ChatGPT vs. Experts:

Can GenAI Develop High-Quality Organizational and Policy Objectives?

Jay Simon
Kogod School of Business
American University
jaysimon@american.edu

Johannes Ulrich Siebert
Department of Business and Management
Management Center Innsbruck
johannes.siebert@mci.edu

Abstract

This paper explores the efficacy of GenAI for value-focused thinking, specifically the ability to generate high-quality sets of objectives for organizational or policy decisions. Overall, we find that most of the GenAI objectives are viable individually, but the sets as a whole are highly flawed. They often include non-essential considerations while omitting important ones. In addition, they are redundant and non-decomposable, often because of a tendency to include means objectives even when explicitly instructed not to. However, the sets of objectives can be improved by implementing best practices in prompting and with decision analysis expertise. The results provide further evidence of the importance of a human in the loop; in this case, GenAI tools are helpful for brainstorming objectives, but an expert with a background in decision analysis is needed before the results are used to support decision making. To facilitate this, a four-step approach incorporating the relative strengths of both GenAI and decision analysts is presented and demonstrated.

Keywords: AI, value-focused thinking, objectives, preferences, human-AI interaction

1. Introduction

Generative artificial intelligence (GenAI) has surged in popularity over the last few years, with seemingly endless applications. It has permeated a wide range of settings; a few examples include health care (Moulaei et al. 2024), cybersecurity (Gupta et al. 2023), strategic planning (Spaniol and Rowland 2023), music (Louie et al. 2020), travel (Wong et al. 2023), and various business decisions (Chuma and De Oliveira 2023). In this paper, we explore its usefulness for one of the most open-ended tasks in decision analysis: generating objectives.

Generating a set of objectives for a decision setting is one of the key steps in the decision analysis process (Clemen 1997). Value-focused thinking (VFT) (Keeney 1996) is broadly used to identify and structure objectives in decision analyses (Borgonovo et al. 2025). It helps ensure that stakeholder preferences are captured thoroughly and in a way that supports high quality decision making and enables clear assessments of tradeoffs (Keeney et al. 2025). VFT can be used for personal decisions; however, because preferences may vary substantially between individual decision makers, it is difficult to establish a credible reference set of objectives for such settings. Higher level decision problems with many stakeholders provide a more viable comparison for objectives produced by GenAI. Therefore, this paper explores GenAI's effectiveness specifically in generating objectives for organizational or policy decisions.

While many applications of GenAI have been published in recent years, there is relatively little prior work on the use of GenAI for generating goals or objectives for specific settings. One example of such work is Waterfield et al. (2025), who compare goals in individualized education programs (IEPs) written by special education experts with those written by ChatGPT. They find no significant difference in experienced teachers' perception of quality between the two. Another is Raman et al. (2024), who evaluate ChatGPT concerning UN Sustainability

Development Goals. They find, in contrast, that ChatGPT displays a high degree of knowledge on the subject but limited intelligence; it is less adept at tasks involving, e.g., critical thinking or integrated problem solving. They advocate limiting its use to information gathering rather than a more active role in decision making. In a similar tenor, Vinuesa et al. (2020) report mixed findings regarding AI's use in developing SDGs and recommend further regulatory oversight for AI to sharpen sustainable development. Lieder et al. (2022) found that subjects preferred randomly selected goals to goals generated with AI across three different settings.

Ghobakhloo et al. (2024) report that GenAI can promote various objectives within the Industry 5.0 framework. GenAI holds promise for advancing key open science (OS) objectives—such as expanding equitable access to knowledge, making infrastructure use more efficient, engaging diverse societal actors, and connecting different knowledge systems. At the same time, its current limitations may undermine core research values like integrity, equity, reproducibility, and reliability. To ensure responsible use, its integration into research must be guided by critical evaluation, robust validation, and appropriate safeguards (Hosseini et al., 2025).

Recently, scholars found that the automated generation of learning objectives for new curricula using large-language models (LLMs) is of comparable quality to learning objectives developed by human instructors. However, limitations arise in terms of, for instance, its achievability (Doyle et al., 2025). In summary, even these few examples of prior work on GenAI's usefulness for producing goals and objectives reach a wide range of conclusions.

The work presented in this paper delves into the issue more directly and with a decision analysis framework. We use GenAI to produce sets of objectives for several organizational and policy decision settings and then evaluate the quality of those sets of objectives. As we will

observe and discuss, there are benefits and limitations to using GenAI for objective generation. We examine potential ways to increase the quality of the results, and then present a suggested hybrid approach that captures the strengths of GenAI while allowing humans to improve the aspects with which GenAI struggles. This work is exploratory. It is not intended to be comprehensive; instead, it offers a foundational understanding of GenAI's basic strengths and weaknesses as a tool to support VFT and offers recommendations for how to obtain high quality sets of objectives. We hope it serves as a valuable starting point for researchers and practitioners who want to incorporate AI into decisions with multiple objectives.

2. Methods

We selected six published papers that present sets of organizational or policy objectives. These specific papers were chosen because they do not merely apply VFT; each one explores the process of objectives generation and makes a thorough case that the results are appropriate for the decision setting. Well-qualified decision analysts contributed to each paper; all six were published in this journal. Thus, we can confidently state that each paper's set of objectives serves as a reasonable benchmark to which GenAI output can be compared. In addition, because each setting involves preferences of organizations or policy makers rather than of personal decision makers, objectives should not be affected by differences in tastes or priorities between individuals.

We used several different GenAI tools to produce objectives, including GPT-4o Mini, Claude 3.7 Sonnet, Gemini 2.5 Pro, and Grok-2. All of the sets of objectives assessed in the paper were produced by GPT-4o Mini; as discussed later, the paper's main qualitative findings are robust to the choice of tool. For each paper, two different prompts were entered. The first

prompt was intended to capture a simple request that might be given by a layperson encountering the problem. It consisted of one sentence asking for objectives for the specific setting, followed by: "*Each objective should be expressed as a simple statement beginning with "maximize" or "minimize."*"[1] For instance, the following prompt was used to elicit objectives for building evacuation based on the setting given by Keeney (2012): "*Create a set of objectives when making decisions on measures for the evacuation of large buildings. Each objective should be expressed as a simple statement beginning with "maximize" or "minimize."*" We purposefully did not use exact wording from the papers to describe the setting. This was to reduce the likelihood of the GenAI tool drawing its response from the paper itself. Instead, we used a description of the setting that reflected the paper accurately but was sufficiently general to allow the response to incorporate a wider body of material.

