Journal of Behavioral and Experimental Economics 121 (2026) 102525

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

A taxonomy of AI experiments☆,☆☆

Aleksandr Alekseev [a],*, Christina Strobel [b]

[a] Department of Economics, University of Regensburg, Universitätsstr. 31, 93040 Regensburg, Germany
[b] Institute for Digital Economics, Hamburg University of Technology, Am Schwarzenberg-Campus 4, 21073 Hamburg, Germany

ARTICLE INFO

JEL classification:
C9
D9
O33

Keywords:
Economic experiments
Experimental design
Automation
Artificial intelligence

ABSTRACT

We introduce a taxonomy of artificial intelligence (AI) experiments. Our taxonomy produces four types of AI experiments: conceptual AI experiments, stylized AI experiments, quasi-natural AI experiments, and natural AI experiments. At the core of our taxonomy is the sophistication of AI used, which we evaluate using a simple and robust proxy test of whether AI is developed exclusively for a research study. We discuss the advantages, disadvantages, and best use cases for each type and illustrate the use of each type in various examples. We provide a guide on how to choose the type of AI experiment that best fits a given research question.

1. Introduction

The number of experimental papers studying interactions between humans and artificial intelligence (AI) has been steadily rising in recent years (see Fig. 1). In fact, almost half of all experimental AI papers reviewed by Chugunova and Sele (2022) and March (2021) were published since 2013.[1] A number of special issues dedicated to AI research have appeared in various journals such as *The Journal of Behavioral and Experimental Economics*, *Experimental Economics*, *Management Science*, *Science*, *Decision Sciences*, and *Research Policy*. With the increasing supply of AI experiments comes a greater demand to organize the literature and make sense of the glut of findings. Recent review articles by Bao et al. (2022), Chugunova and Sele (2022), Langer and Landers (2021), March (2021), Jussupow et al. (2020), and Burton et al. (2020), among others, are valuable contributions aimed at meeting this demand. However, recent literature has also identified an unsatisfied need for greater methodological discipline in AI experiments (Langer et al., 2020; March, 2021). We take a first step by proposing a taxonomy of AI experiments and offering a guide for using AI as a tool in experimental economists' methodological arsenal.[2]

With the growing academic interest in AI experiments, it seems natural to try to define just what *is* an AI experiment. In search of this definition, however, we do not converge on a single ideal type that might be called *the* AI experiment. Rather, we advocate for an inclusive approach and suggest that AI experiments exist on a spectrum that spans quite diverse designs. What defines an AI experiment, in our view, is not so much the design or implementation of AI but rather an underlying research agenda. Therefore, in this paper, we adopt the following broad definition of AI experiments. These are experiments that study interactions between human subjects and computers, algorithms, artificial agents, machines, robots, automated agents, and artificial intelligence agents with the goal of better understanding how the (actual or hypothetical) interaction with, or the presence of, these agents affects human behavior and outcomes in organizations and markets.[3]

Having adopted this broad definition of AI experiments that allows for a variety of designs, we attempt to put some structure on this variety. We identify and label four types of AI experiments: conceptual AI experiments, stylized AI experiments, quasi-natural AI experiments, and natural AI experiments. We discuss the advantages, disadvantages,

☆ This article is part of a Special issue entitled: 'AI/ML in Behavioural Experiments' published in Journal of Behavioral and Experimental Economics.

☆☆ We thank the Editor (Oliver Kirchkamp) and anonymous reviewers whose detailed suggestions helped significantly improve the quality of the paper. All remaining errors are our own.

* Corresponding author.
E-mail addresses: aleksandr.alekseev@ur.de (A. Alekseev), christina.strobel@tuhh.de (C. Strobel).

[1] Despite the recent surge in popularity, some of the ideas in current experimental AI research can be traced back as far as the 1950s (Meehl, 1954).

[2] Although the focus of the present paper is on economics experiments, the majority of our discussion applies to other disciplines that conduct AI experiments, including psychology, management, and computer science.

[3] Since the focus of our paper is social sciences, our definition excludes experiments that study exclusively the interactions between AIs. Including them would have strayed us too far into the realm of computer science and AI research. We include robots in this definition and show examples of their use in experiments because robots can be viewed as a physical manifestation of AI (Russell & Norvig, 2021), with the caveat that not all robots are powered by AI.

https://doi.org/10.1016/j.socec.2026.102525
Received 27 December 2024; Received in revised form 13 January 2026; Accepted 21 January 2026
Available online 22 January 2026

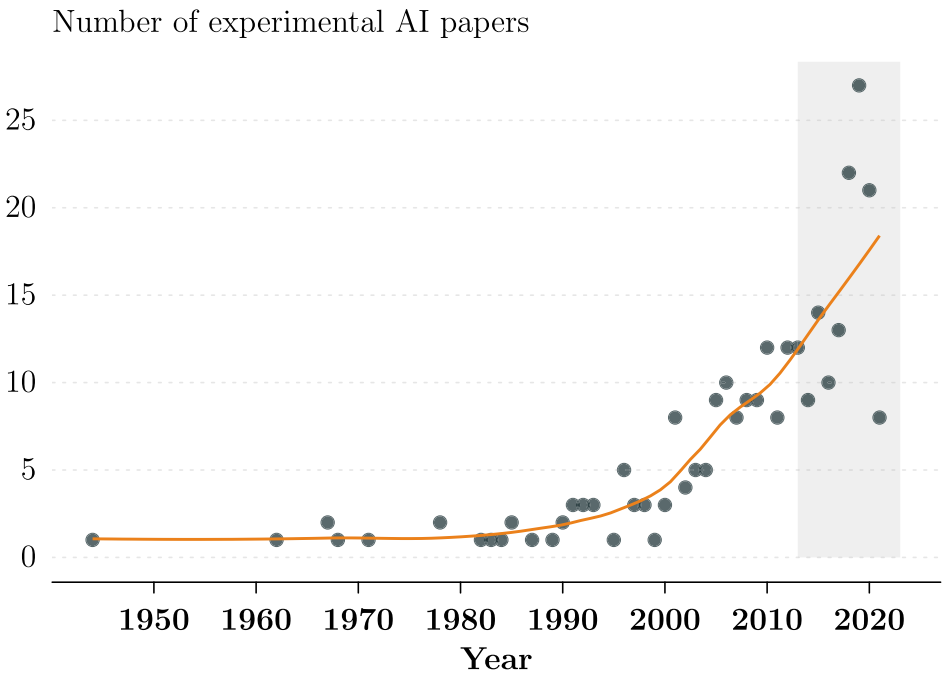

Fig. 1. The growth of experimental AI literature.
Note: The figure shows the total number of experimental AI papers published in a given year. The line shows a loess fit. The shaded region shows the time period during which approximately half of all experimental papers were published. We compiled our list based on the papers reviewed in Chugunova and Sele (2022) and March (2021).

and best use cases for each type and illustrate their usage in various examples.

At the core of our taxonomy is the sophistication of AI used. The sophistication of AI varies from real-world, commercial-grade systems, such as ChatGPT, to AI-as-label used in vignette studies, and everything in between. Our taxonomy, however, does not rely on AI sophistication directly, because that judgment is subjective and requires technical detail impractical for classification purposes. Instead, we propose a simple and robust proxy test: whether AI is developed exclusively for the purpose of conducting a research study. The logic behind this proxy test is that real-world, commercial-grade AI systems are unlikely to be developed for the purpose of simply conducting a study. On the other hand, AI developed specifically for an experiment is unlikely to match the sophistication of real-world systems.[4]

We gauge the levels of AI in an iterative manner. At a first step, we ask whether AI is just a label, which isolates *conceptual* experiments. If the AI is not simply a label, we then ask whether it is implemented exclusively for a study, which isolates *stylized* AI experiments. If AI is not implemented exclusively for a study, we ask if the experiment is conducted in a controlled environment, which isolates *quasi-natural* AI experiments. Finally, if AI is not implemented exclusively for a study and is used in its natural setting, then we identify it as a *natural* AI experiment.

Our taxonomy serves the dual purpose of organizing the existing literature and offering a tool to help researchers design new AI experiments. The organizing role of our taxonomy is particularly valuable for such a fragmented topic as human–AI interaction, which is spread across different fields. Grouping AI experiments into a few well-defined categories helps unify the literature (e.g., by clarifying how different lines of research are related to each other) and reveal underexplored questions (e.g., a certain phenomenon might have been established only for one type of AI experiments while its robustness to other types is unknown). Ultimately, this will help researchers better place their work in relation to existing studies across multiple fields, as well as identify potential directions for future research.

