Markets, Agency, and Trust: AI Agents and the Knowledge Problem

Brennan McDavid
mcdavid@chapman.edu

Lynne Kiesling [*]
lynne.kiesling@northwestern.edu

David P. Chassin
dchassin@eudoxys.com

October 10, 2025

Abstract

Artificial intelligence (AI) is transforming market participation, raising key epistemological questions: Do AI agents enhance or diminish the aggregation of local, private, and tacit knowledge Hayek saw as essential to market processes? How does trust in both markets and AI shape willingness to engage in AI-mediated exchange? This paper examines these issues through market epistemology, agency relationships, and trust epistemology, analyzing how agentic AI reshapes the knowledge problem and principal-agent dynamics. Applying this framework to transactive energy markets, we show that AI shifts decision-making from human cognition to algorithmic processes that require user trust despite epistemic opacity, although it is no substitute for human cognition. Automated market design must therefore align agent actions with user preferences to ensure both efficiency and trust. Through the case of the TESS platform, we explore how epistemic and institutional structures influence agentic exchange, concluding that the future of automated markets hinges not only on technical optimization but also on fostering trust in AI systems.

Keywords: Markets, Automation, Trust, Technology, Epistemology, Electricity, Artificial Intelligence, Transactive Energy

[*] Corresponding author. No external data were used in this research. The authors declare that they received no direct funding for this research. We are extremely grateful to two anonymous reviewers for their comments and suggestions.

1 Introduction

Market processes reveal and rely on decentralized, real-time individual knowledge, as described by Hayek (1945). Individuals make economic decisions based on their immediate circumstances, preferences, and predictions of future conditions. Prices are discovered and emerge from the aggregation of this dispersed, often tacit, knowledge. The advent of artificial intelligence (AI) tools that rely on machine learning and are capable of attending to more information and in finer detail than is possible for human beings are altering the epistemic landscape of market participation. Increasingly, the real-time decision makers in markets are AI agents directed to execute specific tasks by human principals. This paper is concerned with examining how the knowledge problem is being reshaped by this new technological reality and what barriers stand in the way of optimizing the coordination of knowledge that is possible through AI tools.

We address these questions through three frameworks: a model of market epistemology familiar to the Austrian economics tradition, a model of principal-agent relationships that features AI technology in the role of agent, and a model of trust constructed from the philosophical literature on trust epistemology. Market epistemology describes how markets coordinate decentralized knowledge and facilitate discovery, experimentation, and error-correction through price signals. Delegation in a principal-agent relationship is the definitive mode of market participation in our new technological environment. Trust epistemology examines the conditions under which humans are willing to delegate to AI agents, particularly when those agents operate with a level of opacity and unpredictability. Trust is crucial in this context because human principals of AI agents must delegate decision-making while nevertheless retaining ultimate responsibility for market outcomes. Market epistemology and the principal-agent relationship are the frameworks through which we can assess the dramatic impact of technological change on the fundamental processes of knowledge coordination, while trust epistemology affords us a lens through which to predict and prescribe remedies for barriers to optimizing that coordination.

We consider AI in the context of a widely accepted hierarchy of computers and machines that simulate human learning, problem solving, decision-making, creativity, and autonomy (Stryker and Kavlokoglu, 2024). Current "generative AI" technology is built on so-called deep learning systems which in turn rely on machine-learning technology developed in the 1980's. Most AI researchers distinguish between supervised and unsupervised AI technology to recognize the difference between systems trained using labeled versus unlabeled datasets, respectively, the former being the simpler approach to AI and solving classification problems. Generative AI is distinct from previous unsupervised learning technologies insofar as these deep learning models can generate new content that mimics content on which it has previously been trained. In the context of this paper, we consider agent-based AI or *agentic AI* to be autonomous versions of any AI system that can perform tasks and accomplish goals on behalf of a user or another system without human intervention. Agentic AI comprises multiple AI models coordinated to achieve a particular collective goal that no single agent could accomplish alone.

The combination of their autonomy and the system behavior that emerges from their collective action and impact raises complex ethical and governance questions. In response to these questions, AI researchers have identified five common values associated with responsible AI: (i) explainability and interpretability, (ii) fairness and inclusion, (iii) robustness and security, (iv) accountability and transparency, and (v) privacy and compliance. At the core of these values is a desire to ensure that AI systems can be trusted by those who choose to adopt them, and more broadly by a society that will be required to accept their widespread use. It is the need for this trust that concerns us presently.

The focal illustration of our analysis is transactive energy, an application of AI-driven market participation in electricity systems. Transactive energy systems enable devices, such as smart thermostats and battery storage systems, to participate in local energy markets by autonomously bidding and responding to price signals. The TESS (Transactive Energy Service System) platform, developed at Stanford University's SLAC National Accelerator Laboratory and deployed by the Post Road Foundation in partnership with Efficiency Maine Trust with funding from the U.S. Department of Energy's Connected Communities program, provides a concrete example of how automated market participation functions in practice.

We situate transactive energy within the broader context of market epistemology and trust. The deployment of AI agents as market participants in transactive energy systems does not eliminate the knowledge problem but transforms it. Unlike human market participants, limited by their interpretations of market conditions and bound by their ability to attend to only a few salient needs at a time, agentic AI market participants act as "superagents," processing vast amounts of data, registering a full complement of needs, and devising new strategies in real time. This shift has profound but mixed implications. On one hand, AI agents enhance market efficiency in transactive energy through rapid, high-resolution information processing. On the other hand, they introduce epistemic opacity for their human principals by the sheer fact that a human user is unable to discern or keep pace with the reasoning of AI tools.

Trust epistemology illuminates both the positive and the negative here. If the tools are to act on accurate information about human user preferences and devise the best strategies for market bids, human users must entrust their AI agents with information and decision-making. But this act of extending trust places the human user in a fraught epistemic and affective position. The sea of data and complex strategies informing their AI agent's decisions is unintelligible to them, and yet they must relinquish control of decision-making for which they remain responsible (because they pay the bill due for the market transaction). They are vulnerable, and their perception of this vulnerability can disrupt their willingness to trust. Hence, the introduction of AI tools to the domain of markets has the power to alter the knowledge landscape itself and also is highly dependent on human trust of the tools.

The structure of the paper is as follows. Section 2 draws on market epistemology and trust epistemology to explore (i) how agentic technologies change knowledge aggregation and decision-making in markets and (ii) how the trust attitudes of human users impact these changes. Section 3 applies our analysis to transactive energy, illustrating how agentic AI market participants in local electricity markets alter the possibilities for knowledge aggregation in that sphere and how their efficacy can be throttled by poorly calibrated trust relationships. Section 4 extends the analysis by framing automation within the principal-agent model and discussing the implications of AI as a "super agent" in market environments. Finally, Section 5 concludes with a discussion of design considerations for ensuring that agentic market participation enhances rather than undermines market efficiency, human agency, and trust.

By integrating insights from economics, philosophy, and engineering, this paper develops a framework for understanding the epistemic and trust implications of deploying agentic AI in markets. We argue that the rise of agentic AI market participation presents both an epistemic opportunity and an epistemic challenge. The success of these markets will depend not only on their technical efficiency but also on their ability to calibrate and optimize the trust relationships that underpin market legitimacy and coordination.

2 Modeling the Epistemology of Agentic AI Market Participation

Our exploration of how agentic AI technologies are reshaping the epistemic landscape of market participation relies on market epistemology as developed in the Austrian economic tradition, the conception of principal-agent relationships familiar from studies of firms, and a model of trust inspired by the philosophical literature on trust epistemology. AI agents are transforming the aggregation of knowledge that Hayek identified as inherent to market processes, but this transformation is conditioned upon the propensity of human principals to trust or distrust their AI agents. If human principals withhold information, override, or otherwise interfere with the autonomy of decision-making in their AI agents, they will sharply curb the discovery and coordination power of these tools. But the realization of the theoretically most efficacious deployment of AI agents in markets gives rise to a separate problem: complete delegation and non-interference by human principals manifests as epistemic deprivation. Humans will be ignorant of the market process itself as well as what information is being relied upon by the AI agent who represents them in market transactions.

Our analysis begins with a description of the market epistemology of Austrian economics.

