koenen, T., A. Hackethal, J. Koenen, and C. Laudenbach (2023): "Gender Differences in Financial Advice," .

Burt, R. S. (1992): "Structural Holes The Social Structure Of Competition," .

Calvó-Armengol, A. and M. O. Jackson (2004): "The Effects of Social Networks on Employment and Inequality," *The American Economic Review*, 94, 426–454.

Camilleri, A. R. (2023): "An investigation of big life decisions," *Judgment and Decision Making*, 18.

Cheong, I., K. Xia, K. J. Feng, Q. Z. Chen, and A. X. Zhang (2024): "(A)I Am Not a Lawyer, But?: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice," in *2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024*, Association for Computing Machinery, Inc, 2454–2469.

Chituc, V., L. Paul, and M. Crockett (2021): "Evaluating Transformative Decisions," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43.

Curran, S. R., F. Garip, C. Chung, and K. Tangchonlatip (2005): "Gendered Migrant Social Capital: Evidence from Thailand," *Social Forces*, 84, 225–255.

Dimaggio, P. and F. Garip (2012): "Network effects and social inequality," *Annual Review of Sociology*, 38, 93–118.

Fedyk, A., A. Kakhbod, P. Li, and U. Malmendier (2025): "AI and Perception Biases in Investments: An Experimental Study," .

Fieberg, C., L. Hornuf, M. Meiler, and D. J. Streich (2025): "Using Large Language Models for Financial Advice," .

Gallen, Y. and M. Wasserman (2024): "Informed Choices: Gender Gaps in Career Advice," .

Hartmann, J., J. Schwenzow, and M. Witte (2023): "The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation," .

Hechtlinger, S., C. Schulze, C. Leuker, and R. Hertwig (2024): "The Psychology of Life's Most Important Decisions," *American Psychologist*.

Krook, J., E. Schneiders, T. Seabrooke, N. Leesakul, and J. Clos (2023): "Large Language Models (LLMs) for Legal Advice: A Scoping Review," .

Lai, V., C. Chen, Q. V. Liao, A. Smith-Renner, and C. Tan (2021): "Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies," *arXiv preprint arXiv:2112.11471*.

Leighton, M. and J. D. Speer (2026): "Family Background and College Major Choice: Evidence on Major Earnings Growth," *Education Finance and Policy*, 21, 121–145.

Oehler, A. and M. Horn (2024): "Does ChatGPT provide better advice than robo-advisors?" *Finance Research Letters*, 60.

O'malley, A. J. and N. A. Christakis (2011): "Longitudinal analysis of large social networks: Estimating the effect of health traits on changes in friendship ties," *Statistics in Medicine*, 30, 950–964.

Paul, L. A. (2014): *Transformative Experience*, Oxford University Press.

Rivera, M. T., S. B. Soderstrom, and B. Uzzi (2010): "Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms," *Annual Review of Sociology*, 36, 91–115.

Rosenfeld, M. J. (2008): "Racial, educational and religious endogamy in the united states: A comparative historical perspective," *Social Forces*, 87, 1–32.

Schotter, A. (2023): *Advice, Social Learning and the Evolution of Conventions*, Cambridge University Press.

Schwartz, C. R. and R. D. Mare (2005): "Trends in educational assortative marriage from 1940 to 2003," *Demography*, 42, 621–646.

Shool, S., S. Adimi, R. S. Amleshi, E. Bitaraf, R. Golpira, and M. Tara (2025): "A systematic review of large language model (LLM) evaluations in clinical medicine," .

Sloane, C. M., E. G. Hurst, and D. A. Black (2021): "College majors, occupations, and the gender wage gap," *Journal of Economic Perspectives*, 35, 223–248.

Ullmann-Margalit, E. (2006): "Big Decisions: Opting, Converting, Drifting," *Royal Institute of Philosophy Supplement*, 58, 157–172.

Yates, J. F., E. S. Veinott, and A. L. Patalano (2003): *Hard decisions, bad decisions: On decision quality and decision aiding*, Cambridge University Press, 1–63.

Zafar, B. (2013): "College Major Choice and the Gender Gap," *The Journal of Human Resources*, 48, 545–595.

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen (2025): "A Survey of Large Language Models," .