The second prompt for each paper adds more detailed guidance about what constitutes a good set of objectives (Keeney 1996) and an explicit instruction to produce fundamental objectives rather than means objectives. The text added to the second prompt is identical for each of the six papers and is shown in Table 1. Certainly, there is not universal agreement on what additional guidance would be the most helpful or appropriate to include in a prompt. However, the general findings of this paper are robust to modest changes to the prompts.

> The objectives should be fundamental objectives rather than means objectives; that is, they should capture direct concerns of decision makers rather than factors that are important only because of their implications on the degree to which more fundamental objectives can be achieved.

[1] This sentence was added to the prompts after initial experimentation revealed that GenAI tools tended not to express objectives in the manner most common in decision analysis. (For instance, we observed generic high-level objectives followed by groups of single words or phrases.) For use in decision analysis, Keeney (1992) recommends formulating objectives using a verb, an object, and a preference direction. The GenAI tools often did not use this fundamental and straightforward structure.

The set of objectives should follow the principles of value-focused thinking, which means they should have the following properties:

Essential – only objectives that must be considered when determining the best alternative should be included.

Controllable – an objective should only be included if performance on it can be affected by the choice of alternative.

Complete – all considerations that might reasonably affect the choice of alternative should be included.

Measurable – it must be possible to capture each objective with the use of clear and unambiguous attributes.

Operational – it must be possible to obtain the relevant information required to assess alternatives' performance on each objective.

Decomposable – the impacts of an alternative's performance on each objective should be able to be assessed individually, and those individual performances aggregated; that is, there should not be any substantial preference interactions between objectives.

Nonredundant – no aspects of the decision makers' preferences should be double-counted. *Concise* – very similar objectives should be grouped together.

Understandable – objectives should be clear to all stakeholders and use as little jargon as possible.

Table 1. Additional guidance provided in the detailed prompts based on Keeney (1996).

Each GenAI response was evaluated on the nine desirable properties given by Keeney (1996) for sets of objectives, which are also included in the detailed prompts, as shown in Table 1. The evaluations were expressed via a Likert scale from 1 (Strongly Disagree) to 7 (Strongly Agree) for a statement that the set of objectives satisfied the given property. To maximize the accuracy of the assessments, we took the following five-step approach. First, each of the two authors assessed all twelve sets of objectives individually. Second, two holistic discussions were held. One discussion covered each set of objectives to ensure that the authors shared a consistent understanding of each paper's decision setting and any relevant implications on assessment. The other discussion covered each of the nine properties to ensure that the authors were similarly calibrated. Third, after the discussions concluded, each author had an opportunity to revise their

initial ratings. Only a few minor adjustments were made[2]. The resulting two sets of ratings were extremely similar, with a Spearman rank correlation of 0.927 (where ties were assigned their average rank) and a weighted Cohen's kappa of 0.732 using linear weighting or 0.916 using quadratic weighting[3]. Fourth, joint ratings for each set of objectives on each property were determined as follows. When the difference between the two authors' ratings was 1, the midpoint between the two values was used. Because the ratings are ordinal, there is no meaningful difference between choosing, e.g., 5.5 and 5.8. Larger differences were discussed on a case-by-case basis to reach agreement on a joint rating. (There were four such cases.)

Finally, to ensure that the ratings were reasonable, they were reviewed by at least one author of each of the six papers (none of whom is an author of the current paper). These authors are all experienced decision analysts who applied VFT in the paper's setting and thus have a thorough understanding of both the domain and the properties of a high-quality set of objectives. The authors of the original papers confirmed that the assessments were generally reasonable, but did suggest a few adjustments. For example, we increased the *completeness* scores for one paper because an author informed us that the objectives omitted by the GenAI tool were relatively unimportant. We recognize that universal agreement on the final scores is impossible, but we strove to follow a rigorous process that minimized the likelihood of large errors or systematic biases.

Note that the properties require varying degrees of comparison with the published paper. Assessing the *completeness* of a GenAI set of objectives, for instance, is heavily dependent on

[2] For instance, one author increased a pair of ratings for *operational* after learning that the scope of a paper's setting was broader than first believed.

[3] Quadratic weighting leads to a higher value of kappa because it penalizes larger differences more heavily and there were no very large differences between scores.

the published set of objectives. Assessing *nonredundance*, on the other hand, requires some domain knowledge but no direct comparison to the paper's objectives.

3. Results

This section presents our scores for all twelve of the prompts, as well as brief discussions of the sets of objectives. It is divided into three subsections; we present the results for each of the six individual papers first, followed by the overall results. All 108 scores are shown in Table 2. The third subsection contains additional results obtained using further improvements to prompts for two of the six papers.

	Keeney (2012)		Dees, Nestler, and Kewley (2013)		Simon, Regnier, and Whitney (2014)		Couce-Vieira, Rios Insua, and Kosgodagan (2020)		Caballero, Naveiro, and Rios Insua (2022)		Methling et al. (2022)	
Essential	3.5	3	3	3.5	2.5	2	2	2.5	2.5	3	4	4
Controllable	5.5	5.5	5	5	5.5	5	5	4.5	4	4	5	5.5
Complete	1.5	1.5	2.5	3.5	6	5.5	1.5	3	2	2	5	5
Measurable	3.5	3.5	5	5	5.5	5.5	4	4	4.5	5	6	6
Operational	3	3	5	5	5	5	4	4	5	5	6	6
Decomposable	1	2	3.5	3.5	2.5	3	1.5	1.5	2.5	3	3	3
Nonredundant	3	3	3.5	4	2.5	3	2	2	2	3.5	3.5	2.5
Concise	2	2	2.5	3	2	2	2	2	1	2.5	2	2
Understandable	6.5	6.5	6.5	6.5	6	6	5	5	5	5	6.5	6.5

Table 2. All assessment scores. For each paper, the first column reflects the objectives from the simple prompt, and the second column from the detailed prompt.

3.1. Results for Individual Papers

This subsection shows the simple prompt used for each paper (with the sentence about the maximize or minimize format omitted) along with a brief summary of our observations. All

scores are shown in Table 2, and the sets of objectives themselves are in Appendix A in the online supplement.