Our taxonomy also offers a practical guide for designing new AI experiments. By highlighting the main advantages, disadvantages, and best use cases of each type of AI experiment, our taxonomy provides researchers with a well-structured menu of alternative designs. This can help researchers be deliberate about which type (types) to use, rather than defaulting to conventional practices. Making the menu of available designs explicit can also inspire new combinations of designs or help identify underutilized designs (e.g., an existing convention might dictate the use of quasi-natural AI experiments, while a stylized design would suffice). Each study should ultimately settle for a type (or combination of types) that best fits the research question at hand.[5] It is possible that no single type best fits a given research question, in which case we recommend a complementary approach. Just like many experimental studies combine different samples (e.g., lab and online) (Hergueux & Jacquemet, 2015/06/01; Palan & Schitter, 2018) or types of experiments (e.g., lab and field) (Harrison & List, 2004) that complement each other, studies of human–AI interaction can exploit complementarities offered by different types of AI experiments. For example, a study can first establish a result in a specific setting using a quasi-natural AI experiment that offers greater external validity at the cost of a narrower scope. Then it can explore the mechanisms behind this result in a stylized AI experiment that offers a wider scope at the cost of lower external validity. In any case, we argue that researchers should justify their design choices, and we provide guidance for such justification.

[4] It is possible that our proxy test can misclassify some cases. For example, a research team conducting an experiment might develop an algorithm that later becomes a commercial-grade system in practice. We believe, however, that the ease of use and robustness of our proxy test outweigh the dangers of potential misclassification that could be caught with a more elaborate test.

[5] It should be noted, however, that while our focus is on experimental studies, some research questions are naturally more amenable to non-experimental methods. We also acknowledge the value of survey-based approaches for studying the use of AI, such as Carvajal et al. (2024) and Chugunova et al. (2025).

Our research builds upon and contributes to the recent reviews and methodological discussions of experimental AI papers. The paper that is closest in spirit to the present one is March (2021) that focuses on the use of computer players in experimental games spanning a range of topics including auctions, bargaining, and social dilemmas. It not only reviews the results from this vast literature but also provides a useful classification of the reasons for using a computer player in a game (e.g., reducing decision-making noise or inducing certain behavioral types) and the types of algorithms used in those games (e.g., equilibrium or adaptive algorithms). Our paper echoes the implicit message in March (2021) that the type of AI used should be based on the research question and addresses the concern about the lack of methodological standardization and guidance in AI experiments.

Bao et al. (2022) offer a more targeted discussion of AI in strategic interactions that occur in experimental financial markets. Their review provides researchers with a detailed classification of the algorithms used in finance experiments and how those algorithms affect participants' behavior and market outcomes. Consistent with our argument that the type of AI experiments should be based on a research question, Bao et al. (2022) highlight the importance of both finance experiments in which the algorithm is actually implemented (our taxonomy would classify the majority of them as stylized AI experiments) and experiments in which the presence of an algorithm is merely announced (these would fall into our conceptual AI experiments category).

Likewise, a review by Chugunova and Sele (2022) embraces methodological diversity. The study stands out due to its breadth and organizes findings from experiments where human subjects interact with automated agents across a wide range of disciplines spanning economics, psychology, sociology, marketing, medicine, and others. While the authors do not make methodological points explicitly, they do note that in many instances the details of AI implementation matter less for subjects' behavior than the mere notion of interacting with AI instead of humans.

Burton et al. (2020) and Jussupow et al. (2020) summarize the findings of the literature on algorithm aversion. Although both papers offer illuminating insights into the reasons behind algorithm aversion and potential ways to overcome it, they remain largely silent about the methodological issues in the reviewed studies. Jussupow et al. (2020) does comment, however, on the predominance of vignette studies, which fall under conceptual AI experiments according to our taxonomy. Perhaps this predominance is what makes the authors refer to AI in a study like Yeomans et al. (2019) as a "real working algorithm," even though our classification puts it into a stylized, rather than a natural or quasi-natural, category.

The methodological concern about the predominance of vignette studies in AI experiments finds a stronger voice in Langer and Landers (2021). This review is unique because it focuses on people affected by AI who do not interact with it (second parties) and on outside observers (third parties), rather than on people who directly interact with it (first parties). Langer and Landers (2021) lament the over-reliance on vignette studies, which have the downside of lower external validity. Our paper acknowledges this important shortcoming of conceptual AI experiments, while also highlighting their benefits. Importantly, we do not take a stance on what type of AI experiments is "best" but instead argue that the type of AI experiment should fit the research question.

Our study is also related to the literature that explores the broader methodological role of AI in scientific research and experimentation. Charness et al. (2025/03/31) offers a related yet distinct perspective by focusing on the application of generative AI, in particular large language models (LLMs), as a research tool. The authors argue that LLMs can enhance experimental research by improving comprehension, immersion, data collection, and analysis. The review also addresses broader risks and benefits, providing guidance on how generative AI might support open science and enable scalable experimentation in policy and business contexts. In a similar vein, Korinek (2023) investigates the potential of generative AI, such as ChatGPT, to support economists. The study identifies six key areas where generative AI can be beneficial: ideation, writing, background research, data analysis, coding, and mathematical derivations. While these reviews offer practical guidance and examples showcasing the potential of LLMs throughout the research process, our study focuses on how AI is implemented within experiments and explores the diverse approaches for designing AI experiments.

Our paper offers three main contributions to the literature. First, we propose a taxonomy of AI experiments and a simple and robust test to classify studies according to this taxonomy. Second, we discuss the advantages and disadvantages of each type of AI experiments that we identify and compare the types based on these features. Third, we discuss the best use cases for each type along with relevant examples from the literature. Our paper, however, is not a substitute for existing taxonomies, e.g., March (2021) or Bao et al. (2022). Rather, we complement these by providing researchers with a high-level classification of AI experiments and guidance on how to choose the right type to fit their research questions, after which researchers should tailor their designs using narrower taxonomies developed for their topics. We also note that our paper is not a literature review. Our examples of studies using AI experiments are not meant to be exhaustive. Instead, they are meant to illustrate why choosing a particular type of AI experiment makes sense in the context of these studies.

2. A taxonomy of AI experiments

2.1. Classification procedure

We identify and label four types of AI experiments: *conceptual AI experiments*, *stylized AI experiments*, *quasi-natural AI experiments*, and *natural AI experiments*. Fig. 2 shows the decision tree that we use for classification. Conceptual AI experiments, unlike the three other types, are characterized by AI that exists merely as a label or a framing device. In stylized AI experiments, the AI is implemented but is designed specifically for a study. In both quasi-natural and natural AI experiments, the AI used in a study is designed for purposes other than conducting a study. What distinguishes these two types is the environment in which the use of AI occurs. In quasi-natural AI experiments, the use of AI occurs in a controlled environment. Natural AI experiments, on the other hand, take place in an environment in which the use of that AI naturally occurs.[6]

Although we present these types in a certain order, we stress that the order does not reflect their quality. While it might be tempting to claim that one type of AI experiments is superior to another, we argue that this is not the case. Each type of AI experiments has its advantages and disadvantages, and there is no single type that is suitable for all studies. We now proceed to define each type, discussing its advantages, disadvantages, and best use cases. Table 1 offers an overview of each type.

[6] We use the following rationale for our terminology. Natural AI experiments, by analogy with the natural field experiments, occur in an environment that is "natural" for the AI used in them, hence the term. Quasi-natural AI experiments are "almost like" the natural AI experiments since both use the same type of AI, hence the term "quasi." The only difference between the two is the environment in which experiments occur, which we view as minor albeit important. We use the term "stylized" in its dictionary definition sense ("depicted or treated in a mannered and nonrealistic style"), and similarly to "stylized facts." "Stylized" here refers to AI that is deliberately nonrealistic and removed from many real-world implementation details in favor of focusing only on the essential elements. Finally, "conceptual" refers to AI as a concept or an idea rather than an actually implemented device. We do not call conceptual AI experiments "quasi-stylized" because in our view the difference between the conceptual and stylized AI experiments is more substantial than the difference between natural and quasi-natural AI experiments. Conceptual AI experiments have no implementation of AI whatsoever, while stylized AI experiments do.