This account endows us with a predictive framework through which we can set baseline expectations for the aggregation of knowledge that is possible through market participation. Connecting that framework with the model of principal-agent relationships—particularly Human-AI iterations of such relationships—reveals the dependency of knowledge aggregation on the quality of information from which the agent acts and, thereby, the quality of the relationship between principal and agent. Completing our analysis, then, is a model of the trust attitude at the core of the principal-agent relationship. The possibility of AI agents enhancing the epistemic mechanisms of market participation depends on cultivating high quality trust attitudes in human principals.

2.1 Market Epistemology

Following Hayek (1945), we understand markets as institutions that enable large numbers of people to coordinate their actions and plans over time and place. Hayek's seminal insight, that much economic coordination relies on localized, often tacit, and ephemeral knowledge held by the "man on the spot" (Hayek, 1945, p. 524), draws attention to the fact that such knowledge rarely exists apart from the individual who directly experiences it. In a traditional market setting, each individual knows best their own circumstances in real time. Prices, as signals, communicate fragments of this dispersed knowledge, making it possible for strangers to coordinate in ways that no central planner could replicate. Markets are thus institutions that address the knowledge problem (Kiesling, 2015). They are epistemic systems (Koppl, 2006).

An array of complex mechanisms comprise the epistemic system of a market. For example, the preferences of consumers for a novel product are unknown until the product is introduced, purchased, and used. Markets facilitate this discovery by enabling experimentation and exchange. Shared meaning is another key channel through which markets generate epistemic outcomes. As Searle (1998) argued, shared symbols like language and money connect individual subjective realities to create social structures. Likewise, prices and market processes serve as feedback mechanisms that enable agents to coordinate plans and adapt to changing conditions.

The advent of AI technology is transforming these mechanisms. Human users are delegating the task of market participation to AI agents by encoding the technology with preferences for a future array of states, such as future discomfort levels, future willingness to pay for incremental changes in temperature, or future thresholds for resource consumption. AI agents implement these preferences in line with their own evaluation of market signals and other data, acting on the user's behalf but in accordance with their own determination of what information is salient at the moment of time-and-place decision-making. Accordingly, AI agents are increasingly manifesting the discovery procedures that Hayek's observes in market processes.

The epistemic advantages of this development are unclear, however. Hayek emphasized that coordinating actions and plans hinges on participants having personal knowledge relevant to the "here and now". AI agents defy this particular feature of the epistemic mechanisms of markets in two ways. First, the delegation of market participation to an AI agent removes human beings from direct engagement in the real-time context of each transaction. The private knowledge of the human user, on whose behalf the market transaction is being carried out, is replaced with the information that the AI agent determines to be salient and actionable. Knowledge may yet be aggregated through these market processes, but that knowledge is likely stripped of the tacit and ephemeral knowledge borne by human beings acting as direct participants. In this way, AI agents are no more disruptive of the epistemic system of market processes than other iterations of indirect participation, but there is a second way they disrupt that is unique

Human cognition relies on shortcuts, requires rest, and has a narrow scope for attention. AI technology is not bound by human cognitive limits. Modern sensor networks and machine learning algorithms can provide AI agents with a granular, real-time pictures of the environment, more granular than a human would ever notice. For example, an AI agent can simultaneously attend to and integrate high resolution changes in building thermal properties, occupant behavior patterns, external weather forecasts, and price fluctuations to decide on a precise energy bidding strategy. The AI agent far outstrips the human user's knowledge horizon, capturing more detail at more frequent intervals than the user could manage independently. Thus, the idea that AI agents bring the knowledge of their user to bear in market processes is fantastical. AI agents are bringing much more information to market processes than human beings are capable of bringing on their own, but this information is not readily understood as comprising the "personal knowledge" of human users of these tools. So in addition to removing the tacit and ephemeral knowledge of the human user from the market process by replacing direct engagement with indirect engagement, agent AI participation additionally rewrites even the articulable and focal knowledge of the human user by increasing the resolution and frequency of updating to a level that the human being cannot keep pace with.

In place of the Hayekian vision of people aggregating private and tacit knowledge through direct market participation, we must insert a model of market participation carried out by agents representing principals. We must also update our conception of the human-human iteration of the principal-agent relationship to accommodate the changes introduced by AI technology performing the role of agent.

2.2 Human-AI Iterations of Principal-Agent Relationships

The principal-agent model in economics was first articulated by Jensen and Meckling (1976):

> We define an agency relationship as a contract under which one or more persons (the principal(s)) engage another person (the agent) to perform some service on their behalf which involves delegating some decision making authority to the agent. (Jensen and Meckling, 1976, p. 308)

The framework highlights both advantages and potential problems in delegating authority, including the moral hazard inherent in misalignment between principal and agent's interests and adverse selection when the principal cannot discern the agent's motives prior to delegation. Early work in this area built on foundational ideas from information economics and game theory, emphasizing how information asymmetry shapes the interaction between principals and agents.

Instances of the principal-agent relationship that feature AI technology in the role of agent present special advantages and risks beyond classical human-human iterations. They replicate the advantages borne of the cognitive division of labor in these relationships. The agentic technology—just like a human agent—frees the principal to perform other tasks that either are more enriching or require more generalized judgment than the agent possesses.[1] Our principal-agent model thus treats delegation to AI agents as a mode of co-intelligence (Mollick, 2024).

As for the risks, AI agents fundamentally alter the principal-agent problem that gives rise to moral hazard, adverse selection, and information asymmetry risks. Jensen and Meckling (1976, 308) originally set out the principal-agent problem as arising "if both parties to the relationship are utility maximizers." That is, when each has their own private interests, there is reason to doubt that the agent will act in perfect alignment with the principal's interests, and so the principal will incur "agency costs" through monitoring or setting incentives in place for enhancing alignment. AI, lacking private interests of its own, is not a utility maximizer in the same sense and so does not give rise to exactly the same alignment problem. However, the human authors of the algorithm endow it with *their* interests, thus generating a second-order version of the problem. Additionally, Bostrom (2014, 127–129) theorizes that the workings of autonomous technology is sufficiently opaque to us that we cannot rule out

[1] An illustration of this cognitive point is the famous passage from Alfred North Whitehead: "It is a profoundly erroneous truism, repeated by all copy-books and by eminent people when they are making speeches, that we should cultivate the habit of thinking of what we are doing. The precise opposite is the case. Civilization advances by extending the number of important operations which we can perform without thinking about them. Operations of thought are like cavalry charges in a battle—they are strictly limited in number, they require fresh horses, and must only be made at decisive moments." (Whitehead, 1911, p. 45)

the possibility of it developing strategies and behaviors that positively "harm the project's interests." Borch (2022, 3) characterizes the potential harm as arising when "the agent's behavior and decision-making logic are not a result of human instructions." From this insight, implementation of controls are recommended that would preclude deviations (Bostrom, 2014, 129–144). These controls are agency costs peculiar to the human-AI relationship. Thus, the human-AI relationship is burdened with two distinct sets of risks, neither of which involve any private interests being predicated of the technology.

The cognitive distance between the human principal and the market transaction, described in the previous subsection, gives rise to tension in relation to Hayek's framework, too. The principal's direct knowledge of time and place is replaced by a forecasting exercise: the principal tries to imagine possible future states and encode them as rules or preferences (e.g., "I value comfort over cost within a certain budget", or "I will pay up to X for Y degrees of cooling"). Hayek's original framework assumes that decisions emerge from contextual knowledge, from a confluence of immediate personal circumstances, which might shift daily or even hourly. When a principal pre-specifies preferences, any unforeseen change in context runs the risk of misalignment, especially if the principal's preferences evolve but are not updated in the agent's model.[2]

AI agents may have more focused knowledge of certain variables (temperature, price signals) than humans, but they lack the full spectrum of lived human context of the person on whose behalf they are engaging in transactions.[3] They do not literally feel discomfort, nor do they weigh intangible priorities unless those priorities are modeled explicitly. Thus, in place of a single, holistic "man on the spot", we get a cognitively specialized system that is extremely good at certain tasks but that may lack broad human judgment. When coupled with a user's preferences, however, this synergy can yield an expanded epistemic architecture: the principal provides overarching goals informed by her cognitive experience of her lived context while the agent provides enhanced local knowledge of time and place.

