Noname manuscript No.
(will be inserted by the editor)

Ethical Issues for Autonomous Trading Agents

Michael P. Wellman · Uday Rajan

Received: date / Accepted: date

Abstract The rapid advancement of algorithmic trading has demonstrated the success of AI automation, as well as gaps in our understanding of the implications of this technology proliferation. We explore ethical issues in the context of autonomous trading agents, both to address problems in this domain and as a case study for regulating autonomous agents more generally. We argue that increasingly competent trading agents will be capable of initiative at wider levels, necessitating clarification of ethical and legal boundaries, and corresponding development of norms and enforcement capability.

Keywords Autonomous Agents · AI Ethics · Financial Trading

M. P. Wellman
University of Michigan, Ann Arbor, MI, USA
E-mail: wellman@umich.edu

U. Rajan
University of Michigan, Ann Arbor, MI, USA
E-mail: urajan@umich.edu

1 Introduction

Whereas forecasts about the arrival of superintelligence vary widely,[1] autonomous agents are here today, and are likely to become significantly more pervasive and important in the near future. By *autonomous agents*, we refer to computational entities that make decisions and execute actions in response to environmental conditions, without direct control by humans. Although today's autonomous agents operate with relatively narrow scope of competence and autonomy, they nevertheless take actions with consequences for people. Maintaining human control over these agents is thus imperative, and reconciling immediate autonomy with ultimate human authority raises technically challenging problems [Bostrom, 2014, Russell et al., 2015]. Solving the problems of agent regulation in the context of specific domains addresses current social concerns, and may also provide experience informing the ultimate solution of control problems for AI more broadly.

1.1 Algorithmic Trading as a Case Study

One of the first domains where autonomous agents have become ubiquitous is trading in financial markets. Precise figures on the share of trading conducted by agents (in this context often called *trading bots* or *algos*) are not available, but most estimates attribute to algorithms over half of trading volume in US equities. Trading in other financial securities (such as currencies and fixed income assets) and in other jurisdictions is also increasingly automated.

Financial markets have proved fertile ground for autonomous agents for a variety of reasons. The markets themselves now operate almost entirely electronically, over networks with relatively well-scoped and well-defined interfaces. Markets generate huge volumes of data at high velocity, which require algorithms to digest and assess state. The dynamism of markets means that timely response to information is critical, providing a strong incentive to take slow humans out of the decision loop. Finally, and perhaps most obviously, the rewards available for effective trading decisions are large, enabling a commensurate devotion of resources toward talent and effort to develop and analyze technically sophisticated strategies.

The domain of financial markets also provides examples of autonomous agents out of control. A well-known instance is that of Knight Capital Group in 2012. As documented by the US Securities and Exchange Commission (SEC), during the first 45 minutes of the trading day on 1 August 2012, while processing 212 small orders from customers, an automated trading agent developed by and operating on behalf of Knight Capital erroneously submitted millions of orders to the equity markets Securities and Exchange Commission [2013]. Over four million transactions were executed in the financial markets as a result, leading to billions of dollars in net long and short

[1] It is easy to find statements predicting a singularity around the corner [Kurzweil, 2006], as well as those denying its inevitability [Walsh, 2016] or even the possibility of ever achieving human-level AI. Most expert opinion considers superintelligence plausible this century, with significant disagreement about whether many humans alive today will meet machines exceeding their intelligence across the board [Müller and Bostrom, 2016].

positions.[2] The errant orders were generated due to an accidentally misconfigured software upgrade, which allowed invocation of some obsolete code which repeatedly submitted orders without recognizing they had already been filled. The company lost $460 million on the unintended trades (over $10 million per minute that the code was operational), and the value of its own stock fell by almost 75%. It was acquired by a rival trading firm, Getco LLC, a few months later.

Software errors are of course not unusual, though in an autonomous agent they may carry especially great potential for damage. In this instance, the harm accrued almost entirely to the owner of the trading agent. More generally, an out-of-control trading agent could destabilize markets or otherwise harm innocent parties, and indeed the SEC (the applicable government regulator) sanctioned Knight Capital for violating risk management requirements.