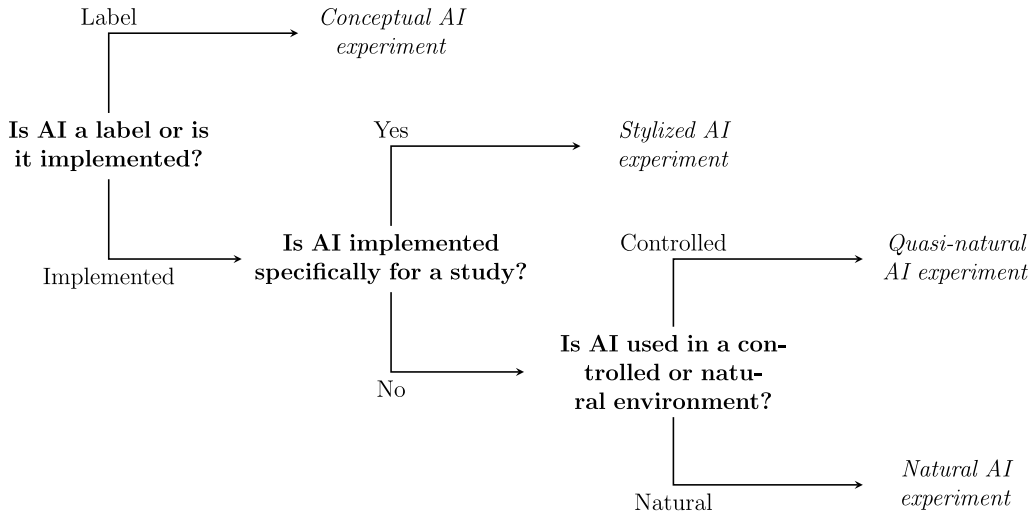

Fig. 2. Decision tree for classification.

Table 1
Overview of the four types of AI experiments.

Type	Naturalness	Control	Feasibility	Scope	Use if
Conceptual	★	NA	★★★★	★★★★	An algorithm cannot be implemented · ease of implementation/scalability and generality of research questions matter much more than naturalness of AI and/or setting
Stylized	★★	★★★	★★★	★★★	Implementation of AI is required · tight control over the algorithm, high feasibility and generality of research questions matter more than naturalness
Quasi-natural	★★★	★★	★★	★★	Implementation of AI is required · naturalistic implementation of AI and ability to relate to applied questions matter more than tight control or ease of implementation
Natural	★★★★	★	★	★	Implementation of AI is required · naturalistic implementation of AI/setting in which it is deployed and ability to relate to applied questions matter much more than control over the algorithm and ease of implementation

The table compares the following features of AI experiments. *Naturalness* (List, 2020) refers to the degree to which an algorithm and/or the setting in which it is deployed are similar to those found outside the context of a study, with lower values meaning "dissimilar" and higher values meaning "similar." *Control* refers to the degree to which researchers have control over the features of an algorithm, with lower values meaning "little control" and higher values meaning "a lot of control." *Feasibility* refers to the ease of conducting, scaling, or replicating a typical study, with low values meaning "easy" and low values meaning "hard." *Scope* refers to the breadth of research questions a typical study is suited for, with low values meaning "narrow or applied" and high values meaning "wide or general." The rankings represent our a priori expectations about a typical study, however, individual studies might deviate from these patterns.

2.2. Conceptual AI experiments

In conceptual AI experiments, AI exists merely as a label or framing device that models the operational principles or consequences of AI, but no AI is actually implemented. These experiments typically take place in a controlled, rather than natural, environment. Not being constrained by implementation allows researchers to study scenarios that are impractical or impossible to test with actual AI, such as ethical dilemmas or changes in a labor market.

A primary advantage of conceptual AI experiments is their high feasibility. They are relatively inexpensive, and researchers can conduct them at scale and replicate them easily. The low cost of these experiments allows researchers to quickly explore mechanisms underlying subjects' behavior or test the generalizability of their findings across different subject pools. The downside of conceptual AI experiments is that the AI and/or choice setting presented to subjects are dissimilar to the ones the subjects typically experience outside a study. Even when choice contexts are similar (e.g., consumer choice), subjects do not interact with actual AI and their choices are inconsequential. Conceptual AI experiments are best suited to studies that do not require AI implementation (e.g., due to ethical concerns) and prioritize ease of implementation, scalability, and generality over the naturalness of AI and/or its setting.

2.3. Stylized AI experiments

Stylized AI experiments actually implement AI, but it is designed specifically for a study. They are conducted in a controlled environment because the AI used in them typically does not occur outside the study. AI used in these experiments is based on a research question. It can thus take a variety of forms: from rule-based algorithms to algorithms that replicate decisions based on historical data to artificial agents trained using reinforcement learning.

The major strength of stylized AI experiments is a tight control over the features of an algorithm. This allows researchers to study a broad range of questions about human–AI interaction in settings with real stakes and to isolate specific behavioral mechanisms. Another benefit of these experiments is that they are feasible (i.e., they can be conducted with standard lab or online samples) and replicable. The downside of stylized AI experiments is lower naturalness: the AI and/or its setting

often differ from what subjects encounter outside a study. Stylized AI experiments are best suited to studies that require an actual AI implementation and prioritize control, feasibility, and generality over the naturalness of AI and/or its setting.

2.4. Quasi-natural AI experiments

Quasi-natural AI experiments feature naturalistic AI that is designed for purposes other than a research study, unlike in stylized AI experiments. They occur in a controlled environment—a feature they share with conceptual and stylized AI experiments. For example, researchers can study how subjects interact with state-of-the-art chatbots, such as ChatGPT, or commercial-grade robots in a controlled setting.[7] Quasi-natural AI experiments can also occur when an organization runs pilot experiments before launching a product, e.g., to test which features of AI lead to better customer satisfaction.

Quasi-natural AI experiments retain the naturalness of natural AI experiments while being more feasible, affordable, and replicable. They can be used to study a broad set of research questions because these experiments are not tied to the natural environments in which the use of AI occurs. A controlled setting enables collecting data on variables that might be impractical or impossible to elicit in natural settings, e.g., emotions or physiological responses. It also enables presenting scenarios that do not occur in natural environments. Their affordability also makes quasi-natural AI experiments a convenient test bed before scaling up a study. The reliance on actual AI, however, can be a limitation, since researchers give up a certain degree of control (although not all) over how the algorithm is constructed. Quasi-natural AI experiments are best suited to studies that require an actual AI implementation and prioritize naturalness and ability to relate to applied questions over control and ease of implementation.

2.5. Natural AI experiments

Natural AI experiments feature sophisticated AI (e.g., computer vision, natural-language processing, decision support, recommender systems, and robots) in the environments where it is actually used. The AI is usually trained on large datasets using machine-learning methods. The defining characteristic of natural AI experiments is that the AI involved is developed for purposes beyond the scope of the research study itself. Natural AI experiments are also natural field experiments (Harrison & List, 2004).

Many natural AI experiments are *A/B tests* (Azevedo et al., 2020) run by tech companies (e.g., Microsoft, Google, Amazon, Spotify, and Netflix) to increase engagement or develop product innovations. In these experiments, AI is not designed for the purpose of conducting a research study but, for example, for the purpose of giving better recommendations to the users of a service or improving the search results on a search engine. These experiments occur on a platform itself during the actual use of the services by its users.

The biggest strength of natural AI experiments is that they possess the highest possible degree of naturalness of AI and the setting in which it is deployed. These experiments produce findings that can be directly applied to an organization. The flip side of that naturalness and the applied nature of research questions is that natural AI experiments can often be too narrow and hard to generalize outside the context of a study. Tech companies run thousands of (often automated) A/B tests, but only a fraction of them produces generalizable knowledge about human–AI interaction. Another downside of natural AI experiments is low feasibility and replicability. Natural AI experiments are rarely public. Even when they are, replication is challenging because it would require access to proprietary platforms. Natural AI experiments are best suited to studies that require an actual AI implementation and prioritize the naturalness of AI and its setting together with the ability to relate to applied questions over control and ease of implementation.

3. The taxonomy in action

In this section, we put our taxonomy into action and illustrate the use cases of each type through various examples. Table 2 provides a summary of the reviewed papers.

3.1. Conceptual AI experiments

The most common application for conceptual AI experiments is vignette studies. A typical vignette study presents subjects with a series of hypothetical situations and asks them to state their preferences about what they would do in those situations. The primary advantage of vignette studies is that they can model any situation of interest to researchers without being constrained by the actual implementation of AI.

For example, Awad et al. (2018/11/01) conduct an online experiment that elicits subjects' preferences for what a self-driving car should do in moral dilemmas (trolley problems). A typical dilemma involves a hypothetical situation with a malfunctioning self-driving car that can either stay on course and kill pedestrians or swerve and kill the passengers. The dilemmas present subjects with various trade-offs such as between humans vs. animals, more lives vs. fewer lives, and young vs. old, and each subject makes choices in 13 different situations. Using vignettes allows the authors to collect an enormous dataset of choices spanning a wide variety of situations and cultures.[8] Moreover, a conceptual AI experiment is the most appropriate choice for this research question. Running a natural AI experiment would have been infeasible: One could never run an experiment where actual self-driving cars implemented subjects' choices and killed people. One could attempt to make subjects' choices consequential by promising to use these choices to inform the programming of actual self-driving cars in the future. The legal or ethical status of such a promise, however, would be unclear. A more feasible alternative would be a stylized AI experiment in the spirit of the "mouse-model" design of Falk et al. (2020). Researchers could program simple devices that would implement subjects' choices that would be consequential for mice. Aside from the ethical issues it might raise, the downside of this implementation is that mice do not possess many of the characteristics of interest of potential victims in Awad et al. (2018/11/01).