In the move from direct involvement in market decisions to a delegated model, market epistemology is not necessarily invalidated; it is reconfigured. Hayek's insight that diffuse private (and often tacit) knowledge must be aggregated for coordination still applies, but the nature of that knowledge changes. The principal's input now involves advanced forecasting and preference articulation, while the agent's input focuses on real-time data processing. The knowledge problem is mitigated by the agent's specialized capabilities. It cannot be eliminated because at its core it is a cognitive characteristic of humans (Kiesling, 2015). But how well this epistemic system operates (as measured by the economic concept of total

[2] A related aspect of market epistemology is the discussion of the feasibility of AI-enabled technosocialism; see Boettke and Candela (2023).

[3] See Foss (2002, 17) for discussion of "the impact of principals becoming increasingly uninformed about the actions open to agents [...] and at the same time becoming increasingly reliant on the knowledge controlled by agents".

surplus) remains dynamic: how the agent learns the principal's preferences, how frequently those preferences are updated, and how the principal trusts the agent all affect the accuracy and timeliness of the final market signal (the bids or offers).

Rather than breaking Hayek's epistemic framework, then, agentic AI market participation reframes it. This dynamic rearrangement requires treating market systems with AI participants not just as technological or economic constructs, but as epistemic systems that either enrich or obscure the "time and place" insights essential to coordination. Trust becomes pivotal: if users trust the agent to handle local decision-making, they can enjoy higher-resolution market signals and more efficient outcomes. If they mistrust, withhold, fail to update, or misinform agents about their preferences, the system's epistemic potential erodes. The feedback process between user expectations and realized outcomes undergirds this epistemic potential, a relationship that the system engineer designing the technology system must incorporate.

2.3 Trust

The promise that agentic AI will enhance the epistemic mechanisms of markets is contingent upon the willingness of human users to entrust algorithms with personal information and discretion. The *telos* of participation in markets is the procurement of resources for enabling the preferred activities of the human user, and the means of achieving that goal is the aggregation of the contextual and private knowledge of the many human beings who come to markets (whether as self-representatives or represented by AI agents) with the same goal. If AI agents are deprived of information that genuinely tracks the preferences of their principals, then the mass of information that AI agents bring to market processes will nevertheless fail to realize their purpose. Thus, the human user must trust the algorithm with a sufficient grasp of her contextual knowledge.

Following the dominant line among philosophers, we conceive of trust as an affective attitude that combines cognitive and emotional content in an optimism about the will and competency of the trustee (i.e. the entity that bears the locus of control but not the locus of responsibility) (Baier, 1986; Holton, 1994; Jones, 1996). A trust attitude, in general, may be optimistic or pessimistic, however. The trust dynamic can manifest between peers and in shared activities, but it is particularly salient in principal-agent relationships which, by definition, uncouple the locus of control over decision-making and task-execution from the locus of responsibility for them. Trust, in our focal principal-agent relationships, is an attitude that the principal has in relation to the agent. The cognitive content in this attitude is grounded in uncertainty about the agent's will and competence. The emotional content stems from a feeling of vulnerability arising from the perception of risk-taking in bearing responsibility but not control (Hawley, 2017; Mele, 2004; Schoenfield, 2014). The optimism or pessimism is a disposition to relinquish control or withhold it despite the uncertainty and vulnerability. This disposition is dynamic. The principal's uncertainty and the perception of risk-taking may diminish, increase, or otherwise change through interaction with the agent. In the case that the principal becomes certain of the agent's will or competence or else no longer perceives risk in the dependence, the relationship shifts away from trust toward reliance.[4]

As smart devices, artificial intelligence tools, and other technologies designed for autonomous action have become increasingly ubiquitous, so has the impetus to extend the philosophical analysis of trust to human-AI iterations of the principal-agent relationship (Glikson and Woolley, 2020). It is easy to see why. Unlike machines and tools that are fashioned in such a way that all of the parameters of the mechanisms necessary to the performance of their function are set and modified by a human designer, devices that rely on machine learning can adjust their own parameters in response to patterns that they detect in data sets. The autonomy in this self-adjustment has engendered uncertainty and a sense of vulnerability among users of these tools. Individuals have, for 20 years now, reported feelings of vulnerability in using AI, and this vulnerability is believed to be caused by the user's uncertainty about how AI operates (Hoff and Bashir, 2015; Lee and See, 2004).

We acknowledge the causal role of that uncertainty, but additionally propose that human users perceive a special vulnerability in abdicating control to AI agents, a vulnerability that is different from or more extreme than the vulnerability felt in abdicating control to fellow humans. Human-AI iterations of agency relationships are fraught with peculiar trust problems. With human trustees, relinquishing the locus of control is accompanied by the intuition likely grounded in fellow-feeling that the trustee will share responsibility for outcomes—both gains and losses—and will permit us to recover or retain some modicum of control. The locus of control is perceived as not entirely relinquished, then, and the locus of responsibility is perceived as not absolutely retained. The corresponding perception of risk and feeling of vulnerability is blunted. We cannot relate to AI agents in this way, however. They have no intuitive appeal as sharers of responsibility or of control. The loci of control and responsibility are perceived as absolutely divorced between the truster and the trustee, then, and the corresponding vulnerability is extreme.

On our analysis, then, human-AI iterations of principal-agent relationships come under a trust analysis just in case the principal (i) is uncertain of either the competence or the will (motivations and objectives) of the AI agent and (ii) perceives risk in relinquishing control to it. Not every instance of relinquishing control to an AI agent involves non-negligible stakes and a perception of risk, so not every instance will qualify as a trust relationship. Dependence on robot vacuums to clean floors is low-stakes even if the mechanism of autonomous decision-making is uncertain. Dependence on autonomous vehicles to drive on highways, however, does

[4] We will treat monitoring as the salient contrary to trust in Section 4.

invoke trust because the perceived risk and accompanying uncertainty are high (Hegner, Beldad and Brunswick, 2019). Not all domains of AI tool deployment involve a sense of meaningful risk-taking, then, and so not all come under a trust analysis.

We also recognize that the trust attitudes in these relationships have complex, not simple, objects. That is, what gives rise to uncertainty and a sense of vulnerability is rarely the AI technology alone, but is a mass of the market process, the user interface, even the company or individuals responsible for producing the tool. Human principals may bear varying trust attitudes toward each of these entities, with the collective emerging as something distinct from the mere sum of parts in the mind of the truster. Likewise, the object of trust (even in all its complexity) is embedded within some context of use and purpose for the principal. Uncertainty and vulnerability in risk-taking within that context layer upon the relationship to the tool. Again appealing to the example of autonomous vehicles, the trust attitude that human users have toward other drivers, toward other vehicles, even toward infrastructure, all shape the context in which an individual develops an affective attitude toward the autonomous vehicle that is making decisions and executing tasks for them. Both the nexus of trustees and the context in which trust is embedded are determinant of the trust attitude itself, then.

Our model of how agentic AI is reshaping the epistemic mechanisms of market processes is thus informed by a complex, scaffolded conception of trust in principal-agent relationships. The basic unit of the trust relationship is shaped by the severance of the locus of control from the locus of responsibility, but this basic unit is, in some Human-AI relationships, multipied by as many trustees as are perceived by the human principal as uncertain and risky delegates in the task at hand. Each application of agentic AI stands in need of this trust analysis for determining the specific entities or attributes of entities that are causing the uncertainty and sense of vulnerability in the principal. Optimizing the outcomes of applications of these trust depends on this analysis, and in the case of applications in market participation for enhancing the aggregation of private knowledge, optimizing the principal's trust is indispensable.

3 Application to Transactive Energy: TESS

The AI technology that occasions this investigation is a tool designed to automate a specific iteration of market participation, an agent embedded in a transactive energy (TE) system for the purpose of automating various tasks within that system. Specifically, the agent acts on behalf of a user, interacting frequently with markets for energy (electricity) and actually operating certain energy-consuming/producing devices. The user depends on the agent to deploy bidding strategies in the power market that maximize the user's benefits from these energy-consuming/producing with due consideration of the user's preferences. These preferences extend over both the final cost of energy consumption and the optimization of use of the energy-reliant devices that are operated by the agent.[5] Along both of these preference dimensions, the user perceives risk in depending on an automated decision-maker, and this risk is enhanced by uncertainty regarding the mechanism guiding the agent's interaction with markets and operation of devices. Accordingly, the user's attitude toward the agent is best characterized by its uncertainty and vulnerability and is, therefore, an attitude of trust.