One might also discount this example as an accident, preventable through improved software quality procedures. Perfect prevention is unlikely, but more to the point, market participants and regulators also need to be concerned with autonomous trading behavior that is not accidental. The effects on markets of autonomous agents is at present unclear, in part because the factors at play are unprecedented. The introduction of autonomous agents to financial markets generates several important new phenomena, not present in human-only trading.

1. Trading agents can *respond to information much faster than human reaction time*. The unprecedented speed renders details of internal market operations—especially the structure of communication channels and distribution of information—systematically relevant to market performance. In particular, the latencies between market events (transactions, price updates, order submissions) and when various actors find out about these events become pivotal, and even the smallest differential latency can significantly affect trading outcomes.
2. The *autonomy and adaptivity* of algorithmic trading strategies makes it challenging to understand how they will perform in unanticipated circumstances. The challenges are exacerbated by the increasing use of sophisticated machine learning techniques to generate and tune trading strategies [Kearns and Nevmyvaka, 2013], and the fundamental multiagent nature of the execution environment.
3. The automated nature of algorithmic trading makes it easy to *replicate and operate at scale*. Once one develops an algorithmic trading technique, that method can be immediately applied to trading many securities on a wide variety of trading venues.

Naturally, these factors go together, as autonomy is necessary for operation at superhuman speed and massive scalability. Some issues, such as interactions among adaptive and data-driven strategies, apply to algorithmic trading even if not conducted at high frequency [Easley et al., 2012].

[2] A *long* position is created when a trader buys a security, generally expecting to sell it later at a higher price. A *short* position is created when a trader sells a security in anticipation that its price will fall, planning to profit in buying it back later at a lower price.

1.2 Ethics for Algos

The regulation of autonomous agents ultimately relies a great deal on legal frameworks and institutions (including governmental) that can implement and enforce rules of behavior, as established through a political process. Establishing laws and policies is a cumbersome and often slow process, however, and may be expected to lag behind the pace of technological development. Generally accepted norms and ethical principles can also serve important functions in regulating behavior: filling in gaps that official rules do not cover, providing guidelines for interpreting conditions specified in existing rules, and characterizing practices that may eventually be ratified in new rules.

Trading algorithms are programmed by people, so naturally the programmers (individuals and their organizations) are responsible for complying with applicable laws and ethical standards.[3] The autonomy of trading agents does not absolve their masters of accountability, however, the indirection of decision making does present some tricky issues. For example, autonomous agents may perform actions—particularly in unusual circumstances—that would have been difficult to anticipate by their programmers. Does that difficulty mitigate responsibility to any degree? Presumably, the likelihood of encountering novel situations is something that itself should have been anticipated and accounted for in the design of the autonomous agent. Regardless of how accountability lines are drawn, market participants and regulators will need ways to assess predictability of trading agent behavior, as well as standards for deploying these agents in environments where unpredictable deleterious behaviors are likely to arise.

Some argue that autonomous agent should have explicit facilities for ethical reasoning [Wallach and Allen, 2009], which would be especially salient in unanticipated circumstances. Tonkens [2009] points out that implementing ethical reasoning may be problematic for autonomous agents, as they may come to believe that their existence itself is unethical. Such concerns should be avoidable for agents with limited scope of autonomy. Designing autonomous agents based on utilitarian calculations presents additional dilemmas, due to conflicts between the agent's client and the multiagent system where it operates [Bonnefon et al., 2016]. Yampolskiy [2013] makes the case against even trying to resolve such philosophical problems, arguing that it will be more effective, particularly in developmental stages, to focus on designing agents for safety, preventing them from taking actions that are potentially very harmful.