Value-focused brainstorming (Keeney 2012)

Simple prompt: "*Create a set of objectives when making decisions on measures for the evacuation of large buildings.*"

Both prompts produced 20 objectives with no grouping, leading to low scores for *concise*. In our sample, this was not the exception but the norm; all subsequent prompts for the other papers, except for one, produced 20 objectives and received similarly low scores for *concise*. Many of the objectives were means objectives, leading to low scores for *decomposable* and *nonredundant*. The low scores for *complete* were due to the lack of any objectives capturing property operations, risks to responders, or impact on friends/relatives. Upon a review of the two sets of objectives, Keeney (personal communication, March 10, 2025) offered the following comment: "Both lists are better than most individuals could create. However, neither list should be used for a quality decision analysis, as you should only include the fundamental objectives in explicitly evaluating alternatives."

Whole Soldier performance appraisal to support mentoring and personnel decisions (Dees, Nestler, and Kewley 2013)

Simple prompt: "*Create a set of objectives to be used to assess the performance of junior enlisted soldiers.*"

The simple prompt produced a set of objectives that omitted motivation, resilience, confidence, self-esteem, and work ethic, leading to a low score on *complete*. The detailed prompt led to several minor improvements, though both sets included only one very general objective in the

physical domain. (The original paper categorizes objectives into three domains: moral, cognitive, and physical.) Both sets also included a few higher-level or long-term aims (e.g., mission readiness) that are difficult to connect directly to a junior enlisted soldier, which reduced otherwise strong performance on *controllable*.

A value-focused approach to energy transformation in the United States Department of Defense (Simon, Regnier, and Whitney 2014)

Simple prompt: "*Create a set of objectives for energy decisions in a department of defense.*"

The two sets of objectives were similar to one another. Unlike several of the other papers, the GenAI sets of objectives, in this case, were close to *complete*. A possible explanation for this anomalous result is that the objectives in the original paper are extracted from publicly available documents that are likely also visible to GenAI tools. There were many means objectives, which led to poor performance on *decomposable* and *nonredundant*. Both sets included scalability as an objective, which is not included in the paper.

Assessing and forecasting cybersecurity impacts (Couce-Vieira, Rios Insua, and Kosgodagan 2020)

Simple prompt: "*Create a set of objectives for an organization to use when making cybersecurity planning or resource allocation decisions.*"

Again, both sets of objectives included a large number of means objectives, leading to low scores for *decomposable* and *nonredundant*, as preference interactions would arise in many combinations of objectives. The simple prompt produced only one general catch-all objective

related to the actual impacts of cybersecurity incidents, leading to a low score for *complete*. Again, both sets included scalability as an objective, which was not in the original paper.

Modeling ethical and operational preferences in automated driving systems (Caballero, Naveiro, and Rios Insua 2022)

Simple prompt: "*Create a set of objectives for decisions involving the development and implementation of automated driving systems (ADSs).*"

The simple prompt produced 20 objectives, while the detailed prompt produced 15. In both cases, the objectives did not capture speed or comfort. As previously mentioned, many of the GenAI objectives were means objectives (e.g., "Minimize Response Time" and "Minimize False Positives and Negatives"), leading to low scores for *decomposable* and *nonredundant*. There were fewer in the detailed prompt, however, leading to substantial increases in the *nonredundant* and *concise* scores.

Supporting innovation in early-stage pharmaceutical development decisions (Methling et al. 2022)

Simple prompt: "*Create a set of objectives for pharmaceutical companies when making early-stage pharmaceutical portfolio decisions.*"

The sets of objectives produced by the simple and detailed prompts were very similar. Both ignored the objective of reducing the economic cost to society and included several redundant objectives regarding discontinuation costs and potential return on investment. Unlike the paper, the GenAI sets of objectives included portfolio-level objectives (e.g., "Maximize Portfolio

Diversity"). We attributed this to a reasonable difference in scope and did not penalize the GenAI objectives for deviating from the paper in this way.

3.2. Overall Results

Several general observations about the sets of scores can be made, the most striking of which is that GenAI performed substantially better on some properties than on others. The GenAI sets of objectives received high scores for *controllable*, *measurable*, *operational*, and *understandable*, with medians of 5, 5, 5, and 6.25, respectively. They received lower scores for *essential*, *complete*, *decomposable*, *nonredundant*, and *concise*, with medians of 3, 2.75, 2.75, 3, and 2, respectively. A Kruskal-Wallis test, an analog of one-way ANOVA suitable for ordinal data, revealed significant differences in the medians of the nine properties ($p < 0.0001$).

In many cases, the lower scores on *decomposable* and *nonredundant* were due to the prevalence of means objectives. Because the original papers provide thorough discussions of objectives in each setting, means objectives were often easy to identify. For instance, both prompts for Simon, Regnier, and Whitney (2014) produced an objective related to logistical requirements, which are explicitly discussed as a means objective in that paper. If an objective does not appear in the original paper, it can be considered a means objective if it is relevant only or primarily due to its influence on fundamental objectives. For instance, the prompts for Caballero, Naveiro, and Rios Insua (2022) produced an objective to minimize response time, the relevance of which is clearly due to its impact on fundamental objectives regarding safety.

Surprisingly, the differences between the sets of objectives produced by the simple and detailed prompts were minor. Despite the detailed prompts including explicit instructions to ensure that the objectives were fundamental and satisfied the nine properties, improvements

occurred in only 15 out of 54 possible scores, and ten of those were increases of 0.5. (Six decreases in scores occurred as well.) Cliff's delta was 0.060, suggesting a small overall improvement. Multiple authors of the original papers remarked on the similarity of the two sets of objectives. Overall, the detailed prompts appeared to produce slightly better sets of objectives, but the improvement was minimal.

3.3. Results from Improved Prompting Techniques

In this subsection, we present the results of two additional approaches intended to improve GenAI output. The first is chain-of-thought (CoT) prompting (Wei et al. 2022). There is a wide variety of CoT methods, but the core concept is to include step-by-step or intermediate reasoning in the prompt to help guide how the task is accomplished.