Transactive energy applies principles of market economics and decentralized coordination to the management of modern power distribution systems, enabling more flexible and dynamic ways to match supply and demand compared to traditional centralized control methods. The GridWise Architecture Council defines transactive energy as "... a system of economic and control mechanisms that allows the dynamic balance of supply and demand across the entire electrical infrastructure using value as a key operational parameter" (GridWise Architecture Council, 2019). Transactive energy systems are engineering control systems in which economic signals are used to coordinate heterogeneous assets that consume, produce, or store electricity. As such, it does not provide a solution to the knowledge problem, but instead transactive energy provides an example of how agentic AI can reveal previously unobservable private knowledge.

As an example, consider a neighborhood where each household has a smart, digital thermostat connected to a shared digital communications network. The residents are able to provide their thermostats with preferences that reflect their willingness to pay for heating or cooling at different times of day or under varying circumstances; typically this programming is implemented by the resident choosing a user profile that best matches their preferences. That profile is designed by the technology system engineers. Through the transactive energy system, these smart thermostats autonomously submit bids into the local energy market at frequent intervals, based on the household's programmed preferences. A particular thermostat's bid price being lower than the current market clearing price signals that the household is willing to curtail energy consumption for heating or cooling to save money at that time, and the thermostat will adjust the household's temperature settings automatically to reduce energy use. Conversely, if a thermostat's bid price is higher than the market clearing price, this outcome communicates that the household places a premium on maintaining thermal comfort and is willing to pay more for heating or cooling during that period. The thermostat will then keep temperature settings within the preferred range, allowing higher energy consumption. By responding to real-time market prices, the networked thermostats optimize energy usage across the neighborhood based on aggregated household preferences. In this sense, TE embodies Hayek's concept of decentralized coordination and markets as knowledge ecosystems.

The first transactive energy project, the GridWise Olympic Peninsula Testbed Demon-

[5] These preferences may also include environmental preferences, from which we abstract for this analysis.

stration Project, was a field experiment in 2006-2007 that demonstrated using market signals to coordinate a system of thermostats autonomously (Hammerstrom et al., 2008; Chassin and Kiesling, 2008). The TESS (Transactive Energy Service System) platform deployed by Post Road Foundation is a current TE pilot project in Maine, with funding support from the U.S. Department of Energy's Connected Communities program. This pilot involves the integration of a diverse range of DERs and flexible loads, including rooftop solar, battery energy storage, heat pumps, electric vehicles, and smart appliances. The goal is to demonstrate how the TESS market-based coordination mechanism can optimize distribution system operation in rural communities while also providing valuable insights and lessons for future deployments.

3.1 TESS Technologies and Design

Several key technologies underpin the automation, market participation, and control essential for transactive energy systems. Digital meters provide high-resolution data on consumer-level energy consumption and production, enabling automated bidding in transactive markets based on real-time conditions. Control and optimization algorithms streamline this bidding process by incorporating inputs such as customer preferences, grid conditions, weather, and market prices, while optimizing the dispatch of distributed energy resources (DERs) according to market clearing prices. Internet of Things (IoT) devices and home/building automation integrate sensors, thermostats, appliances, lighting, energy storage, and EV chargers, enabling coordinated demand response through transactive signals. DERs—such as rooftop solar, battery storage, flexible loads (e.g., HVAC, water heaters), and vehicle-to-grid systems—support localized supply and demand management using dynamic pricing. Reliable, low-latency communication networks link grid operations with DERs, smart devices, and market platforms, ensuring real-time coordination and control. Cloud platforms aggregate and process large data streams needed for optimization algorithms, while edge computing devices provide localized intelligence. Robust cybersecurity measures, including encryption, access controls, and intrusion detection, are critical for securing the distributed and automated nature of transactive systems. Together, these technologies enable key capabilities such as automated demand response, DER integration, real-time pricing, and market operations, fostering active participation from consumers, utilities, retail providers, aggregators, and system operators. Machine learning and other AI-related methods enable learning of and adaptation to patterns in user behavior and market outcomes.

TESS supports a variety of devices, including: (i) feeders, which manage electricity distribution and supply-demand balance; (ii) heat pumps, which provide building heating and cooling; (iii) water heaters, especially heat-pump models that facilitate price-based energy management; (iv) rooftop solar panels for on-site generation; (v) EV chargers, which autonomously adjust charging based on electricity prices; and (vi) stationary batteries, which store and release energy as needed. TESS optimizes the operation of these devices to enhance energy efficiency and grid stability.[6]

Transactive energy systems can operate differently depending on their design. Here we adopt a taxonomy of transactive systems, allowing us to compare and contrast the benefits and costs of each system, as shown in Figure 1.

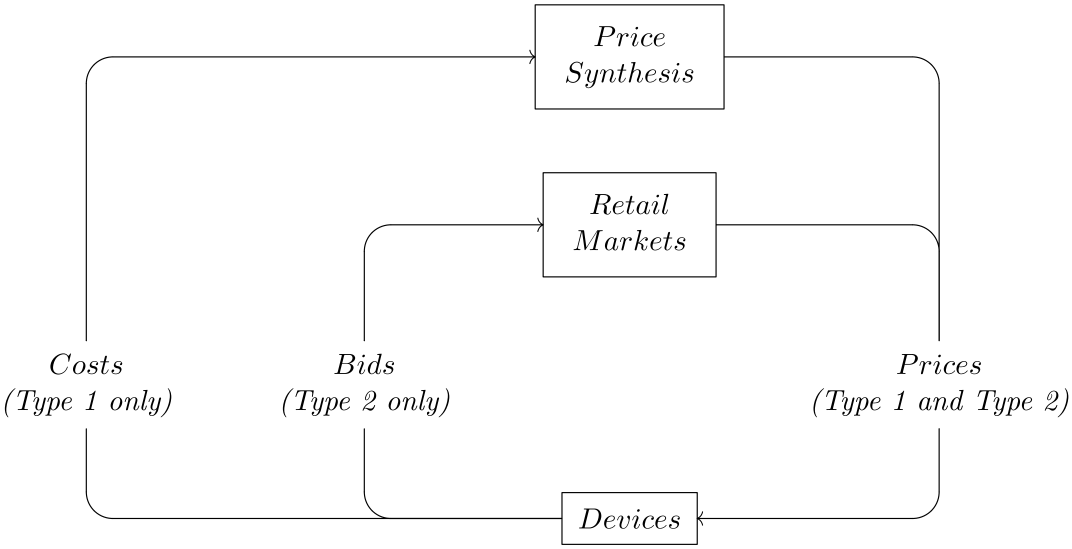

Figure 1: Autonomous system diagram of Type 1 and Type 2 transactive energy system behaviors.

Type 1: In a Type 1 Transactive System, economic signals regarding upcoming actions only flow to assets prior to control actions. This is often referred to as "*prices to devices*" and is analogous to traditional posted pricing or administrative pricing. Cost and/or historical performance data are used to generate a price. No information is collected from assets about their willingness or ability to produce or consume at any particular price prior to the price being revealed and there is no guarantee or even expectation that quantity supplied will equal quantity demanded. For example, Southern California Edison's Retail Automated Transactive Energy System (RATES) project communicated real-time wholesale power market prices to the thermostats of participants, whose devices would autonomously change settings or not depending on whether the user's set price was above or below the market-clearing price.

Type 2: In a Type 2 Transactive System, economic signals regarding upcoming actions flow from assets prior to control actions, and is often referred to as "*prices from devices*"

[6] Related complementary developments are provisions guaranteeing data privacy and cybersecurity, which are embedded in the TESS design.

and is analogous to market pricing arising from the interaction of buyers and sellers. Assets provide information about their willingness or ability to produce or consume various quantities at various prices prior to the price being revealed. TESS is a Type 2 system because it allows assets to provide the markets with information about their willingness and ability to forgo consumption or production at a particular price prior to the actual price being revealed.

3.2 TESS Markets

The key feature of transactive energy systems is that they have one or more pricing mechanisms to generate prices in real time. Type 2 *price discovery* mechanisms are typically called markets.[7] The markets may be of varying types, such as auctions or order books, and usually require input information such as demand forecasts, available resources, reservation prices, etc. to determine the price at which supply will equal demand. The economics literature provides numerous examples of price discovery mechanisms, most of which can be used in transactive energy systems, and there is no single market design that ideally suited for all transactive systems.