Autonomy also complicates questions of lawful or ethical behavior that hinge at all on intent. For example, financial regulations generally proscribe *market manipulation*, which is often defined in terms of the intent behind market actions. If these actions are taken by autonomous agents, do regulators need to determine whether the action was intended by the agent to have manipulative effects, or whether the programmer intended the agent to take such actions for such purposes? Examining

[3] As Davis et al. [2013] point out, the ethical responsibilities of traders, computer engineers, and quantitative analysts are each determined by separate professional organizations, creating a need for an organizational-level understanding of standards.

these distinct propositions may yield different answers, or the available evidence may support an answer to one but not the other.

In the rest of this paper we explore such issues for autonomous trading agents. We do not offer broad resolutions to the fundamental questions, but propose a framework for considering the possibilities. Our framework posits an ambitiously broad architecture for trading agents (broader than we expect current trading agents employ), based on automated discovery and exploitation of arbitrage opportunities. The framework allows us to characterize distinct levels of initiative, which might provide plausible ethical boundaries. We illustrate its application through a discussion of market manipulation and intent.

2 An Arbitrage-Based Framework for Reasoning about AI Traders

In an *arbitrage* operation, a trader takes advantage of a discrepancy in prices for an asset across multiple markets, in order to achieve a near-certain profit. The concept of arbitrage is central in finance theory, which commonly takes the absence of arbitrage as a general criterion for market efficiency [Varian, 1987]. It also turns out that a broad range of automated trading behaviors can be viewed as seeking out and exploiting arbitrage opportunities. In this section, we illustrate a variety of forms of arbitrage, and sketch a generalized AI trading architecture—the ARB-BOT—based on the search for arbitrage. We then show how to characterize ethical boundaries in trading behavior in terms of the scope of this search.

2.1 Arbitrage and Identities

Consider a situation where a good can be sold in one market at a price higher than it can be bought at another. This meets the very definition of an arbitrage opportunity: exploiting a price discrepancy across markets.

In general, an arbitrage in the financial markets may involve a combination of transactions, which together yield no net change in an agent's effective asset position. For example, one of the most pervasive forms of arbitrage is based on *index securities*. An index security is defined by equivalence to a bundle of underlying assets. The Standard & Poor's Depository Receipt (SPDR), for instance, is an index security that tracks the S&P 500 index, which in turn is a capitalization-weighted average of stock prices of 500 of the largest US public corporations. For an investor, buying or selling a share in the SPDR is as straightforward as trading a share in a company such as Apple.

In practice, one SPDR share is valued at very close to 1/10 of the value of the S&P 500 index; for discussion purposes, let us assume this ratio is exact. Then, in principle one could buy ten SPDR shares and effectively own a (non-integer) number of shares of each of the S&P 500 companies, corresponding to their weight in the index. By selling these exact numbers of shares (in 500 separate transactions), one could then return to a neutral asset position.

We can formalize this concept of arbitrage as follows. Let X be a set of goods (e.g., financial securities), and $\mathbf{x}$ a transaction vector, with x_i the quantity purchased

(or if negative, sold) of good i. We refer to $\mathbf{x}$ as a *position-neutral transaction* if $\mathbf{x} \neq \mathbf{0}$ and executing trades for the specified quantities would be neutral with respect to the trader's net asset position.

For example, if goods 1 and 2 are the same (e.g., a security tradeable on different exchanges), then $(1, -1)$ is a position-neutral transaction, as is any vector (x_1, x_2) such that $x_1 \neq 0$ and $x_1 + x_2 = 0$. For the SPDR index example, let goods $1, \ldots, 500$ denote the stocks of the S&P 500, and let good 501 denote the SPDR security. Further let n_i denote the number of shares of stock i in a portfolio with the same stock weights as the S&P 500 index, and with a value equal to the index. Then $(-n_1, \ldots, -n_{500}, 10)$ is a position-neutral transaction, as is any positive or negative multiple of that vector.