Of course, there are many different processes by which objectives can be generated. We implement CoT prompting by including adapted text from Table 7.3 in Keeney (2007), which lists and explains nine different devices that can be useful in helping individuals articulate their objectives (for example, constructing a wish list of what would be desired of an ideal alternative). This text is appended to the detailed prompts used previously.

The second approach is to provide the GenAI tool with a critique of its initial set of objectives and request a revised set that addresses specific shortcomings. We refer to this approach as critique-and-revise (CaR). Note that CaR in this context requires expertise in VFT; a layperson cannot easily replicate it. In addition, for the purposes of this paper, we must be careful not to offer critiques that are so specific as to spell out an intended "correct" solution. For example, we can tell it that Objective A and Objective B are not decomposable, but we cannot tell it how to restate those objectives.

We conducted a 2x2 factorial experiment using the detailed prompt for each of Couce-Vieira, Rios Insua, and Kosgodagan (2020) and Methling et al. (2022), varying whether or not CoT and CaR are used. The scores are shown in Table 3. The full prompting explanations and sets of objectives are in Appendix B in the online supplement. More refined prompting appears to produce smaller sets of objectives; for each of the two papers, the output obtained using both CaR and CoT contained only six objectives. Unsurprisingly, these smaller sets received higher scores for *essential* and *concise*. They were also more *decomposable* and *nonredundant*, largely due to the removal of several means objectives.

	Couce-Vieira, Rios Insua, and Kosgodagan (2020)				Methling et al. (2022)			
	Previous Detailed Prompt	Critique and Revise (CaR) Only	Chain of Thought (CoT) Only	Both CaR and CoT	Previous Detailed Prompt	Critique and Revise (CaR) Only	Chain of Thought (CoT) Only	Both CaR and CoT
Essential	2.5	4.5	4.5	5.5	4	4.5	4	5.5
Controllable	4.5	5.5	5.5	5.5	5.5	6	6	6
Complete	3	3.5	4	4	5	4.5	4.5	4.5
Measurable	4	5	5	5	6	6	6	6
Operational	4	5	5	5	6	6	6	6
Decomposable	1.5	2	1.5	4.5	3	3	2.5	4.5
Nonredundant	2	2.5	2.5	5	2.5	3	2.5	5
Concise	2	3.5	3	4.5	2	3	2	5
Understandable	5	6	6.5	6.5	6.5	6	6	6

Table 3. Scores for objectives obtained using two additional prompting techniques.

Table 3 suggests that, similarly to the inclusion of VFT principles in the prompt, each of CoT and CaR yields a modest benefit. Cliff's delta for CoT and CaR individually (relative to the detailed prompts using neither CoT nor CaR) is 0.167 and 0.231, respectively, indicating some

benefit. However, the combination of them results in even further improvement, with a Cliff's delta of 0.386 when compared to the initial detailed prompt. This suggests a cumulative effect: while no one individual technique will overcome all of the challenges of VFT, following best practices of both prompting and decision analysis may be able to produce a high quality set of objectives with assistance from GenAI.

4. Discussion and Recommendations

Perhaps the biggest shortcoming observed in the use of GenAI was its proclivity toward producing means objectives. Means objectives can provide further context and detail regarding how fundamental objectives might be achieved, but they should not be used to evaluate the actual tradeoffs that decision makers are willing to make (Keeney 2002). Notably, the detailed prompt contained an explicit direction not to include means objectives, which was largely ignored.

The impact of means objectives on several properties is straightforward; for instance, such objectives are often not *decomposable* when they are means to the same fundamental objective(s). However, for other properties, the impact is less clear-cut. Should a set of objectives be considered *complete* and *essential* if a fundamental objective is excluded but a comprehensive and non-overlapping set of means objectives supporting it is included? Our judgment was yes: provided none of the means objectives is superfluous, and they capture the fundamental objectives thoroughly, the set would still be considered *complete* and *essential*. (Despite this, the GenAI objectives struggled with both of these properties.)

The GenAI sets of objectives tended to be *controllable*, *measurable*, *operational*, and *understandable*. That is, broadly, GenAI did quite well in terms of the properties that relate to

objectives individually. The objectives were easy to understand, and it was clear in most cases that they could be influenced by the choice of alternative, captured with measurable attributes, and their performance assessed. The only exception was *essential*; many objectives appeared that were not crucial to consider when evaluating alternatives. These findings suggest that GenAI could be valuable as a brainstorming tool, as it produced many objectives that could be part of a viable set used to compare alternatives. In his assessment of the GenAI objectives' scores for two of the six papers, Rios Insua (personal communication, April 10, 2025) observed that the objectives have shortcomings but could serve as a useful starting point. Bond, Carlson, and Keeney (2008, 2010) found that decision makers often struggle to identify all of their objectives; GenAI, while not a panacea, can undoubtedly help mitigate that problem.

On the other hand, GenAI sets of objectives fared poorly on *complete*, *decomposable*, *nonredundant*, and *concise* (and *essential*, as mentioned) unless more advanced prompting techniques and DA expert feedback were both applied. While GenAI produced many high quality individual objectives, it struggled to produce coherent sets of objectives. This suggests that GenAI can be valuable for brainstorming objectives but cannot replace a decision analysis expert when conducting VFT.

This is consistent with assessments of GenAI's strengths and weaknesses in other contexts. For instance, Barcaui and Monat (2023) provide a detailed comparison between expert and GenAI performance in project planning. They recommend a collaborative approach and suggest using GenAI output as a starting point. Similarly, Spaniol and Rowland (2023) find that GenAI tools help produce many scenarios quickly and cheaply but are unlikely to outperform experts. Kocoń et al. (2023) find that while ChatGPT performs well on straightforward semantic tasks, it consistently struggles with more complex pragmatic tasks that require subjective

judgment. This finding is supported by earlier research exploring the integration of AI in human-centric sales processes, where both AI and human experts should take key roles (Paschen, Wilson, and Ferreira, 2020). Even on a strategic level, AI was found to change marketing strategies and customer behaviors, but it is more effective if it augments rather than replaces human managers (Davenport et al. 2019). A broad consensus appears to be emerging that GenAI tools are handy for producing raw content efficiently but are not an adequate substitute for human expertise in challenging real-world tasks that can be nuanced, murky, and ill-defined.