TESS employs a synthesis of economic market design principles and control systems engineering, with a strong emphasis on enabling more expressive, preference-based bidding from both the demand-side and DER owner/operator perspectives. At the heart of the TESS project is the goal of creating a decentralized, market-based coordination mechanism that can manage the growing complexity of modern electricity grids. As renewable energy sources and DERs like rooftop solar, energy storage, and electric vehicles become more widespread, traditional top-down control approaches are struggling to manage and balance supply and demand. TESS aims to address this challenge by empowering end-users to participate actively in the optimization of the system. Figure 2 presents a schematic diagram of the design of the TESS platform's electricity flows (double lines) and data flows (single lines), both periodic (solid) and intermittent (dashed).

The TESS approach integrates Austrian economics, auction theory, and mechanism design, emphasizing that individuals have unique, private preferences and valuations. In transactive energy, consumption bids should reflect consumers' marginal willingness to pay rather than just engineering cost minimization. DER owners face subjective and heterogeneous opportunity costs, which complicate coordination. The price system addresses this challenge by aggregating dispersed knowledge within a market framework, allowing prices to reflect both supply and demand dynamics. A well-designed institutional structure enables these prices to coordinate participant actions efficiently.

TESS combines advanced market design with sophisticated device control. Its double-

[7] In Type 1 systems the pricing mechanisms may be called something else.

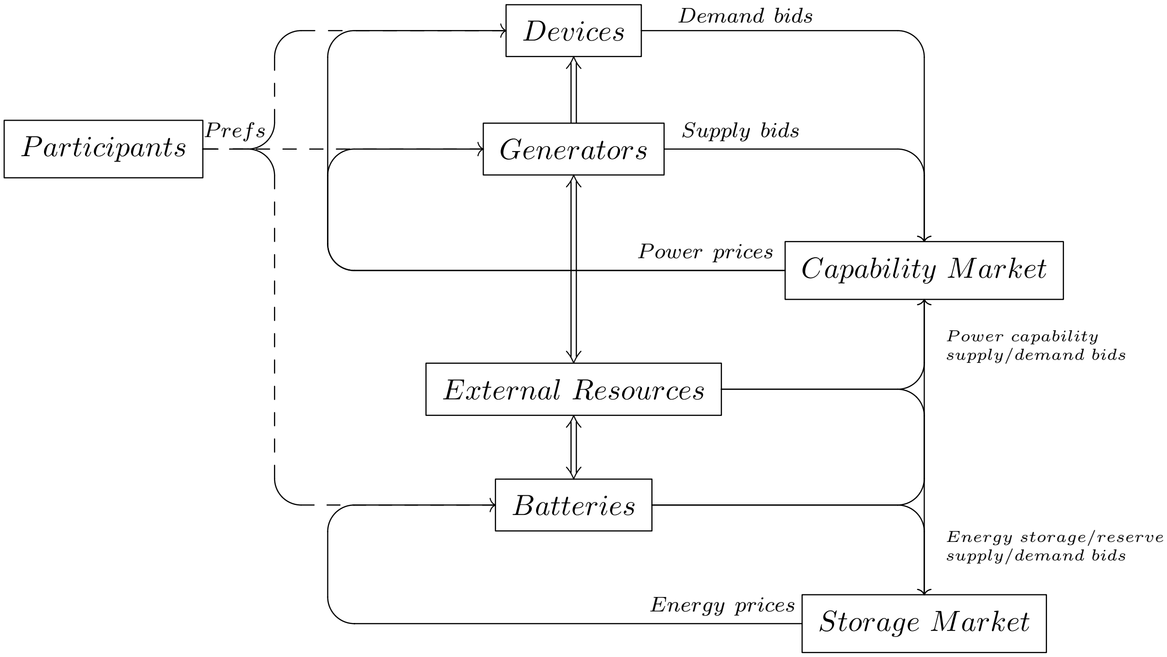

Figure 2: TESS Transactive Energy Platform Diagram

auction mechanism allows consumers and DER operators to submit bids and offers simultaneously. Buyers and sellers provide bid and offer schedules, forming market demand and supply curves. The platform clears the market, determining a price that reflects collective value and opportunity costs. Resources clearing the market must operate as bid, with deviations subject to penalties. Settlements are based on actual quantities at the market-clearing price.

These principles shape TESS's market design and bidding functions, allowing users to communicate dispatch preferences and flexibility directly, rather than relying solely on cost-based or heuristic methods. By aligning market outcomes with users' values and constraints, TESS reduces incentives to deviate from dispatch instructions, enhancing system efficiency.

3.3 TESS Agents

Agents implement the bidding and response strategies for all devices in TESS. An agent is responsible for the operation of one or more devices' within the context of one or more markets given the participants' preferences, as shown in Figure 3. They interact with participants through a user interface specifically designed to elicit preferences in a minimally intrusive and highly intuitive manner. These preferences are combined with historical and forecast data as well as current device status to form bids that are submitted to the market. When

$$\begin{tikzcd}
 & Data \arrow[d, leftrightarrow] & \\
Participant \arrow[r, leftrightarrow] & Agents \arrow[r, leftrightarrow] \arrow[d, leftrightarrow] & Devices \\
 & Markets &
\end{tikzcd}$$

Figure 3: Agent structure in transactive energy system.

the markets reveal the current prices, the agent uses the new price to dispatch the devices according to the bids submitted and the dispatch strategies preferred by the participant.

The key feature of agents is that they can act on behalf of the participants to implement frequent market interactions while not requiring the participant to engage directly in those interactions very often. Although agents must monitor devices, collect data, submit bids and respond to prices as often as every minute, participants need only interact with agents when preferences change or an exceptional or unforeseen condition arises. The "fire-and-forget" aspect of agents is an essential characteristic of transactive agents and one of the foundational elements of the value proposition of a transactive energy system.

Agent strategies can be arbitrarily designed and implemented so long as they provide the following capabilities.

1. Every agent must consider the participant's preferences, if any, as part of the bidding strategy and device dispatch.

2. Every agent must consider historical and forecast data, if relevant and available, as part of the bidding strategy and device dispatch.

3. Every agent must consider the device's current status, if relevant and available, as part of the bidding strategy.

4. Every agent must consider the market price(s), if relevant and available, as part of the device response.

5. When operating in more than one market, every agent must coordinate the bids and responses in each market with those in each other market in which the agent interacts.

Consider one of the simplest agents conceivable, a heat pump water heater with a backup heating coil with power $\hat{Q}$. The agent for this scenario is responsible for controlling the backup heating element based on the market price P relative to a buy limit price $\hat{P}$, i.e., use the coil when electricity is cheap and avoid using the coil when electricity is expensive, but otherwise do not interfere with the heat pump's operation.

The water heater can be in one of three possible discrete states $\hat{S} = \{OFF, HEATPUMP, COIL\}$, the latter two referring to the modes of the normally operating heat pump and the backup heating coil, respectively. The bidding strategy is

$$P_{buy} = \hat{P} \qquad \text{and} \qquad Q_{buy} = \hat{Q} \tag{1}$$

and the dispatch strategy for the water heater's state S when the price P is revealed by the market is given by

$$S = \begin{cases} COIL & : P \leq \hat{P} \text{ and } \bar{S} \neq OFF \\ HEATPUMP & : P > \hat{P} \text{ and } \bar{S} = HEATPUMP \\ OFF & : otherwise \end{cases} \tag{2}$$

where $\bar{S}$ refers to the state specified by the normal operating strategy of the waterheater.

In typical conventional price-based control systems, the participant is expected to determine the price limit $\hat{P}$ that controls when the coil is turned on or off. However, thermostatic controls always end up using the same amount of energy in the long run. Consequently, the difference in the efficiencies of the heat pump and the coil is a significant factor that must be considered when determining the price $\hat{P}$. This determination is something that is likely outside the capabilities of typical participants, making any user interface that asks the participant for their price preference unlikely to yield the best performance in the long run. This aspect of TESS bidding strategies makes agent performance superior to that of a human setting the bidding strategy directly based on their comfort preferences and cost objectives.

3.4 Verification and Validation

The last stage of engineering system design and development requires verification and validation that the design principles and objectives are satisfied by the implementation. Verification can be viewed as a top-to-bottom check that the outputs of the system, subsystems, and components are correct given specified inputs. This check serves to ensure that they meet the specifications, designs, and relevant regulatory and statutory requirements. Verification usually involves testing, analyzing, and reviewing functionality, reliability, resilience, and quality. Typically verification is performed on design documents, models, and early engineering prototypes and are intended to answer the question "Did we design the product correctly?"