At the heart of any arbitrage situation is an *identity relation*, characterizing the equivalence of goods or their combination. Suppose that a unit of good i can be bought or sold at the same price p_i. Consider a transaction $\mathbf{x} = (\ldots, x_i, \ldots)$, and divide it into a *buy vector* $\mathbf{x}^+$ and *sell vector* $\mathbf{x}^-$.

$$\mathbf{x}^+ \equiv (\ldots, \max(x_i, 0), \ldots)$$
$$\mathbf{x}^- \equiv (\ldots, \min(x_i, 0), \ldots)$$

If $\mathbf{x}$ is a position-neutral transaction, then $\mathbf{x}^+$ is essentially equivalent to $-\mathbf{x}^-$, in that buying $\mathbf{x}^+$ and selling $-\mathbf{x}^-$ is net neutral on asset position. An arbitrage opportunity would exist if the price of buying the vector $\mathbf{x}^+$ is different from the amount obtained on selling the vector $\mathbf{x}^-$.

More formally, let $\mathbf{p}$ be a vector of prices, one for each good. Then, an *arbitrage opportunity* exists at prices $\mathbf{p}$ if and only if $\mathbf{p} \cdot \mathbf{x} > 0$. Transacting at the arbitrage opportunity represents pure gain, obtaining a strictly positive monetary payment with no effective change in asset position. Note that such a transaction requires simultaneous trading in the assets that comprise the vector $\mathbf{x}$. For example, to reap such gains in the SPDR domain, a trader monitors the 501 relevant prices, and simultaneously issues orders to trade in several securities whenever the combined execution would be profitable.

Before jumping at apparent opportunities like this, however, a trader needs to account for several factors, two of which are generically important in financial markets. First, there are *transaction costs*: trading may incur commissions or fees, or other fixed or variable costs of maintaining trading operations (computational resources, information access, communications, etc.). Adjusting the profit condition to account for transaction costs is relatively straightforward. Second, there is *execution risk*: the chance that available prices may change between the triggering observation and the transaction itself. Execution risk is a factor whenever there is latency between price information and the trade initiation (i.e., virtually always), and is therefore a significant driver of the race to reduce latency. Properly accounting for execution risk would entail assessing the probabilities and consequences of short-term price movements that threaten the arbitrage situation. In principle a trader could model these and apply the profitability condition with respect to expected returns. In practice, arbitrage algorithms often address both transaction costs and execution risk simply by adding a threshold gross profit margin for triggering the arbitrage operation.

As another example, consider currency trading, where the identity relation is described by exchange rates. Let $\rho_{\$€}$ denote the exchange rate between dollars and euros. That is, the price for obtaining one euro is $\rho_{\$€}$ dollars. Similarly, let $\rho_{€¥}$ denotes the price of yen in euros. Then, having one yen is effectively like having $\rho_{€¥}$ euros. Finally, let $\rho_{¥\$}$ denote the price of dollars in yen, which also means that the price of yen in dollars is $1/\rho_{¥\$}$. If at any point $\rho_{¥\$}\rho_{\$€}\rho_{€¥} < 1$, an arbitrageur can profit by purchasing $\rho_{€¥}$ euros for $\rho_{\$€}\rho_{€¥}$ dollars. These euros are equivalent to one yen, which can be sold at a profit for $1/\rho_{¥\$}$ dollars. In our transaction notation, with goods $(€, ¥)$, the arbitrage transaction is $(\rho_{€¥}, -1)$ with price vector $(\rho_{\$€}, 1/\rho_{¥\$})$. Under the trigger condition above, this transaction is profitable. Of course, the qualifications about transaction costs (typically expressed as a spread between buy and sell exchange rates) and execution risk still apply. This reasoning can be extended to an arbitrary number of currencies related by pairwise exchange rates, and arbitrage opportunities can be identified in a computationally efficient manner by applying variants of shortest-path algorithms to the exchange-rate graph.