The limitations of GenAI performance in this paper's examples were likely due to multiple factors. The decision settings were complex and thus not encountered very often (Thaler and Sunstein 2009). They were not standard problems discussed frequently online and thus might not be covered thoroughly in the corpus of text used to train LLMs. It is also possible that certain concepts might be well understood by decision analysts but have not gained sufficient traction throughout the population. This could explain, for instance, why GenAI struggles with the distinction between means objectives and fundamental objectives[4], and with less tangible properties such as *decomposability*. It is also consistent with the findings of Acerbi and Stubbersfield (2023), who observe that LLMs retain and can even magnify human biases regarding what information is and is not retained when conveying content.

Because all six of the VFT examples explored in Section 3 were published in *Decision Analysis*, it is also helpful to ensure robustness of the main findings across publication outlets. We generated and scored sets of objectives using both simple and detailed prompts for decision settings of three additional papers that were published in other journals. Information about the

[4] Runge and Walshe (2014), e.g., state that: "Organizations with a strong emphasis on science or evidence-based decision-making are commonly distracted by the drivers of system dynamics rather than fundamental objectives and key value-based trade-offs" (p. 31).

papers, prompts, and results are in Appendix C in the online supplement. Overall, the results are similar. GenAI tools tended to perform well on properties related to individual objectives and poorly on properties regarding the set as a whole.

Another potential concern is overreliance on GenAI tools (Buçinca, Malaya, & Gajos 2021). It is possible that decision makers will simply accept the initial GenAI output as sufficient and proceed with that set of objectives. We strongly caution against that, and urge the adoption of a more thorough approach that captures the relative benefits of both GenAI and human experts more fully.

Based on this paper's findings, we recommend the following four-step approach for leveraging GenAI to create valuable sets of objectives, as illustrated in Figure 1:

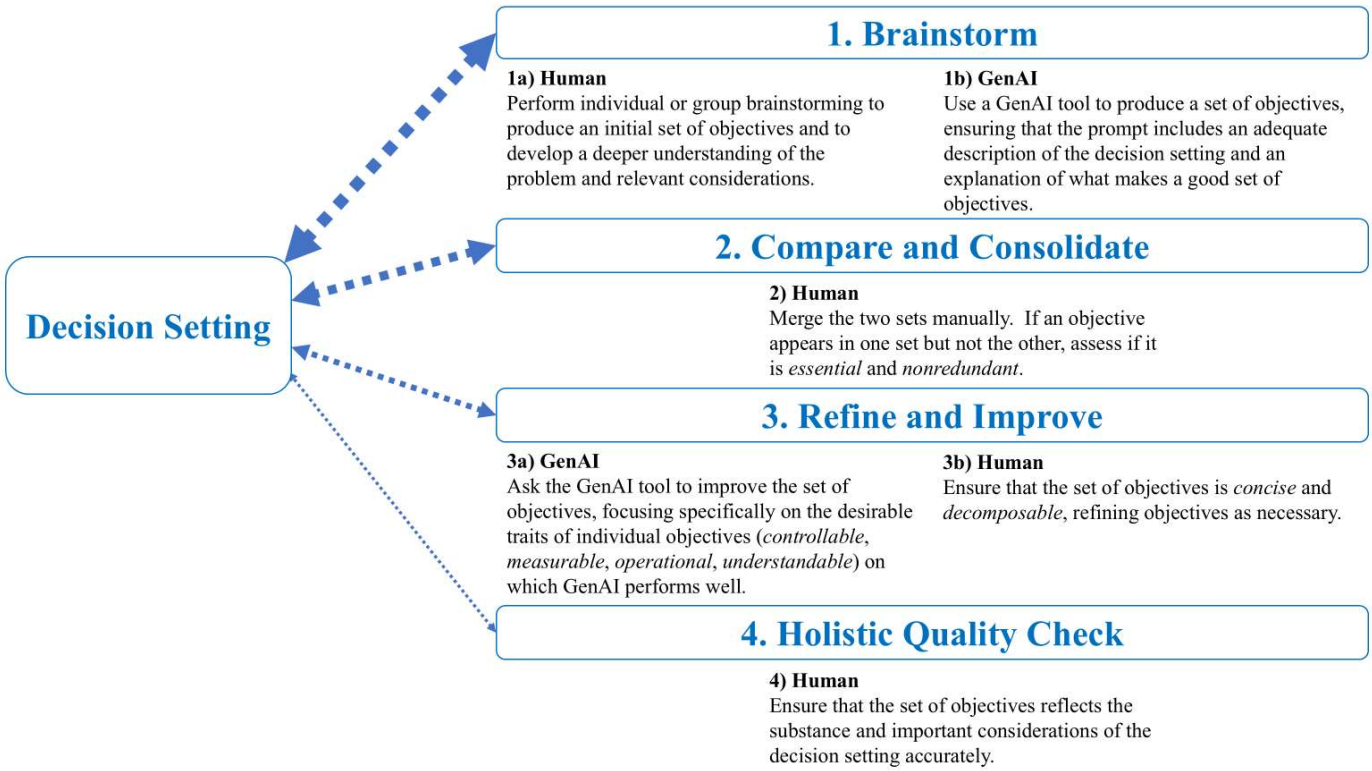

Figure 1. Identifying a high quality set of objectives using human and artificial intelligence.

1a) Human Brainstorm: Perform individual or group brainstorming to produce an initial set of objectives and to develop a deeper understanding of the problem and relevant considerations. We recommend assessing the extent to which the objectives match the decision setting. It is possible that intensive consideration of the objectives will lead to an improved understanding of the real-world problem. This is illustrated by the largest bi-directional arrow in Figure 1. Further insights about the decision setting might be obtained at any step of the process, but we emphasize it here especially, as it reflects the first effort to generate potential objectives and may lead to high-level changes in how the problem is framed. We recommend starting with human brainstorming to avoid the decision makers being anchored on or otherwise biased by the objectives identified with Gen AI.