AI vignettes often present subjects with the option to choose between an AI and a human. For example, Lee (2018) presents subjects with different managerial decisions and asks them about their perceptions of those decisions when implemented by either AI or a human. Castelo et al. (2019) offer subjects a variety of tasks and ask them whether they prefer an AI or a human to complete those tasks. Similarly, Granulo et al. (2021) ask subjects whether they prefer an AI or another human to replace a human worker. In addition to exploring a variety of choice situations, these studies highlight two other key advantages of conceptual AI experiments. First, these experiments allow researchers to quickly explore the mechanisms underlying subjects' behavior, for example, whether a decision requires mechanical or human skills (Lee, 2018), whether a task is objective or subjective (Castelo et al., 2019), and whether the replaced worker is the subject herself or a third party (Granulo et al., 2021). Second, researchers can quickly

[7] To provide a counter-example, if the developer of ChatGPT, OpenAI, conducts an experiment on the website that people use to access ChatGPT, that would count as a natural AI experiment.

[8] It might appear that this example illustrates the low feasibility of conceptual experiments. However, high feasibility is precisely what enabled researchers to conduct such a large-scale study. Conducting a natural AI experiment of such scale, apart from ethical issues, would have been much less feasible.

Table 2
Summary of reviewed papers.

Classification	Experiment	Subfield	AI implementation
Conceptual AI experiments	Awad et al. (2018/11/01)	Social Preferences	Vignettes
	Lee (2018)	Technology Acceptance	Vignettes
	Castelo et al. (2019)	Technology Acceptance	Vignettes
	Granulo et al. (2021)	Technology Acceptance	Vignettes
	Wu (2022)	Behavioral Public Policy	Vignettes
	Zhang (2022)	Behavioral Public Policy	Vignettes
	Gallego et al. (2022)	Behavioral Public Policy	Vignettes
	Jeffrey (2021)	Behavioral Public Policy	Vignettes
	Farjam and Kirchkamp (2018)	Financial Markets	Rule-based
	Jacob Leal and Hanaki (2023)	Financial Markets	Rule-based
Stylized AI experiments	Strobel (2025)	Worker Performance	Timing
	Alekseev (2025)	Worker Performance	Rule-based
	Angerer et al. (2023)	Financial Markets	Rule-based
	Kirchkamp and Strobel (2019)	Social Preferences	Historical averages
	Corgnet et al. (2023)	Worker Performance	Historical averages
	Gogoll and Uhl (2018)	Social Preferences	Rule-based
	Dietvorst et al. (2015)	Technology Acceptance	Regression
	Dargnies et al. (2024)	Social Preferences	Regression
	Klockmann et al. (2022)	Social Preferences	Supervised learning
	Werner (2021)	Markets and Competition	Reinforcement-learning
	Schauer and Schnurr (2023)	Markets and Competition	Reinforcement-learning
Quasi-Natural AI experiments	Cominelli et al. (2021)	Social Preferences	Image recognition
	Gorny et al. (2023)	Social Preferences	Robot
	Leib et al. (2023)	Social Preferences	GPT-J
	Dell'Acqua et al. (2023)	Worker Performance	GPT-4
Natural AI experiments	Luo et al. (2019)	Worker Performance	Chatbot
	Brynjolfsson et al. (2025)	Worker Performance	GPT-3
	Paravisini and Schoar (2013)	Financial Decision-Making	Scoring model
	Bundorf et al. (2019)	Financial Decision-Making	Scoring model
Edge cases	Dell'Acqua (2022)	Worker Performance	Regression
	Cox et al. (2016)	Worker Performance	Probit model
	Bai et al. (2022)	Worker Performance	Rule-based

test the replicability and generalizability of their results across different subject pools. Unlike the (Awad et al., 2018/11/01) study, the studies cited above could be feasibly run as natural, quasi-natural, or stylized AI experiments. Some scenarios, such as the work assignment scenario in Lee (2018), have even been implemented as natural AI experiments (Bai et al., 2022). However, even the studies that use natural AI implementations, such as that by Bai et al. (2022), acknowledge the complementary value of conceptual AI experiments.

Another useful application for conceptual AI experiments is the manipulation of subjects' beliefs through priming. For example, the political science experiments by Wu (2022), Zhang (2022), Gallego et al. (2022), and Jeffrey (2021) use priming to evaluate the effects of subjects' beliefs about their automation exposure on their tendency to support various public policies. In a typical priming experiment, subjects first read a passage containing some information about new technologies, which is supposed to affect their beliefs, or some neutral information, which serves as a control condition, and then have to state their support for given policies. The conceptual AI experiments are ideal for these studies because researchers are typically interested in the general effects of automation. Running a natural AI experiment in this context would be challenging because it is difficult to experimentally manipulate the automation exposure for actual workers.[9] Running a stylized or quasi-natural AI experiment would be an option, and in fact, experiments that elicit distributional preferences in controlled environments are not uncommon in economics (Cappelen et al., 2023).

Conceptual AI experiments are not limited to vignette studies. For example, finance experiments by Farjam and Kirchkamp (2018) and Jacob Leal and Hanaki (2023) leverage the possibility of interacting with AI to study actual changes in subjects' behavior. In these experiments, subjects participate in stylized financial markets in which they can buy and sell assets for several rounds. The treatment condition informs subjects that the markets in which they participate may include algorithmic traders. The control condition does not give subjects this information. Importantly, no algorithmic traders exist in either condition.[10] These studies employ conceptual AI experiments mainly to disentangle the mechanisms underlying subjects' behavior. The presence of algorithmic traders in a market can affect the behavior of human traders and the resulting market outcomes through two distinct channels. The first channel is mechanical: The algorithms implement certain strategies and trade faster, which directly influences market outcomes. The second channel is behavioral: The beliefs about the presence of algorithms and what strategies they employ may alter the behavior of human traders independently of what the algorithms actually do. By focusing on the possibility of interacting with algorithmic traders, the studies cited above are able to isolate the behavioral channel. None of the other types of AI experiments would have been suitable for achieving this research goal because the actual implementation of algorithmic traders would have confounded the two channels.

3.2. Stylized AI experiments

Unlike conceptual AI experiments, stylized AI experiments actually implement AI in some form. The ability to design their own algorithms presents researchers with a variety of implementation options. One of the simplest options is a rule-based algorithm. Such algorithms are

[9] Although experimental methods allow for a more precise way of manipulating and measuring beliefs, in this case a more feasible alternative to a natural AI experiment could be an observational study that uses an exogenous shock as an instrument for automation exposure (Anelli et al., 2019; Webb, 2019).

[10] Unlike some AI experiments in computer science and psychology, economic and finance experiments do not use deception. Even though the studies cited here do not focus on the treatments that feature interactions with algorithmic traders, there was an actual possibility of interacting with algorithms in each study.

easy to explain to subjects and offer a high degree of control over the performance of the algorithm.

For example, Strobel (2025) studies whether automated bonus evaluation affects worker performance using a modified one-shot principal–agent game. In the game, subjects in the role of workers choose their performance levels. Subjects in the role of principals set performance thresholds for assigning bonuses. The performance threshold can either be set before the performance is known, which models the mechanism of an automated bonus evaluation process, or after the performance is known, which models the mechanism of a non-automated process. The automated process is implemented by simply comparing the predetermined performance threshold with the actual performance and assigning a bonus if the threshold is exceeded, and subjects are aware of each process. The experiment varies whether the bonus evaluation process is determined by a principal or randomly. The results show that performance is significantly lower under the automated process. However, whether automation is determined by a principal or randomly has no significant effect on performance. The study argues that lower performance under the automated process is not driven by fairness or trust concerns, but rather by misaligned expectations about how generous the threshold should be.

Alekseev (2025) is another example of a rule-based algorithm in a labor setting. It studies preferences for working with an algorithm that induces a task-switching environment. In the experiment, subjects perform real-effort tasks and make a choice of whether to work manually and complete the tasks themselves or work with an algorithm and delegate some of the tasks to it. Delegating to the algorithm enables subjects to work on new tasks, which is always better than completing all the tasks themselves in terms of monetary payoffs. The algorithm, however, is programmed to periodically interrupt subjects' work and ask for help. Subjects are aware of the interruption rule — the algorithm interrupts only when it encounters certain tasks, the frequency of which is identical whether subjects work by themselves or with an algorithm — but not of the actual interruption frequency, which varies between subjects. The study finds that as the frequency of interruptions increases, subjects are less likely to delegate to the algorithm, which suggests that task switching has tangible utility costs to the subjects.