Futures markets may also enable arbitrage, in this case through transactions across time. A futures contract is an agreement to trade a good at a specified time in the future, at a price agreed upon today. For example, consider a contract to deliver a good at a certain price one year from now. We conceptualize this contract as two distinct goods, one the spot (i.e., delivered now) version and the other a future (one-year-later) version. If the future price exceeds $1 + r$ times the current (spot) price, where r is the interest rate over the next year, then an arbitrageur can borrow funds to buy the good now, and sell the good a year later. In one year, after paying off the loan, the transaction delivers a profit. As above, any extra costs such as that of storing the good pending delivery must be accounted for. Here the identity relation is that having the good now plus storing it is equivalent to having it in the future.

2.2 Arbitrage Agents

The search for an arbitrage opportunity, therefore, involves searching for the violation of a given identity relationship in asset prices at a particular point of time. As is clear from the examples above, there are many such identity or near-identity relationships that should hold in efficient financial markets, and some of these relationships can involve as many as hundreds of different assets. The problem of constantly searching for possible deviations is therefore ideally suited to automation, as computers are adept at monitoring large streams of data and verifying well-specified conditions. After a deviation is found, the requisite buy or sell orders for each asset need to be submitted to the financial markets in a very short period of time (before prices move and the arbitrage vanishes). There are therefore substantial gains to automating order submission as well.

Index arbitrage is a perfect example where automation of trading strategies is necessary and sufficient for effective performance. It is necessary because the instantaneous valuation of the portfolio that comprises an index, with interest and dividend adjustments to be made, is too complex for rapid and reliable human calculation. Moreover, executing a simultaneous trade of the index contract and all of the un-

derlying securities requires automated order submission through electronic market interfaces to achieve the speed associated with acceptable levels of execution risk. Arbitrage agents are sufficient to identify opportunities for index arbitrage because the formula to determine the value of an index is straightforward, and computationally simple given the prices of the underlying securities. As a result, index arbitrage accounts for a significant segment of trading on US equity markets. The New York Stock Exchange reports that the fraction of transaction volume attributed to program trading (which it defines as transactions involving 15 or more simultaneous securities and with a minimum dollar volume) is well over a half, and much of that is attributable to index arbitrage specifically.

Autonomous trading agents can be very effective at finding arbitrage opportunities. They pick up on the smallest deviations from the identity relationship, and do so very quickly. Because computers are cheap, there has been a substantial increase in the use of such autonomous agents over time. The end result is that arbitrage opportunities are rendered slight and scarce. In consequence, the leading edge in arbitrage strategy is to spread out in two ways. First, the search for an arbitrage opportunity is extended to more complicated identity relationships that involve larger sets of assets. Second, the very notion of identity is stretched to *statistical arbitrage*, that is, to relationships among asset prices that hold probabilistically (as induced by historical observation) but not by definition. Autonomous agents are well-suited to identifying patterns in past data that have led to particular trades being profitable. Statistical arbitrage represents a major category of automated trading strategies in financial markets.

2.3 ARB-BOT

The arbitrage automation discussed in the preceding section is limited to the execution of predefined arbitrage transactions, based on monitoring of prices in preidentified markets. A more ambitious automation of arbitrage would extend to the construction of arbitrage transactions, essentially generating new arbitrage strategies based on market reasoning and observation. For example, if an autonomous agent finds a statistical arbitrage, it may then also be able to determine what kinds of market conditions lead to the data patterns that imply the profitability of a given trade, and can try and induce those conditions in the market. We sketch a general architecture for this kind of arbitrage trading automation, which we call the ARB-BOT. Having such an architecture enables us to discuss more concretely the possible behaviors of sophisticated trading agents, in a broader set of conditions than typically contemplated.

The primary operating mode of ARB-BOT is seeking out arbitrage opportunities, and developing trading strategies that exploit them. As argued in Section 2.1, arbitrage is inevitably associated with an underlying identity relation (possibly approximate or stochastic) among goods. Finding an arbitrage opportunity therefore reduces to constructing an identity relation that can be applied across a set of disparate markets. We consider two approaches to constructing identities, described in sections below. Given the defining identity, developing an arbitrage trading strategy further requires models of transaction costs and execution risk (perhaps through ex-

plicit automated experimentation), and design or learning [Kearns and Nevmyvaka, 2013] of trading rules optimized for those models.