1b) GenAI Brainstorm: Use a GenAI tool to produce a set of objectives, ensuring that the prompt includes an adequate description of the decision setting, a detailed explanation of what makes a good set of objectives, and suggestions for how to help articulate objectives. We recommend adding an explanation on how to formulate an objective in the prompt: *Each objective should be expressed as a simple statement beginning with "maximize" or "minimize."*

2) Compare and Consolidate: Merge the two sets manually. If an objective appears in one set but not the other, assess if it is *essential* and *nonredundant*. This process simply involves merging the unique objectives from both sets into a single consolidated list of objectives and is similar to combining sets of objectives from different stakeholders (e.g., Keeney et al. 1987, 1995; von Winterfeldt 1987), from different sources (Siebert, von Winterfeldt, and John 2016), or from different individual brainstormings (Keeney 2012). To carry out this task, the decision analyst(s) should have a thorough understanding of the sets of objectives under consideration (Siebert and von Winterfeldt 2020). As in the first step, an in-depth consideration of the

objectives could lead to changes in the decision setting. However, these changes are likely to become smaller as the process advances; this is conveyed in Figure 1 by increasingly narrow lines.

3a) GenAI Refine and Improve: Ask the GenAI tool to improve the set of objectives, focusing specifically on the desirable traits of individual objectives (*controllable*, *measurable*, *operational*, *understandable*) on which GenAI performs well. If a decision analysis expert is participating in the process, consider also identifying any particular weaknesses of the set of objective and asking the GenAI tool to address them in this step. It is not strictly necessary to adopt the updated objectives generated here; changes that detract from the quality of the set should be rejected. Again, we recommend adding an explanation on how to formulate an objective in the prompt: *Each objective should be expressed as a simple statement beginning with "maximize" or "minimize."*

3b) Human Refine and Improve: Ensure that the set of objectives is *concise* and *decomposable*, refining objectives as necessary. Conciseness can be improved by grouping similar objectives; e.g., Couce-Vieira, Rios Insua, and Kosgodagan (2020) group all of the objectives that describe harm to individuals together. Decomposability requires aggregating considerations with significant preference interactions; e.g., Simon, Regnier, and Whitney (2014) combine the likelihood of attacks and disruptions and the magnitude of their impacts into a single vulnerability objective. The use of an additive value function requires decomposability.

4) Holistic Quality Check: Ensure that the set of objectives reflects the substance and important considerations of the decision setting accurately. For decisions involving several decision makers, we recommend presenting the set of objectives to the group for discussion and final confirmation (Methling et. al. 2022). It is also important to be mindful of any other challenges or

biases that might hurt the quality of the objectives; for example, the various areas of consideration should be presented in a balanced way to avoid a splitting bias (von Winterfeldt and Edwards 1986). If deficiencies are identified, it may be necessary to revisit earlier steps; simple issues can be resolved via specific improvements to the set of objectives as in Step 3, but a foundational concern that influences understanding of the decision setting could even warrant returning to Step 1.

This approach is relatively straightforward and incorporates the contributions of both GenAI and human subject matter experts. For organizational and policy decisions, several of the steps are likely to require decision analysis expertise. These tasks involve refining and updating objectives and, thus, familiarity with the desirable properties. However, for smaller individual decisions, this approach is likely to improve decision making even in the absence of a decision analyst. As mentioned previously, individual decision makers often fail to identify many of their relevant objectives, and GenAI can help mitigate that problem. It can also help train individuals in the skill of identifying objectives for decisions faced in daily life. It should be noted that the cost of applying additional prompts in a GenAI tool at the scale required by the four-step approach is trivial relative to the stakes of organizational decision-making; at the time of this writing, it is easily within the limitations of the free version of ChatGPT 4o mini.

The remainder of this section provides a brief demonstration of the approach using the defense energy setting of Simon, Regnier, and Whitney (2014).

Step 1a) Human Brainstorm. This initial step is notional in this example, and contains a set of objectives that is not intended to be of high quality. The objectives are shown in the left column of Table 4.

Step 1b) GenAI Brainstorm. For this step, we use the detailed prompt used previously for this paper, but with the CoT text from Keeney (2007) added. The resulting objectives are shown in the right column of Table 4.

Objectives from Step 1a) Human Brainstorm	Objectives from Step 1b) GenAI Brainstorm
1. Minimize energy cost	1. Maximize mission readiness enabled by energy availability
2. Minimize energy consumption	2. Maximize continuity of operations during energy supply disruptions
3. Minimize environmental impact	3. Maximize safety of personnel in energy production, transport, and use
4. Minimize dependence on foreign sources	4. Maximize security of energy supply against hostile interference
5. Maximize operational capability	5. Maximize ability to meet operational energy needs in all environments
6. Maximize reliability	6. Maximize resilience of critical systems to energy infrastructure failure
7. Minimize personnel impact of an attack on energy infrastructure	7. Maximize cost-effectiveness of meeting operational and facility energy needs
8. Maximize domestic energy jobs	8. Maximize environmental protection from energy-related activities
9. Minimize political friction	9. Maximize compliance with applicable laws, regulations, and treaties related to energy
	10. Minimize dependence on vulnerable or adversary-controlled energy sources
	11. Minimize adverse health impacts from energy production and use
	12. Minimize logistical burden of energy supply on operations

Table 4. Objectives produced in Steps 1a and 1b for Simon, Regnier, and Whitney (2014).

Step 2) Compare and Consolidate. The following six objectives appear in both lists (the GenAI versions are shown):

- Maximize continuity of operations during energy supply disruptions.
- Maximize environmental protection from energy-related activities.
- Minimize dependence on vulnerable or adversary-controlled energy sources.
- Maximize mission readiness enabled by energy availability.
- Maximize safety of personnel in energy production, transport, and use.
- Maximize cost-effectiveness of meeting operational and facility energy needs.

The following three objectives, appearing in only one of the two lists, were deemed to be *essential* and *nonredundant*:

- Minimize energy consumption
- Maximize security of energy supply against hostile interference
- Minimize logistical burden of energy supply on operations

Thus, at the end of Step 2, we have a set of nine objectives.

Step 3a) GenAI Refine and Improve. In this step, the GenAI tool was asked to improve the set of objectives. The full prompt is shown in Table 5. It produced the following set of objectives:

1. Maximize ability to maintain operations during energy supply disruptions
2. Maximize protection of air, water, and land from energy-related harm
3. Minimize reliance on energy sources controlled by adversaries or prone to disruption
4. Maximize operational readiness supported by reliable energy access
5. Maximize safety of personnel involved in energy generation, transport, and use
6. Maximize value delivered per dollar spent on energy for operations and facilities
7. Minimize total energy use without reducing mission capability
8. Maximize protection of energy systems from sabotage, attack, or interference
9. Minimize the weight, volume, and complexity of energy systems required in operational settings

Please improve the following set of objectives: • Maximize continuity of operations during energy supply disruptions. • Maximize environmental protection from energy-related activities. • Minimize dependence on vulnerable or adversary-controlled energy sources. • Maximize mission readiness enabled by energy availability.