Rule-based algorithms can be fairly sophisticated, especially when implemented in dynamic environments. For example, Angerer et al. (2023) study the effect of different arbitrage-seeking algorithms on the outcomes in experimental financial markets. In the experiment, human subjects buy and sell assets, across several periods, whose dividends are correlated between the two markets. The algorithms are programmed to seek arbitrage opportunities across the two markets and make buy or sell orders when such opportunities arise. The subjects know they may interact with a "computerized participant," however, they have no information about whether they actually interact with one, what its strategy is, or to whom its earnings accrue. The study finds that the presence of such algorithms moves the markets closer to the law of one price.

An alternative to rule-based algorithms are algorithms that are trained on past data. A common way to train such algorithms is to simply replicate the distribution of decisions from past human-only sessions. Automating the decisions of other players with such AI allows researchers to isolate the behavioral effects of mechanisms such as social preferences, intentionality, or peer pressure (March, 2021). By replacing a human player with an AI that plays like an average human while not possessing the relevant characteristics of a human, researchers can shut down the mechanisms of interest and see how it affects the behavior of human subjects.

An example of a stylized AI experiment that uses historical averages to automate decisions is Kirchkamp and Strobel (2019). The study investigates the role of the perceived responsibility and guilt of others on one's own responsibility, guilt, and selfish choices. In the experiment, human subjects play a one-shot dictator game in which the dictator's decision to split money equally or unequally is implemented by a pair of players. The pairs of players consist of either two human players or a human player and a passive human player whose decisions are automated by an algorithm. Human subjects in the latter case know that a computer automates the decisions of a passive player by replicating past choice frequencies but not the actual frequencies. The researchers expect human dictators to feel more responsible for an outcome, feel more guilt for an unequal split, and make fewer selfish choices when they are paired with an algorithm than with another human because an algorithm cannot be responsible and feel guilt in the same way a human does. The researchers, however, find no such effects.

Corgnet et al. (2023) is another example of using historical averages, this time in a labor setting, that examines the effects of social pressure on workers' performance. In the experiment, a team of three workers repeatedly performs, over five rounds, a sequential task mimicking an assembly line. The team consists of either only human workers or two human workers and one algorithm. The researchers calibrate the productivity of the algorithm to be the same as that of an average human worker. Subjects working with an algorithm are aware of neither the algorithm's actual productivity nor the calibration rule. They know that an algorithm exists and can observe its performance. The researchers isolate the social-pressure effect because the algorithm cannot impose social pressure on other workers in the same way a human can. The study finds that subjects who work with the algorithm underperform relative to those who work in human-only teams, which highlights the importance of social pressure for team performance.

Gogoll and Uhl (2018) is an example of using historical averages to train AI in an experiment at the intersection of labor and moral domains. It studies preferences for delegating to an algorithm tasks that affect third parties. In the experiment, subjects make a choice of whether to delegate a numerical task to another human subject or to an algorithm. The performance on the task, however, affects not the payoff of the subject who solves the task but that of another subject, adding a moral component to the delegation choice. The algorithm is programmed to reproduce the performance distribution of human subjects in previous sessions. As in Corgnet et al. (2023), subjects are aware neither of the actual performance of the algorithm nor of the calibration rule, however, they observe a snapshot of the algorithm's performance (along with that of human subjects) before they make their delegation decisions. The study finds that subjects are three times more likely to delegate the task to another human than to an algorithm, despite identical ex-ante performance. The researchers find that neither the perceived differences in performance between an algorithm and humans nor trust in the algorithm can explain the reluctance to delegate to an algorithm.

A more sophisticated approach for training AI on past data is to estimate a prediction model, such as an ordinary least squares (OLS) regression. This approach provides less experimental control over the performance of the algorithm, which is typically better than that of human subjects, and is more difficult to explain to subjects. However, the superior performance of such algorithms is often a desirable feature that they share with commercial-grade AI tools, which ultimately increases the external validity of a study.

For example, Dietvorst et al. (2015) study subjects' preferences for performing a forecasting task themselves or delegating the task to an algorithm. In the experiment, subjects have to predict over ten rounds how successful an MBA student would be using such variables as a student's undergraduate degree, GMAT scores, years of work experience, and education. The subjects can either make all predictions themselves or delegate all predictions to an algorithm. To build an algorithm, the researchers estimate an OLS regression on the data from 115 students using the same explanatory variables as the subjects themselves can use. The instructions give subjects general information about the algorithm ("The model is based on hundreds of past students, using the same categories of demographic data you are receiving") but not the actual implementation. Even though the algorithm is better than

humans at prediction, a significant fraction of subjects does not choose the algorithm. The study additionally proposes and explores a potential mechanism behind such algorithm aversion: Observing an algorithm perform may increase aversion to it. Consistent with this hypothesis, the fraction of subjects who choose an algorithm drops if the subjects observe the algorithm's performance.

Dargnies et al. (2024) use a similar approach to study the preferences of workers and managers for using an algorithm to evaluate workers and make hiring decisions. In the experiment, subjects in the role of workers perform real-effort tasks. They then make a choice of whether they prefer the hiring decision between themselves and another worker be made by another human subject in the role of a manager or by an algorithm. Subjects in the role of managers first make 20 hiring decisions from among pairs of workers and then make a choice for whether they want to delegate their hiring decisions to the algorithm. The managers can use workers' task performance and gender to make their hiring decisions. To build an algorithm, the researchers estimate an OLS regression on the data from 200 workers using the same explanatory variables that are available to managers (i.e., task performance and gender). The subjects know that the algorithm is designed to predict performance based on the data from previous workers, and that it hires the worker with the highest predicted performance, while the information about implementation details varies by treatment. The study finds that both the workers and the managers prefer human evaluation over algorithmic evaluation, despite the algorithm being better at picking the better-performing worker. However, when the algorithm does not use a worker's gender for prediction and workers know this, they choose the algorithm more often. The study additionally finds that explaining how the algorithm works does not increase either workers' or managers' preference for it.

Researchers are not limited to using simple prediction models, such as OLS. However, more sophisticated models, while offering better predictions, also require better explanations. For example, Klockmann et al. (2022) study how subjects' behavior is affected if they know that their choices train an algorithm that later makes a decision that has consequences either for them or for other subjects. In the experiment, subjects in the role of dictators first make repeated choices in a dictator game for 30 periods. The researchers then create an algorithm for each dictator to predict and make the choice in the final 31st period. To create the algorithm, the researchers train a random forest model using such features as the payoffs and the sum and difference of points allocated to a receiver in the dictator game. The instructions inform subjects that the dictators' choices are used to "train an artificially intelligent Random Forest algorithm," and give a summary of how the algorithm works. The prediction of the algorithm is implemented for either a receiver with whom a dictator was paired, a receiver in a different pair, or, with some probability, the dictators themselves. The study finds that the behavior of dictators does not differ between the cases when the algorithm makes a decision for a receiver in the dictator's own pair or for a receiver in a different pair. However, if there is a chance that the algorithm determines the payoff of the dictators, the dictators behave more prosocially: The share of egalitarian decisions increases.

Yet another step in the complexity of algorithms used in stylized AI experiments is artificial agents trained using reinforcement-learning. Such agents can be used to study complex decisions in dynamic strategic environments. A popular reinforcement-learning algorithm in experimental research is Q-learning. For example, Werner (2021) and Schauer and Schnurr (2023) study market outcomes and strategies in experimental oligopoly markets populated by either human participants, Q-learning agents, or both. In both studies, subjects are aware of whether they are playing against other humans or algorithms while the strategy of algorithms is not disclosed.[11] Werner (2021) additionally informs subjects that an algorithm "acts in the interest of another participant" who does not make any decisions but receives profits earned by the algorithm. The results show that both humans and algorithms learn to collude with each other, even without communication. However, collusion is highest when only human or only Q-learning agents are present in a market, while in hybrid markets humans and algorithms fail to coordinate.

In all the above examples, stylized AI experiments are the most fitting design choice for answering the posed research questions. This design allows researchers to create custom AIs that are tailored to their research questions and to easily evaluate different potential mechanisms behind subjects' choices. Conceptual AI experiments cannot capture the consequential nature of decisions often desired, e.g., for moral choices, or the dynamic choice environment of market experiments. Quasi-natural AI experiments would not have provided adequate control over the features of algorithms. Natural AI experiments, on the other hand, would be challenging to implement because these studies are typically interested in general patterns of human behavior. For some research questions, e.g., the ones that are motivated by labor settings that are of interest to organizations, stylized AI experiments provide a convenient first step for testing ideas before scaling them up to a level of natural AI experiments.