Validation ensures that the entire system satisfies the needs of the users and fulfills the intended use-cases envisaged by the designers. These tests are performed on final product to ensure functionality, safety, and performance. Validation tests answers the question "Does the product meet the user's needs?"

However, in the case of agentic systems like TESS, neither of these test procedures directly answers the question “Will the user trust the product?” unless trust is prescribed in the system requirements, strategies for establishing and ensuring trust are implemented in the system, and verification/validation tests are applied to the product. In the next section we analyze the TESS system with the goal of identifying what these requirements, strategies, and tests must achieve to ensure successful deployment of agentic systems such as TESS.

4 Analysis: TESS As Super Agent

Our account of the power of agentic AI to enhance knowledge aggregation in markets emphasizes the importance of analyzing the trust attitude of human principals and optimizing that attitude through design choices. Section 2 provided a general description of the mechanisms by which engagement with market platforms enhances our epistemic condition, explored how the emergence of principal-agent relationships in market participation has altered the epistemic landscape, and specified a basic model of trust that can be deployed in analyzing the layered, embedded, and otherwise complex realities of trust in Human-AI relationships. Section 3 provided a specific application of market participation via agentic systems embodied in the TESS transactive energy platform. Here we synthesize the two to apply the model to analyze specific issues arising in the TESS system and draw conclusions about how one goes about engineering agentic systems that users will trust.

A crucial focus here is the type of agency that is manifest in Human-AI iterations of principal-agent relationships. On our understanding, the technology under examination combines goal-based agency and utility-based agency with learning capabilities, per the range of agent programs recognized by Russell and Norvig (2021). Agentic AI market participants do not enter bids in auctions merely on the basis of current perception and so cannot be simple reflex agents (Russell and Norvig, 2021, 67). And though they are deployed for the purpose of realizing a determinate human goal of, say, procuring the amount of energy required for serving the energy consumption of a particular household, they are not straightforwardly goal-based agents either (Russell and Norvig, 2021, 71-72). Their performance is measured both by the binary success or failure of procuring the requisite commodity, but also by their optimization of costs in market participation. Accordingly, they bring both goal-based agency and utility-based agency together (Russell and Norvig, 2021, 71-72). They are also endowed with several elements of learning agency: they possess learning elements and critics that enable them to be responsive to internalized performance standards, and it is at least possible that they can be built with problem generators that allow them to explore suboptimal bidding strategies in the interest of learning (Russell and Norvig, 2021, 74-75).

In this section, we describe the transformation of human agency that is achieved through delegation to AI agents that endowed with these forms of agency. We then analyze the epistemic issues arising in the relationships between human principals and their TESS agents, the TESS automated market participation system, and the system designer choosing the functionality and capabilities of TESS and how to communicate them to the user.

4.1 Super Agency in Markets

In applying market and trust epistemology to agentic AI market participation, we define the concept of a *super agent*, a system comprised of a human principal delegating the task of realizing specific market goals to an AI agent that is bound to represent the human principal's interests but also possesses capabilities that extend beyond human cognitive limits. The super agency is descriptive of the principal-agent system rather than only the AI agent because the AI tool's objectives, and *a fortiori* its agency, are informed by the human user's preferences. This concept entails more than straightforward "specialization" by including a capacity to gather, process, and act on information at temporal and spatial scales that humans cannot manage realistically. By virtue of its algorithms, data-processing power, and rapid decision-making, the super agent can identify patterns, respond to price signals, and implement strategies at a granularity that would overwhelm any human attempting to do the same.

As indicated above, the AI agent that serves as a component of the super agent is itself endowed with agency both in its performance element (bringing together both goal-based agency and utility-based agency) and in the design of its learning element (Russell and Norvig, 2021, 75). The endowment of agency within the AI agent is generative of independency of decision-making that enables it to maximize the gains to be had from its distinctive, and distinctively non-human, powers of high resolution, high frequency data perception and processing. Through the objective-setting of the human principal and the processing and tactical capabilities of the the tool, the super agent system is poised to enrich price discovery processes, potentially discovering and exploiting opportunities humans would likely miss.

The knowledge problem emphasizes the decentralized and private nature of knowledge in markets. In the traditional Hayekian account, humans bring distinct time-and-place knowledge to the market, and the price system aggregates these dispersed bits into information. Super agents complicate this framework by interposing themselves between human principals and the market process. Instead of humans supplying their private knowledge directly, super agents translate user preferences—provided incompletely in advance—into rapid, data-intensive actions within the market.

This dynamic alters how knowledge is generated and transferred. AI agents do not merely convey a user's preferences; they also interpret, refine, and augment those preferences in light of real-time data. This transformation means that price discovery and market outcomes become a product of AI-driven knowledge processing at scale. As multiple super agents interact, they intensify the pace at which offers, bids, and prices can and do update. They might detect micro-level changes in supply and demand, thereby improving allocative efficiency, but they also introduce new epistemic blind spots. The human principal may not grasp fully how the AI agent arrives at its decisions. Knowledge shifts from being embedded only in the mind of the person "on the spot" to being embedded in the AI's algorithms and data streams.

In Hayek's model of the knowledge problem, the principal's direct interaction with the institution allows the full development of necessary knowledge to solve the problem in both, as shown in Figure 4.

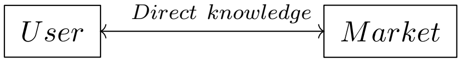

Figure 4: Knowledge problem in the absence of a super agent

However, with the AI agent interposed between the principal and institution, the principal's exchange of knowledge with the market is limited by the agent, while the agent gains access to the full knowledge once available to the principal albeit constrained by the designer's model of knowledge, as shown in Figure 5.

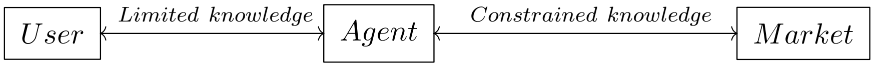

Figure 5: Knowledge problem in the presence of a super agent

The super agent system that is realized though the principal-agent framework is pivotal. While humans theoretically retain responsibility and ultimate control, they often "set it and forget it," allowing the AI to operate autonomously for extended periods. This condition invites two concerns, preference alignment and trust vulnerabilities. The super agent must implement the human principal's preferences accurately, even if those preferences are imperfectly defined, evolve over time, or cannot be anticipated fully in advance. Misalignment can arise from design biases, flawed algorithms, or unrealistic assumptions about human behavior. The human user must trust that the AI tool will continue to act in their best interest. The human user also must accept partial opacity—constant monitoring and interference diminishes the advantages of the super agent system.[8] This requirement to "not pay attention" underscores a tension: the most effective super agent may be one allowed to operate with minimal human interference, yet the user's trust must be sufficiently robust for them to grant such autonomy.

[8] Though interference can diminish AI efficacy through forcing it to suboptimal decisions, recent studies indicate underperformance of AI when left unmonitored as well. See for example Lee and See (2004). Overreliance is the contrary to monitoring and interference Klingbeil, Gr"utzner and Schreck (2024). We recognize that this can diminish super agency as well as interference because the former omits the human objective while the latter dulls the AI capabilities.

Because AI tools wield such powerful capabilities, the human user is newly vulnerable to design decisions and hidden incentives embedded in the AI. A human principal's epistemic position shifts from personally grappling with the complexity of real-time data to relying on an AI that, in effect, sees much more but is also constrained by the system's design. Designer choices are determinative of functionality, user interface, and learning algorithms. In other words, they are determinative of the very agency of the AI agent. These decisions are highly consequential: the designer's assumptions, biases, and interpretations of "optimal outcomes" inevitably constrain the super agent's learning and actions.

Consequently, the principal's risks and uncertainties shift. The traditional limitation of individual cognition—Hayek's emphasis on humans not knowing enough—remains but becomes less relevant, replaced by the constraint of how thoroughly and accurately the super agent's design captures the principal's real preferences and constraints, although the pervasiveness of the knowledge problem means that the agent cannot capture them fully. In practice, for example, this might mean a user who values short-term budget certainty above all else but fails to communicate that preference clearly to the super agent. If the super agent is designed to optimize different objectives (e.g., maximizing long-term savings by taking on near-term risk), the user could experience an outcome that feels misaligned, even if it is "rational" from the system's standpoint.