- Maximize safety of personnel in energy production, transport, and use.
- Maximize cost-effectiveness of meeting operational and facility energy needs.
- Minimize energy consumption
- Maximize security of energy supply against hostile interference
- Minimize logistical burden of energy supply on operations

Each objective should be expressed as a simple statement beginning with maximize or minimize.

Focus specifically on making sure that each objective satisfies the following properties:
Controllable: an objective should only be included if performance on it can be affected by the choice of alternative.
Measurable: it must be possible to capture each objective with the use of clear and unambiguous attributes.
Operational: it must be possible to obtain the relevant information required to assess alternatives' performance on each objective.
Understandable: objectives should be clear to all stakeholders and use as little jargon as possible.

Table 5. The prompt used to obtain an improved set of energy objectives.

Step 3b) Human Refine and Improve. For conciseness, Objectives 1 and 8 were combined into the following objective: "Minimize vulnerability of supply and logistics." Objective 6 was changed to "Maximize net economic impact" to avoid redundance, as the nonmonetary benefits are captured in other objectives. Objective 7 was then simplified to "Maximize mission capability," again to avoid redundance, as the impacts of energy usage are captured elsewhere. Objective 4 was removed because it is a means objective; the relevance of operational readiness is due to its influence on mission capability, and including it would reduce decomposability. Objective 9 was also removed, as the costs and operational impacts of logistical complexity are captured elsewhere.

Step 4) Holistic Quality Check. The previous three steps yielded the following set of six objectives:

1. Minimize vulnerability of supply and logistics

2. Maximize protection of air, water, and land from energy-related harm
3. Minimize reliance on energy sources controlled by adversaries or prone to disruption
4. Maximize safety of personnel involved in energy generation, transport, and use
5. Maximize net economic impact
6. Maximize mission capability

The final step in this example, like the initial human brainstorming step, is hypothetical. However, based on the discussion of objectives in Simon, Regnier, and Whitney (2014), it is possible that a few of these objectives might be modified to reflect more fundamental concerns of high-level decision makers in defense. For instance, minimizing vulnerability of supply and logistics is a fundamental objective for many decision makers, but at an organizational level, it is a means objective for maximizing mission capability and safety of personnel.

5. Conclusion

This paper examines the strengths and weaknesses of GenAI for developing objectives for organizational and policy decisions. It also offers a suggested approach for leveraging the strengths of these tools while incorporating human expertise to ensure that the set of objectives is viable for conducting decision analysis thoroughly. While the work in this paper offers a foundation for the use of GenAI in VFT, it is exploratory. There are several possible avenues for additional research on the topic. One is a thorough comparison of different GenAI tools' relative strengths and weaknesses. We ran a subset of the prompts using several tools for robustness, and the main results held broadly across tools. The only exception was Gemini, which consistently produced smaller sets of objectives; an example from the simple prompt for Simon, Regnier, and Whitney (2014) is shown here:

1. Maximize energy efficiency in military operations.
2. Minimize energy consumption in non-combat activities.

3. Maximize the use of renewable energy sources.
4. Minimize reliance on fossil fuels.
5. Maximize energy resilience in critical infrastructure.
6. Minimize energy waste in logistics and supply chains.

The GenAI landscape is dynamic and changes rapidly; new tools appear frequently, as do new versions of existing tools, and thus, we purposefully examined results across tools only as a robustness check and not as a primary focus of the paper.

Another possible approach is to conduct more prolonged interactions with more sustained back-and-forth conversations between a decision analyst and a GenAI tool. We were able to achieve some improvements to sets of objectives with a few simple approaches, but more refined prompting techniques and structures could potentially lead to more substantial improvements.

Finally, observing the results of implementing GenAI-aided objectives in a real organization or policy setting would be valuable. As demonstrated in this paper, there are clear and tangible differences between sets of objectives created by GenAI and by decision analysts. However, less transparent differences may become apparent only when the objectives are used in practice.

Acknowledgments

We are very grateful to Ralph Keeney, Florian Methling, Scott Nestler, Eva Regnier, and David Rios Insua for reviewing the GenAI objectives and our scores, as well as providing helpful comments and suggestions. We also appreciate the valuable feedback given by two anonymous reviewers and an Associate Editor that helped us improve the paper.

References

Acerbi, A., & Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, **120**(44), e2313790120.

Barcaui, A., & Monat, A. (2023). Who is better in project planning? Generative artificial intelligence or project managers?. *Project Leadership and Society*, **4**, 100101.

Bond, S. D., Carlson, K. A., & Keeney, R. L. (2008). Generating objectives: can decision makers articulate what they want?. *Management Science*, **54**(1), 56-70.

Bond, S. D., Carlson, K. A., & Keeney, R. L. (2010). Improving the generation of decision objectives. *Decision Analysis*, **7**(3), 238-255.

Borgonovo, E., Jose, V. R. R., Knowlton, M., Shachter, R., Siebert, J. U., & Ulu, C. (2025). Fifty years of decision analysis in operational research: A review. *European Journal of Operational Research*, https://doi.org/10.1016/j.ejor.2025.05.023.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, **5**(CSCW1), 1-21.

Caballero, W. N., Naveiro, R., & Rios Insua, D. (2022). Modeling ethical and operational preferences in automated driving systems. *Decision Analysis*, **19**(1), 21-43.

Chuma, E. L., & De Oliveira, G. G. (2023). Generative AI for business decision-making: A case of ChatGPT. *Management Science and Business Decisions*, **3**(1), 5-11.

Clemen, R. T. (1996). *Making hard decisions: an introduction to decision analysis* (Vol. 2). Belmont, CA: Duxbury Press.

Couce-Vieira, A., Rios Insua, D., & Kosgodagan, A. (2020). Assessing and forecasting cybersecurity impacts. *Decision Analysis*, **17**(4), 356-374.

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, *48*, 24-42.

Dees, R. A., Nestler, S. T., & Kewley, R. (2013). WholeSoldier performance appraisal to support mentoring and personnel decisions. *Decision Analysis*, **10**(1), 82-97.