3.3. Quasi-natural AI experiments

Quasi-natural AI experiments, similar to stylized ones, actually implement AI in a controlled environment. The AI in them, however, is designed for purposes other than a research study. An example of a quasi-natural AI experiment is Cominelli et al. (2021) who employ a robot to study the effects of promises made by a robot or a human on trust in human counterparts in one-shot games. The researchers use the Facial Automaton for Conveying Emotions (FACE) robot designed for social robotics and, in particular, therapy for autism (Pioggia et al., 2004). The FACE robot has a human-like appearance and is capable of showing emotional states, empathy, and nonverbal communication. In the experiment, a human subject enters a room containing either a robot, a professional actress, or a computer, each of which makes a verbal promise to take a cooperative action. After that, the subject decides whether to trust them. The study finds that receiving a promise from the robot increases trust in participants who perceive the robot as human-like, but not in those who do not perceive the robot as such. A similar pattern occurs when either an actress or a non-human-like computer makes a promise.

Gorny et al. (2023) conduct a lab experiment at a learning factory to study how the presence of robots in a team of human workers affects their prosociality towards each other and the valuation of their products. A learning factory is a real, albeit simplified, production system featuring state-of-the-art robotics where university employees and students receive hands-on training on production technologies. In the experiment, two human workers produce electronic motor components by operating two different production stations at the beginning and end of a three-station production line. The middle station is operated by either two robots or a station that performs the same steps but with the robots switched off and hidden. The study finds that the presence of robots increases sharing behavior among human workers, however, it does not change workers' valuation of the rewards they earn from production.

Leib et al. (2023) use a large language model to study how AI- versus human-generated advice affects dishonest behavior among human participants in a one-shot die-rolling game. The experiment employs the GPT-J model developed by EleutherAI in 2021 (Wang & Komatsuzaki, 2021). The researchers fine-tune the model using text advice generated by human participants. In the experiment, subjects observe either AI-generated advice, human-generated advice, or no advice and then decide whether to take a dishonest but self-serving action. The instructions explained to subjects the basics of what the language

[11] Schauer and Schnurr (2023) study has one treatment where subjects do not know the identity of their AI competitor, while the other four treatments reveal this information.

model is and how it was trained. The results show that advice that promotes dishonesty increases dishonest behavior, while advice that promotes honesty does not increase honest behavior. This pattern holds regardless of whether an AI or a human gave that advice and regardless of whether a subject knows the exact source of the advice.

Dell'Acqua et al. (2023) study the effect of access to a generative AI tool among highly skilled knowledge workers on their productivity and quality of work. The researchers conduct an experiment with the consultants at Boston Consulting Group, a leading consulting firm, who are randomly assigned to either use the AI tool with no guidance, the AI tool with prompt-engineering training, or no AI at all. The AI tool in the experiment is the GPT-4 model by OpenAI. In the experiment, subjects complete a series of stylized tasks, representative of typical consulting activities at the company, and are then scored on the quality of their responses. The tasks differ by whether they are within the "AI capability frontier" (tasks that AI can reliably perform) or not (tasks that AI cannot reliably perform). The study finds that for tasks inside the frontier, access to the AI tool increases the number of completed tasks by 12.2% and quality by over 40%. These effects are most pronounced among lower-performing subjects, suggesting that AI acts as a performance equalizer. On the other hand, for tasks outside the frontier, access to AI reduced accuracy by 19 percentage points.

In all these examples, a quasi-natural design allows researchers to achieve the right balance between naturalness, on the one hand, and generality of research questions and feasibility, on the other. It enables studying scenarios (e.g., trusting, lying, or sharing behavior) and elicit variables that would be difficult to study in a natural environment. Additionally, it may act as a convenient test bed before potentially scaling up to the natural level. A conceptual or a stylized design would have been too simplistic for that purpose, since researchers in these studies are interested in subjects' behavior in naturalistic environments featuring actual AI or robotic systems.

3.4. Natural AI experiments

Natural AI experiments are often conducted in collaboration with an organization that is interested in deploying a new technology on its platform. For example, Luo et al. (2019) conduct a field experiment with a large Chinese internet-based financial-services company. The researchers study the effects of using a chatbot on the outcomes of sales calls to the company's customers who are eligible for loan extensions. The company uses a sophisticated voice AI chatbot trained on the voice data of the best-performing human workers that can conduct natural-sounding conversations indistinguishable from human conversations. The study finds that the chatbot is as effective in making sales as the best human workers and four times more effective than inexperienced workers. However, revealing that the caller is a chatbot reduces sales by about 80%, a drop mostly driven by customers' biases against machines.

Brynjolfsson et al. (2025) presents another example of using AI in a customer-support setting. The authors collaborate with a Fortune 500 company that sells business-process software to study the effects of the deployment of an AI assistant among over 5000 customer support workers on their productivity and service quality. The AI assistant is based on the GPT-3 model by OpenAI and is designed to monitor conversations in real-time and provide suggested responses and links to technical documentation. The design of the AI assistant allows for worker discretion over whether to follow its recommendations. The study finds that access to the AI assistant leads to 15% more issues resolved per hour, the study's main measure of productivity. The authors also report treatment effect heterogeneity: less experienced and lower-skilled workers benefit the most, while skilled workers see marginal speed gains. The results suggest that generative AI functions by transferring tacit knowledge from high-performing agents to less experienced ones, effectively leveling up the baseline capability of the workforce.

Paravisini and Schoar (2013) is a further example of a natural AI experiment but in the banking sector. The researchers conduct a field experiment with a for-profit bank in Colombia to examine the impact of a credit-scoring model used to evaluate prospective borrowers on the behavior of loan officers. The study finds that the committees that observe credit scores spend more time per loan application and reach decisions more often than the committees that do not observe the scores. The percentage of non-decisions also drops by more than 40%, with the effect being concentrated in difficult-to-evaluate applications. Interestingly, the committees who observe the scores only after making interim decisions also increase their output, by 75%, even though observing the scores never changes their interim decisions, and the quality of decisions is similar to that of committees who observe the scores before making their decisions, which suggests an incentive effect of scores availability. These findings suggest a coordinating role of an algorithmic score: human workers who have access to the score work harder and reach decisions more often.

Although natural AI experiments can benefit an organization's bottom line, as the previous examples illustrate, sometimes they are used for helping customers or improving the overall efficiency of a system. A case in point is Bundorf et al. (2019) who conduct a field experiment in collaboration with the Palo Alto Medical Foundation, a large multi-specialty physician group in California. The researchers develop and evaluate an online decision-support tool designed to help older adults choose a drug insurance plan. The tool uses a proprietary scoring technology from a third-party provider and assigns an expert score to each plan, which is a combination of an estimated total cost of the plan and the plan's "star rating." The study finds that the subjects who had access to the tool were more likely to select the plans suggested by it, with a more pronounced result for the group that had access to the expert scores in addition to the list of plans ordered by those scores. These findings echo the ones in Paravisini and Schoar (2013) on the coordinating role of algorithmic scores.

In these examples, natural AI experiments are the most fitting design choice. While other types of AI experiments could have reached similar conclusions, only natural AI experiments achieve that in contexts that are directly applicable to organizations. The use of AI technologies is complex and deeply embedded in organizational structures — culture, employee dynamics, specific business goals — something that controlled settings might struggle to replicate. AI usage in organizational settings can also reveal unanticipated behavioral responses and heterogeneity that are unlikely to emerge in simulated environments. Finally, the ethical and compliance considerations that affect AI deployment in organizations are context-dependent and are difficult to replicate in controlled settings.

3.5. Edge cases

We chose the examples in the previous sections to highlight the clear-cut cases in each category. In this section, by contrast, we turn to the cases that may present a challenge for classification to stress-test the logic behind our taxonomy. We start with Dell'Acqua (2022) who studies how the quality of an algorithm affects workers' reliance on it and the resulting quality of decisions. The subjects in the study are freelance recruiters hired on an online platform to screen stylized job applications. The subjects' task is to choose whether to invite an applicant for an interview by trying to guess the applicant's math ability based on other characteristics, such as education and employment. The subjects are split into four treatment groups, depending on an algorithm they have access to. In the control group, subjects do not have access to an algorithm, while in the remaining three groups, subjects can either have an almost perfect algorithm, a "good" algorithm with 85% accuracy, or a "bad" algorithm with 75% accuracy. The algorithms are based on a related study (Cowgill et al., 2020) in which software engineers predict a person's math ability based on their education and employment characteristics. The study finds that

subjects who have access to higher-quality AI are less accurate and spend less effort than subjects who have access to lower-quality AI, which suggests that a higher-quality AI may not always be beneficial for human decision-making.