In a super agent-driven market, users must externalize their preferences in advance, often under uncertainty and with incomplete self-knowledge. Only by observing the agent's performance over time can the user refine these inputs. This iterative feedback loop becomes an important aspect of trust-building and preference discovery.

Indeed, there may be times when the user would prefer not to notice every decision. "Optimal inattention" suggests that delegating attention-intensive tasks to a super agent is valuable precisely because it frees humans from micromanagement. However, too much disengagement can erode informed oversight. Designers can mitigate this tension through careful user interface design that delivers insight without requiring constant human intervention.

The super agent concept reconfigures the knowledge problem by transferring cognitive burden from the human principal to an agentic mechanism that operates with greater speed and granularity. This transfer has the potential to increase consumer surplus and market efficiency, expedite price discovery, and reduce human error. At the same time, it introduces new challenges: trust must be cultivated toward a machine whose inner workings may appear opaque, and principal-agent misalignment may emerge if the super agent's design or algorithms deviate from user preferences.

This dynamic underscores a fundamental tradeoff. On one hand, super agents promise an evolutionary advance in market-based resource allocation by enriching the epistemic environment and discovering opportunities humans would miss. On the other hand, these benefits hinge on thoughtful design and management of user trust. Institutional structures, transparency mechanisms, and robust preference-setting interfaces must be developed to ensure that super agents truly extend, rather than distort, human agency in the market.

By framing super agents within both an epistemological and principal-agent perspective, we see that the question is not whether AI can solve Hayek's knowledge problem outright (it cannot) but how it reshapes the very nature of that problem. The super agent represents a powerful new instrument for harnessing dispersed information, yet it also changes the fundamental trust relationships and error risks that arise when humans delegate market participation to intelligent machines.

4.2 Application to TESS

TESS is a system of super agents. The platform operates by autonomously bidding and optimizing energy transactions on behalf of users, leveraging data analytics to respond to fluctuations in supply, demand, and price with minimal human oversight. By doing so, TESS embodies the primary traits of a super agent: it acquires knowledge at a level of detail that humans typically cannot match, and it acts with a frequency and speed that outstrips human capacity.

Despite these strengths, TESS's effectiveness depends on the user's trust in both the system and in the broader market context. This is not a trivial requirement. Unlike a purely mechanical intervention (e.g., a passive device that adjusts temperature based on a user-set schedule), TESS actively navigates market signals and attempts to optimize on behalf of users, often in ways the user may not fully understand.

In the particular case of the super agent in an automated transactive energy system, the user's trust attitude is made complex by the presence of at least three trust relationships that converge in its vicinity: (i) user trust of markets: the user's willingness to rely on market price signals and accept market outcomes as generally fair, efficient, or reliable; (ii) user trust of technology generally, the user's broader stance toward technological solutions, which can range from technophilia to technophobia; and (iii) user trust of the super agent specifically, the user's belief in TESS's capacity to interpret and fulfill their preferences without introducing unacceptable risks. These relationships overlap in ways that can amplify or dampen one another's effects. A user who already feels uneasy about markets, for instance, may be even more cautious about an AI-driven system that transacts in those markets on their behalf. Conversely, a user who enjoys experimenting with new technologies may feel more inclined to accept TESS's decisions, despite harboring doubts about the underlying market structure. Modeling the relationship between user and agent requires mapping the uncertainty and vulnerability in each of the constituent trust relationships, and optimizing the user experience requires sensitivity to the interdependence of uncertainties and vulnerabilities in the convergence.

The goal of optimizing user experience also cannot be anchored solely to the affective attitude of the user. The importance of user *engagement* to the functioning of the tool also figures in the goal. In the case of autonomous cars, for example, user experience crucially depends on specific, even frequent user engagement due to the dependency of the tool on such engagement. In the case of smart devices for managing heating and cooling systems in residences, however, user experience may be optimized in part through inducing minimal user engagement with the tools, allowing the tool maximal autonomy. User experience is indexed to the function of the tool, then, in much the same way that the user's trust attitude is indexed to their perceptions of uncertainty and risk-taking. Because user experience takes the trust attitude as a component part, the function of the tool doubly figures in the project of optimizing the user experience.

Transactive energy systems can realize the epistemic benefits of market platforms in the absence of user monitoring, which means that the AI technology deployed in the agent of those systems may function best when its autonomy in decision-making is left undisturbed by fine-tuning of the principal's preferences, especially if the principal's adjustments are to remedy deviations of the agent's performance from the user's objectives. This deviation is most likely to occur when the choices presented to the user are misaligned with either the user's preferences or the agent's true capabilities. Thus the user interface and user experience design itself becomes crucial to the success of the agents in ways that may not be immediately apparent to the system designers and differ considerably from the conventional engineer design and development approaches used for other parts of the system.

The agent in TESS dials in its execution of tasks most efficiently if it is permitted to set its own parameters. Indeed, in the abstract, user experience of TESS would seem to be optimized through the agent's maximally efficient execution of tasks, and so this may seem to be an AI tool for which the trust attitude falls into the background of any strategy for experience optimization. TESS reveals that we can be our own worst enemies when we try to act on our own behalf, and that it's possible for a broad range of use cases that the agent can create what we want more effectively without our action/interference.

This is where the convergence of multiple trust relationships is illuminating. The user, depending on her standing affective attitudes toward (i) markets, (ii) technology, and (iii) the super agent in particular, will be more or less likely to depend confidently upon the agent in TESS to execute its tasks desirably. Depending on the user's perceptions of uncertainty and risk-taking in each of the constituent trust relationships (i), (ii), and (iii), she may be ready to rely on the agent and engage minimally or else she may be inclined to co-opt the agent's autonomous decision-making by engaging frequently.

Fully profiling the attitude of a high-engagement user requires examination of each constituent trust relationship to discern which, if any, introduces the predominant uncertainty and which the predominant vulnerability to the overall attitude of low trust. In turn, moving the user toward an optimal experience requires mitigating the relevant uncertainty and vulnerability, which may vary from user to user. A strategy that, say, deploys explainable AI (XAI) particularly for mitigating uncertainty about the agent's market interaction could raise the optimism of a user whose uncertainty is derived principally from the market trust relationship. But this strategy would be suboptimal for the user whose uncertainty is grounded in the AI technology relationship. And, in any case, the frequency of the agent's market interaction renders impossible any semblance of completeness or usability of an XAI intervention.

4.3 Agentic System Engineering

Engineering best practices for system verification and validation require performance benchmarks. A simple approach models two reference systems: a "baseline" (business-as-usual) and an "ideal" (optimal) system. The actual system's performance should fall between these benchmarks[9]. This method standardizes improvement measurement by defining the baseline at 0.0 and the ideal at 1.0, allowing meaningful comparisons across solutions. Systems performing worse than the baseline receive negative scores, proportional to their relative underperformance.

For TESS systems, performance can be evaluated across metrics such as total surplus, user satisfaction, energy efficiency, emissions, and resource utilization. While user trust lacks a direct observable metric, its impact can be inferred by comparing system performance under two conditions: (1) all agents disabled and (2) agents with perfect knowledge of user preferences. Since most users' willingness to share preferences falls between these extremes, the resulting performance comparison provides indirect but valuable insights into trust's role.

To isolate trust's impact, a baseline could model all engineering behaviors except trust. This would quantify the improvement from fostering trust alone, though no reports confirm this method's use.

A key challenge remains modeling agents' "perfect knowledge" of users. Engineers typically specify user preferences to simulate ideal agent behavior, but this approach may underestimate the ideal benchmark, as it relies on the same models used in system design. This could lead to overestimating actual system performance, though the bias would be consistent across design comparisons. If a system scores above unity, it suggests the ideal model is flawed or the proposed system has achieved an unforeseen optimal strategy, warranting further analysis.

To improve benchmarking, agentic system designers may adopt user-centric strategies that enhance trust within a principal-agent framework. One effective approach is UI/UX design informed by behavioral analysis, such as TESS's user study (Lim, Baltaduonis and

[9] While an actual system may perform worse than the baseline, it cannot exceed the ideal.

Chassin, 2024), which identified four user types based on learning and action preferences:

- *Explorers* are users who mainly seek to learn as much as possible about the system, i.e., learn from the system.
- *Achievers* are users who mainly want to maximize the benefits of using the system, i.e., act on the system.
- *Socializers* are users who mainly want to learn as much as possible about other users of the system, i.e., learn from other users.
- *Influencers* are users who mainly want to impact how others users engage with the system, i.e., act on other users.