Doyle, A., Sridhar, P., Agarwal, A., Savelka, J., & Sakr, M. (2025). A comparative study of AI-generated and human-crafted learning objectives in computing education. *Journal of Computer Assisted Learning*, *41*(1), e13092.

Ghobakhloo, M., Fathi, M., Iranmanesh, M., Vilkas, M., Grybauskas, A., & Amran, A. (2024). Generative artificial intelligence in manufacturing: opportunities for actualizing Industry 5.0 sustainability goals. *Journal of Manufacturing Technology Management*, *35*(9), 94-121.

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. *IEEE Access*, **11**, 80218-80245.

Hosseini, M., Horbach, S. P., Holmes, K., & Ross-Hellauer, T. (2025). Open Science at the generative AI turn: An exploratory analysis of challenges and opportunities. *Quantitative Science Studies*, *6*, 22-45.

Keeney, R. L., Gregory, R., & Slovic, P. (2025). Proactively planning for possible future crises. *Decision Analysis*, https://doi.org/10.1287/deca.2024.0293.

Keeney R. L., McDaniels, T. L., & Swoveland, C. (1995). Evaluating improvements in electric utility reliability at British Columbia Hydro. *Operations Research*, **43**(6), 933–947.

Keeney R. L., Renn, O., & von Winterfeldt, D. (1987). Structuring West Germany's energy objectives. *Energy Policy*, **15**(4), 352–362.

Keeney, R. L. (1996). *Value-focused thinking: A path to creative decisionmaking*. Cambridge, MA: Harvard University Press.

Keeney, R. L. (2002). Common mistakes in making value trade-offs. *Operations Research*, **50**(6), 935–945.

Keeney, R. L. (2007). Developing objectives and attributes. In *Advances in Decision Analysis*, W. Edwards, R.F. Miles, and Detlof von Winterfeldt (Eds.), Cambridge University Press, 104-128.

Keeney, R. L. (2012). Value-focused brainstorming. *Decision Analysis*, **9**(4), 303-313.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Milkowski, P., Oleksy, M., Piasecki, M., Radliński, L., Wojtasik, K., Woźniak, S. & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*, **99**, 101861.

Lieder, F., Chen, P. Z., Stojcheski, J., Consul, S., & Pammer-Schindler, V. (2022). A cautionary tale about AI-generated goal suggestions. In *Proceedings of Mensch und Computer 2022* (pp. 354-359).

Louie, R., Coenen, A., Huang, C. Z., Terry, M., & Cai, C. J. (2020, April). Novice-AI music co-creation via AI-steering tools for deep generative models. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-13).

Methling, F., Borden, S. A., Veeraraghavan, D., Sommer, I., Siebert, J. U., von Nitzsch, R., & Seidler, M. (2022). Supporting innovation in early-stage pharmaceutical development decisions. *Decision Analysis*, **19**(4), 337-353.

Moulaei, K., Yadegari, A., Baharestani, M., Farzanbakhsh, S., Sabet, B., & Afrash, M. R. (2024). Generative artificial intelligence in healthcare: A scoping review on benefits, challenges and applications. *International Journal of Medical Informatics*, 105474.

Paschen, J., Wilson, M., & Ferreira, J. J. (2020). Collaborative intelligence: How human and artificial intelligence create value along the B2B sales funnel. *Business Horizons*, *63*(3), 403-414.

Raman, R., Lathabai, H. H., Mandal, S., Das, P., Kaur, T., & Nedungadi, P. (2024). ChatGPT: Literate or intelligent about UN sustainable development goals?. *PLOS One*, **19**(4), e0297521.

Runge, M. C., & Walshe, T. (2014). Identifying objectives and alternative actions to frame a decision problem. In *Application of threshold concepts in natural resource decision making* (pp. 29-43). New York, NY: Springer New York.

Siebert, J., Von Winterfeldt, D., & John, R. S. (2016). Identifying and structuring the objectives of the Islamic State of Iraq and the Levant (ISIL) and its followers. *Decision Analysis*, **13**(1), 26-50.

Siebert, J. U., & Von Winterfeldt, D. (2020). Comparative analysis of terrorists' objectives hierarchies. *Decision Analysis*, **17**(2), 97-114.

Simon, J., Regnier, E., & Whitney, L. (2014). A value-focused approach to energy transformation in the United States Department of Defense. *Decision Analysis*, **11**(2), 117-132.

Spaniol, M. J., & Rowland, N. J. (2023). AI-assisted scenario generation for strategic planning. *Futures & Foresight Science*, **5**(2), e148.

Thaler, R. H., & Sunstein, C. R. (2009). Nudge: Improving decisions about health, wealth, and happiness. Penguin.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., ... & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, **11**(1), 233.

von Winterfeldt, D., & Edwards, W. (1993). Decision analysis and behavioral research.

von Winterfeldt, D. (1987). Value tree analysis: An introduction and an application to offshore oil drilling. Kleindorfer P, Kunreuther H, eds. *Insuring and Managing Hazardous Risks: From Seveso to Bhopal and Beyond* (Springer, New York), 349–377.

Waterfield, D. A., Coleman, O. F., Welker, N. P., Kennedy, M. J., McDonald, S. D., & Cook, B. G. (2025). IEPs in the age of AI: Examining IEP goals written with and without ChatGPT. *Journal of Special Education Technology*, 01626434251324592.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, **35**, 24824-24837.

Wong, I. A., Lian, Q. L., & Sun, D. (2023). Autonomous travel decision-making: An early glimpse into ChatGPT and generative AI. *Journal of Hospitality and Tourism Management*, **56**, 253-263.

Jay Simon is an associate professor of information technology and analytics in the Kogod School of Business at American University. His background is in decision analysis. His research focuses primarily on modeling multiattribute preferences in challenging settings where a simple elicitation and application of standard methods would be impossible or impractical.

Johannes Ulrich Siebert is a professor of decision sciences, behavioral economics, and SCM at Management Center Innsbruck, The Entrepreneurial School®. His research advances decision-making at individual, organizational, and interorganizational levels. He has published in leading journals, including *Operations Research*, advises public and private organizations in Europe and the United States, and is a three-time finalist for the prestigious Decision Analysis Society Practice Award of INFORMS.