This study is a field experiment, involving subjects in their natural roles, i.e., recruiters. It might appear, therefore, that this experiment belongs to a natural AI category. Our taxonomy, however, places it in the stylized AI category. The key consideration here is the implementation of AI. The algorithms used in the study are developed specifically for the purpose of a (related) study and are not intended to be used in actual hiring decisions. Hence, the study does not fall into quasi-natural or natural AI categories, according to our definition. The study does implement an algorithm, hence, it is not a conceptual AI experiment either.

For another example, we consider Cox et al. (2016) who collaborate with a large U.S. hospital to study the uptake of a clinical decision-support system among resident physicians and fourth-year medical students. The decision-support system makes hospital discharge recommendations based on a probit model estimated on the data from the hospital's electronic medical records. In the experiment, subjects view patient charts from the database used to develop the decision-support system. They make discharge decisions based on either typical information or that information supplemented by a recommendation from the decision-support system. In the group that has access to the system, the default decision is either generated by the system or based on the current practices. The results show that the decision-support system is more effective if the default is generated by the system and subjects who override it have to explain their choice than if the information provided by the system is simply available.

It might appear that this study belongs to a conceptual AI category because subjects' choices are not consequential for actual patients. Subjects' choices only affect their own payoffs: they receive a monetary bonus for correct discharge decisions. Our taxonomy, however, places it in the quasi-natural AI category. The key consideration here, again, is the implementation of the algorithm. The study actually implements an algorithm — a decision-support system — hence, it is not a conceptual AI experiment according to our definition. The algorithm is developed to be ultimately used in an actual hospital and not for the purpose of conducting a study, hence, it is not a stylized AI experiment. Since researchers conduct a controlled experiment in artificial conditions, albeit with actual doctors, this is not a natural AI experiment. The researchers, in fact, justify their choice of a quasi-natural AI design over a natural one: "it is a practical and ethical requirement before application of the system on patient wards in hospitals" (Cox et al., 2016)[P. 2].

Finally, we consider the case of Bai et al. (2022) who collaborate with a large Chinese warehouse operator owned by Alibaba, the largest retailer in China, to study the effects of human versus computer task assignment on workers' fairness perceptions and productivity. In the experiment, warehouse workers receive lists of items they need to pick from the warehouse (pick lists) either from a human supervisor or from a computer terminal. A list of items for each worker to pick is simply selected at random from a pool of available pick lists, previously generated by a logistical algorithm, regardless of whether a human supervisor or a computer distributes the list. Although the underlying rule for generating a pick list is the same, the workers who receive pick lists from a computer perceive their assigned tasks as fairer than do workers who receive pick lists from a human supervisor. This effect results from a concern that a human supervisor might have been biased towards or against some workers. Greater fairness perception of the computer assignment translates into productivity gains of about 18%.

This study presents a challenge for classification because, while it is a field experiment, the algorithm for generating pick lists is a simple random assignment, there is no sophistication one would expect to see in a natural AI experiment. We think, however, that this study does belong to a natural AI category. The key consideration here is the purpose of designing an algorithm, even a simple one like random assignment: whether it is designed merely for the study or whether it will actually be implemented in a warehouse. The study is silent about that, however, given the high-stakes environment of a natural field experiment, it seems plausible that the algorithm is tested for the purpose of being deployed, perhaps with some modifications, in an actual work setting.

We conclude by noticing that our classification relies on how AI is implemented in a study, rather than on how a study is implemented, e.g., whether it is a field experiment or not. It follows that the same type of study, in principle, can be implemented using different AI experiments. Take, for example, correspondence/audit studies (Verhaeghe, 2022) in which researchers typically send fictional CVs of potential job applicants to recruiters. Researchers can implement a correspondence study as a conceptual AI experiment, e.g., if they vary whether an applicant has experience with generating prompts for LLMs. Alternatively, they can implement it as a natural AI experiment, if they vary whether they send job applications to recruiters who are known to use algorithms for screening candidates or not.

4. Discussion

We have presented the four types of AI experiments in a neutral manner, balancing their strengths with their weaknesses. In practice, however, these types of AI experiments vary in prevalence, as illustrated in Fig. 3. All four types experienced increased usage since the 1990s, however, after the early 2010s, their trajectories began to diverge. The popularity of conceptual AI experiments peaked around mid-2010 and has subsequently declined. Stylized and quasi-natural AI experiments initially decreased in popularity during the mid-2010s but have shown a resurgence in recent years. Natural AI experiments have enjoyed a steady growth throughout the observation period.

The strengths and weaknesses of the four types of AI experiments that we identify offer insights into possible reasons behind these trends. Conceptual AI experiments, while valuable for rapid generation of new insights, likely saw a decline as researchers sought to validate initial findings with experiments where AI is actually implemented. The relative ease of implementation, combined with their ability to model dynamic interactions beyond what simple vignettes allow, has likely contributed to the renewed interest in stylized and quasi-natural AI experiments. The increasing accessibility and capabilities of LLMs have likely further fueled interest in quasi-natural AI experiments (Charness et al., 2025/03/31). A consistent growth in natural AI experiments likely reflects the recent wave of AI adoption and automation within organizations (Agrawal et al., 2022).

Based on these trends, we anticipate continued growth in AI experiments that implement AI in some form. The proliferation of LLMs and the ability to fine-tune them will likely drive further adoption of quasi-natural AI experiments, enabling researchers to investigate interactions with sophisticated conversational agents (Leib et al., 2023). Ongoing adoption of AI tools and automation into workplaces will likely stimulate the use of natural AI experiments, e.g., to study the productivity effects of new technologies (Brynjolfsson et al., 2025; Dell'Acqua et al., 2023). Stylized AI experiments will likely still find place in future research agendas. These experiments enable researchers to explore underlying behavioral mechanisms by controlling how an algorithm is constructed and presented, which distinguishes them from the largely "black box" nature of LLMs. Moreover, while LLMs, at least currently, are best suited for research questions involving textual data and conversational interactions, stylized AI experiments can address a broader range of questions. It is conceivable, however, that the heightened interest in LLMs will inspire new research questions that can be answered with this tool (Charness et al., 2025/03/31; Korinek, 2023). Finally, conceptual AI experiments may assume a complementary role, supplementing the findings from the other types of AI experiments.

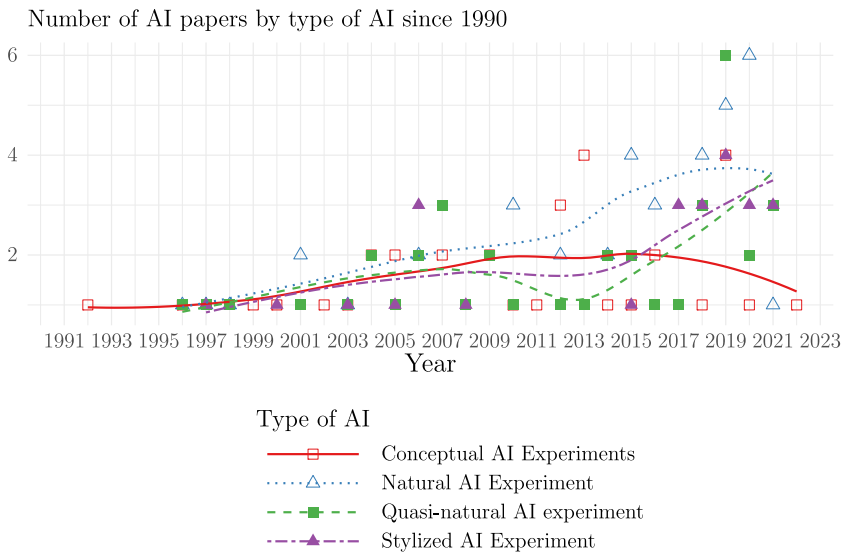

Fig. 3. Popularity of AI experiments by type.
Note: The figure shows the total number of experimental AI papers published in a given year since 1990, clustered by type. The line shows a loess fit. The list of papers is based on Chugunova and Sele (2022) and March (2021).

5. Conclusion

In this paper, we propose a taxonomy of AI experiments. Our taxonomy features four types of AI experiments: conceptual AI experiments, stylized AI experiments, quasi-natural AI experiments, and natural AI experiments. At the core of our taxonomy is the sophistication of AI used. To evaluate the sophistication, we propose a simple and robust proxy test of whether AI is developed exclusively for a research study. We provide a guide on the advantages, disadvantages, and best use cases for each type, illustrated via various examples.

Our taxonomy is designed to be easy-to-use and robust to the emergence of new technologies. However, there will be experiments that are difficult to classify using our procedure. We illustrate a few of such edge cases to stress-test our classification. We hope that our taxonomy will prove to be a useful tool for organizing the existing literature and will help researchers design new experiments.