Lim *et al.* found that most TESS users were explorers or achievers, suggesting UI/UX should prioritize their needs over those of socializers and influencers.

User typology, while not directly addressing trust, provides a framework for testing it in AI-driven systems. Initial trust depends on whether the UI/UX aligns with user goals—explorers expect tools for learning, while achievers seek optimization features. Misalignment can erode trust in the agent's ability to serve user interests.

Trust also evolves over time. Unlike home-energy products that monetize social engagement, TESS prioritizes alignment between user interests and system performance. Systems that fail to do so risk trust erosion, particularly if users perceive agents as serving vendors rather than themselves. Once trust declines, broader system confidence follows.

The consequences extend beyond a single vendor, potentially destabilizing entire ecosystems. The collapse of FTX, for example, damaged trust in cryptocurrency exchanges overall, forcing even uninvolved platforms to rebuild credibility. Similar risks exist for autonomous vehicles, drone defense systems, and transactive energy markets—one high-profile failure could discredit an entire field. A failed business model may serve as a cautionary tale that becomes integral to our cultural narratives about agentic technologies. These narratives inform expectations and so must be managed carefully.

5 Conclusion

The rise of automated, agentic market participation challenges epistemological and institutional assumptions about how knowledge is aggregated in markets, how trust is established between human users and automated agents, and how principal-agent relationships evolve in environments where decision-making is increasingly delegated to intelligent systems. Our examination of market epistemology and trust epistemology suggests that agentic markets do not eliminate the knowledge problem, but rather transform its structure. The deployment of super agents—automated systems capable of high-resolution data processing, real-time decision-making, and adaptive optimization—reconfigures the process of knowledge creation and transmission within markets, raising new questions about user trust, preference alignment, and system design.

As an example of transactive energy markets, TESS illustrates both the promise and complexity of this transformation. By allowing automated agents to bid into electricity markets and optimize energy consumption based on user-defined preferences, TESS extends market coordination mechanisms beyond human cognitive limits, making it possible to achieve finer-grained, real-time responses to supply and demand fluctuations. As we have suggested, though, the success of such systems depends not only on their technical efficiency but also on their epistemic and institutional robustness. Users must trust the system's ability to interpret and execute their preferences faithfully, even as those preferences remain dynamic, sometimes ill-defined, and contingent on evolving real-world conditions.

The super agent model complicates traditional principal-agent dynamics. Unlike human agents, agentic market participants operate with speed and granularity that far exceed human cognition, raising the stakes for trust and preference articulation. Users must conceptualize and forecast their preferences *ex ante*, rather than engaging directly in each decision point. This change in the locus of decision-making introduces a form of epistemic opacity: decisions are made on their behalf by an entity whose optimization logic may not always be fully transparent. This shift requires a rethinking of trust not only as an affective attitude but as a necessary component of epistemic delegation—users must trust that their agents will act in their best interest, even in situations where the decision-making process remains partially inscrutable.

At the same time, system designers must recognize that user trust is not a given, but is instead a dynamic variable that must be cultivated. Poorly aligned agent behavior, opaque optimization functions, or design choices that obscure user preferences can erode confidence in automated markets, undermining both market efficiency and participation. The TESS platform offers valuable insights into how user trust can be fostered through thoughtful system design. By incorporating mechanisms for preference refinement, feedback loops that allow for iterative trust-building, and transparency measures that make agentic decision-making more interpretable, TESS demonstrates how agentic markets can enhance rather than disrupt human agency in market interactions.

Ultimately, our analysis suggests that AI cannot resolve Hayek's knowledge problem, but it does redistribute it. Markets have always been epistemic systems, aggregating dispersed knowledge to enable decentralized coordination. The introduction of super agents amplifies this process, extending the reach of markets into domains where human cognition alone would be insufficient. It also changes it qualitatively, introducing new epistemic tradeoffs where trust, transparency, and user agency must be managed to ensure that agentic market participation enhances, rather than erodes, the knowledge ecosystem function of markets. Finally, it changes how we design and deploy systems that depend on such epistemic complexities and forces engineers to take on new challenges and responsibilities they have not previously been required to engage in.

References

Baier, Annette C. 1986. “Trust and Antitrust.” *Ethics*, 96(2): 231–260.

Boettke, Peter J, and Rosolino A Candela. 2023. “On the feasibility of technosocialism.” *Journal of Economic Behavior & Organization*, 205: 44–54.

Borch, Christian. 2022. “Machine learning, knowledge risk, and principal-agent problems in automated trading.” *Technology in Society*, 68: 1–10.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Chassin, David P, and Lynne Kiesling. 2008. “Decentralized coordination through digital technology, dynamic pricing, and customer-driven control: The Gridwise Testbed Demonstration Project.” *The Electricity Journal*, 21(8): 51–59.

Foss, Nicolai J. 2002. “’Coase vs Hayek’: Economic Organization and the Knowledge Economy.” *International Journal of the Economics of Business*, 9(1): 9–35.

Glikson, Ella, and Anita Williams Woolley. 2020. “Human Trust in Artificial Intelligence: Review of Empirical Research.” *Academy of Management Annals*, 14(2): 627–660.

GridWise Architecture Council. 2019. “GridWise Transactive Energy Framework v.1.1.” Pacific Northwest National Lab.(PNNL), Richland, WA (United States).

Hammerstrom, Donald J, et al. 2008. “Pacific Northwest GridWise™ Testbed Demonstration Projects; Part I. Olympic Peninsula Project.” Pacific Northwest National Lab.(PNNL), Richland, WA (United States).

Hawley, Katherine. 2017. “Trustworthy Groups and Organisations.” In *The Philosophy of Trust.* , ed. Paul Faulkner and Thomas Simpson, 215–234. Oxford:Oxford University Press.

Hayek, Friedrich August. 1945. “The Use of Knowledge in Society.” *American Economic Review*, 35(4): 519–530.

Hegner, Sabrina M., Ardion D. Beldad, and Gert J. Brunswick. 2019. “In Automatic We Trust: Investigating the Impact of Trust, Control, Personality Characteristics, and Extrinsic and Intrinsic Motivations on the Acceptance of Autonomous Vehicles.” *International Journal of Human–Computer Interaction*, 35(19): 1769–1780.

Hoff, K. A., and M. Bashir. 2015. “Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust.” *Human Factors*, 57(3): 407–434.

Holton, Richard. 1994. “Deciding to Trust, Coming to Believe.” *Australasian Journal of Philosophy*, 72.

Jensen, Michael C., and William H. Meckling. 1976. “Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure.” *Journal of Financial Economics*, 3(4): 305–360.

Jones, Karen. 1996. "Trust as an Affective Attitude." *Ethics*, 107(1): 4–25.

Kiesling, Lynne. 2015. "The Knowledge Problem." In *Oxford Handbook of Austrian Economics.* , ed. Peter Boettke and Christopher Coyne, 45–64. Oxford: Oxford University Press.

Klingbeil, Artur, Cassandra Gr"utzner, and Philipp Schreck. 2024. "Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI." *Computers in Human Behavior*, 160: 108352.

Koppl, Roger. 2006. "Epistemic systems." *Episteme*, 2(2): 91–106.

Lee, J. D., and K. A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Computers in Human Behavior*, 20(6): 765–781.

Lim, Lijing, Rimvydas Baltaduonis, and David P. Chassin. 2024. "Gamification Archetypes Validation for Energy Applications." *submitted to IEEE Transactions on Energy Generation and Power Development*, in review.

Mele, Alfred R. 2004. *Trusted Agents: The Role of Trust in Social Cognition.* Oxford:Oxford University Press.

Mollick, Ethan. 2024. *Co-Intelligence.* Penguin Random House.

Russell, Stuart, and Peter Norvig. 2021. *Artificial Intelligence: A Modern Approach.* . 4th ed., Pearson.

Schoenfield, Miriam. 2014. "Trust and the Problem of Uncertainty." *Social Epistemology*, 28(4): 321–337.

Searle, John R. 1998. *Mind, Language and Society: Philosophy in the Real World.* New York: Basic Books.

Stryker, Cole, and Eda Kavlokoglu. 2024. "What is artificial intelligence (AI)?"

Whitehead, Alfred North. 1911. *An Introduction to Mathematics.* Oxford University Press.