CRediT authorship contribution statement

Aleksandr Alekseev: Writing – review & editing, Writing – original draft, Visualization, Methodology. **Christina Strobel:** Writing – review & editing, Writing – original draft, Visualization, Methodology.

References

Agrawal, A., Gans, J., & Goldfarb, A. (2022). ChatGPT and how AI disrupts industries. *Harvard Business Review*, *12*, 1–6.
Alekseev, A. (2025). The economics of babysitting a robot. Working paper. University of Regensburg.
Anelli, M., Colantone, I., & Stanig, P. (2019). We were the robots: Automation and voting behavior in western europe. CReAM Discussion Paper Series No. 17/19.
Angerer, M., Neugebauer, T., & Shachat, J. (2023). Arbitrage bots in experimental asset markets. *Journal of Economic Behavior and Organization*, *206*, 262–278.
Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018/11/01). The moral machine experiment. *Nature*, *563*(7729), 59–64.
Azevedo, E. M., Deng, A., Montiel Olea, J. L., Rao, J., & Weyl, E. G. (2020). A/b testing with fat tails. *Journal of Political Economy*, *128*(12), 4614.
Bai, B., Dai, H., Zhang, D. J., Zhang, F., & Hu, H. (2022). The impacts of algorithmic work assignment on fairness perceptions and productivity: Evidence from field experiments. *Manufacturing & Service Operations Management*, *24*(6), 3060–3078.
Bao, T., Nekrasova, E., Neugebauer, T., & Riyanto, Y. E. (2022). Chapter 23: Algorithmic trading in experimental markets with human traders: A literature survey. In S. Füllbrunn, & E. Haruvy (Eds.), *Handbook of experimental finance* (pp. 302–322). Cheltenham, UK: Edward Elgar Publishing.
Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, *140*(2), 889—942.
Bundorf, M. K., Polyakova, M., & Tai-Seale, M. (2019). How do humans interact with algorithms? Experimental evidence from health insurance. http://dx.doi.org/10.3386/w25976, Working paper series 25976.
Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239.
Cappelen, A. W., Cappelen, C., & Tungodden, B. (2023). Second-best fairness: The trade-off between false positives and false negatives. *American Economic Review*, *113*(9), 2458–2485.
Carvajal, D., Franco, C., & Isaksson, S. (2024). Will artificial intelligence get in the way of achieving gender equality? *SSRN Electronic Journal*.
Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825.
Charness, G., Jabarian, B., & List, J. A. (2025/03/31). The next generation of experimental research with LLMs. *Nature Human Behaviour*.
Chugunova, M., Harhoff, D., Hölzle, K., Kaschub, V., Malagimani, S., Morgalla, U., & Rose, R. (2025). Who uses AI in research, and for what? Large-scale survey evidence from Germany. *Research Paper No. 25-11 25–11*, Max Planck Institute for Innovation & Competition.
Chugunova, M., & Sele, D. (2022). We and it: An interdisciplinary review of the experimental evidence on how humans interact with machines. *Journal of Behavioral and Experimental Economics*, *99*, Article 101897.
Cominelli, L., Feri, F., Garofalo, R., Giannetti, C., Meléndez-Jiménez, M. A., Greco, A., Nardelli, M., Scilingo, E. P., & Kirchkamp, O. (2021). Promises and trust in human–robot interaction. *Scientific Reports*, *11*(1), 9687.
Corgnet, B., Hernán-González, R., & Mateo, R. (2023). Peer effects in an automated world. *Labour Economics*, *85*, Article 102455.
Cowgill, B., Dell'Acqua, F., Deng, S., Hsu, D., Verma, N., & Chaintreau, A. (2020). Biased programmers? Or biased data? A field experiment in operationalizing AI ethics. In *Proceedings of the 21st ACM conference on economics and computation* (pp. 679–681). New York, NY, USA: Association for Computing Machinery.
Cox, J. C., Sadiraj, V., Schnier, K. E., & Sweeney, J. F. (2016). Higher quality and lower cost from improving hospital discharge decision making. *Journal of Economic Behavior and Organization*, *131*, 1–16.
Dargnies, M.-P., Hakimov, R., & Kübler, D. (2024). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science*.
Dell'Acqua, F. (2022). Falling asleep at the wheel: Human/AI collaboration in a field experiment on HR recruiters. Working paper, Harvard Business School.
Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Working Paper No. 24-013. Harvard Business School.
Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126.
Falk, A., Neuber, T., & Szech, N. (2020). Diffusion of Being Pivotal and Immoral Outcomes. *Review of Economic Studies*, *87*(5), 2205–2229.

Farjam, M., & Kirchkamp, O. (2018). Bubbles in hybrid markets: How expectations about algorithmic trading affect human trading. *Journal of Economic Behavior and Organization*, *146*, 248–269.

Gallego, A., Kuo, A., Manzano, D., & Fernández-Albertos, J. (2022). Technological risk and policy preferences. *Comparative Political Studies*, *55*(1), 60–92.

Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, *74*, 97–103.

Gorny, P. M., Renner, B., & Schäfer, L. (2023). Prosocial behavior among human workers in robot-augmented production teams—A field-in-the-lab experiment. *Frontiers in Behavioral Economics*, *2*.

Granulo, A., Fuchs, C., & Puntoni, S. (2021). Preference for human (vs. robotic) labor is stronger in symbolic consumption contexts. *Journal of Consumer Psychology*, *31*(1), 72–80.

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, *42*(4), 1009–1055.

Hergueux, J., & Jacquemet, N. (2015/06/01). Social preferences in the online laboratory: a randomized experiment. *Experimental Economics*, *18*(2), 251–283.

Jacob Leal, S., & Hanaki, N. (2023). Algorithmic trading, what if it is just an illusion? Evidence from experimental asset markets. http://dx.doi.org/10.2139/ssrn.4620189, Working paper.

Jeffrey, K. (2021). Automation and the future of work: How rhetoric shapes the response in policy preferences. *Journal of Economic Behavior and Organization*, *192*, 417–433.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. *Research Paper No. 168*, AIS Electronic Library.

Kirchkamp, O., & Strobel, C. (2019). Sharing responsibility with a machine. *Journal of Behavioral and Experimental Economics*, *80*, 25–33.

Klockmann, V., von Schenk, A., & Villeval, M. C. (2022). Artificial intelligence, ethics, and intergenerational responsibility. *Journal of Economic Behavior and Organization*, *203*, 284–317.

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, *61*(4), 1281–1317.

Langer, M., König, C. J., & Hemsing, V. (2020). Is anybody listening? The impact of automatically evaluated job interviews on impression management and applicant reactions. *Journal of Managerial Psychology*, *35*(4), 271–284.

Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior*, *123*, Article 106878.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), Article 2053951718756684.

Leib, M., Köbis, N., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2023). Corrupted by algorithms? How AI-generated and human-written advice shape (dis)honesty. *The Economic Journal*, *134*(658), 766–784.

List, J. A. (2020). Non est disputandum de generalizability? A glimpse into the external validity trial. In *Working paper series*, (27535), National Bureau of Economic Research.

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*, *38*(6), 937–947.

March, C. (2021). Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. *Journal of Economic Psychology*, *87*, Article 102426.

Meehl, P. E. (1954). *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Paravisini, D., & Schoar, A. (2013). The incentive effect of scores: Randomized evidence from credit committees. (19303), Working paper series.

Pioggia, G., Ahluwalia, A., Carpi, F., Marchetti, A., Ferro, M., Rocchia, W., & Rossi, D. D. (2004). FACE: facial automaton for conveying emotions. *Applied Bionics and Biomechanics*, *1*(2), 91–100.

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th). Pearson.

Schauer, A., & Schnurr, D. (2023). Competition and collaboration between human and artificial intelligence in digital markets. Working paper.

Strobel, C. (2025). The impact of process automation on performance. *Journal of Behavioral and Experimental Economics*, Article 102377.

Verhaeghe, P.-P. (2022). Correspondence studies. In K. F. Zimmermann (Ed.), *Handbook of labor, human resources and population economics* (pp. 1–19). Cham: Springer.

Wang, B., & Komatsuzaki, A. (2021). GPT-j-6B: A 6 billion parameter autoregressive language model.

Webb, M. (2019). The impact of artificial intelligence on the labor market. Working paper.

Werner, T. (2021). Algorithmic and human collusion. Working paper.

Wu, N. (2022). Misattributed blame? Attitudes toward globalization in the age of automation. *Political Science Research and Methods*, *10*(3), 470–487.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414.

Zhang, B. (2022). No rage against the machines: Threat of automation does not change policy preferences. In *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society* (pp. 856–866). New York, NY, USA: Association for Computing Machinery.