2.3.1 Reasoning about security descriptions

Once concepts like exchange rates, futures, options, indices, interest, and other standard derivative constructs have been formalized, defining inference rules to generate or verify identities is conceptually straightforward. A key prerequisite for recognizing instances of standard arbitrage patterns is that the goods be labeled sufficiently to draw the connections across markets. For example, an index security must be linked to the constituent goods defining the bundle. Similarly, a futures contract must link to the spot-market good it concerns. If these links are in place, the corresponding arbitrage operation follows directly from rules already well codified in the standard texts on derivative securities [Hull, 2000].

Indeed, generation of arbitrage identities in standardized markets is already routinely applied. For example, much options trading is based on the Black-Scholes pricing formula, which is itself derived from arbitrage reasoning [Black and Scholes, 1973]. Options exchanges are structured so that it is easy to extract the defining information about an options security (underlying spot security, strike date, strike price), which with some additional parameters enables calculation of the Black-Scholes price. Given the boilerplate extraction of the needed information from the security description, we would probably not even call this "reasoning."

More complex cases, however, may benefit from automation that would clearly qualify as reasoning. In some cases a financial security is defined as a composition of constructs. The S&P 500 index mentioned above is an index or basket of 500 stocks. Derivative securities such as futures contracts can be traded on this basket. When considering such index futures, arbitrage analysis must account for both the underlying security prices and the interest rate between now and the time when the transaction will take place. It is also necessary to factor in dividend payments [Hull, 2000]. Similarly, it is not uncommon to add idiosyncratic option features to contracts, or define complex derivatives that bring together combinations of existing securities in ad hoc ways.

More fundamentally, any financial security can be described in terms of the stream of cash flows it provides across time, contingent on states of nature [Duffie, 1992]. Given such descriptions for securities, it is likewise possible in principle to verify identity relations: vectors of financial goods are equivalent if they provide the same aggregate income streams in every state of nature. Moreover, for any non-identity, this kind of description provides a representation for the residual difference, in the form of a conceivable security that could be sought or constructed.

Given a database of security descriptions, there are many ways to organize a search for arbitrage identities, exact or approximate. Search procedures should exploit explicit links (e.g., between index securities and their constituents) when they exist, but also have a means for bottom-up generation of candidate combinations.

This search problem has not been generally formulated or tackled to our knowledge to date,[4] but we consider it a plausible approach to developing ARB-BOT.

2.3.2 Automated discovery by machine learning

Having admitted the applicability of statistical arbitrage and other forms of approximate identity, it makes sense to consider non-deductive means to evaluate arbitrage opportunities. Markets and especially financial markets generate huge volumes of price information, which can be exploited to find identities based on statistical evidence. Indeed, some data mining and machine learning approaches to algorithmic trading can be viewed as implicitly identifying arbitrage relations. For example, Gatev et al. [2006] apply simple rules to find pairs of stocks that tend to move in tandem, which can then be exploited in a *pairs trading* statistical arbitrage method.

Making the arbitrage identification explicit may focus the development of machine learning algorithms. The statistical approach addresses a more open-ended computational problem than the reasoning approach described above, but the two are complementary with respect to the overall ARB-BOT architecture.

3 Levels of ARB-BOT Initiative

The description of arbitrage above focuses on search for profit-making opportunities that exist independently from anything that ARB-BOT does. An autonomous trading agent that limits its behavior to such search is essentially *passive*, and typically benign as discussed above. Of greater concern is that the ARB-BOT might take *initiative* to leverage its capabilities, by causing the generation of arbitrage opportunities that otherwise might not exist.

Consider the following levels of increasingly proactive ARB-BOT behavior.

1. Passive search for arbitrage opportunities.
2. Attempts to instigate arbitrage opportunities via existing channels, through purposeful instigation of market movements, for example through spoofing or other forms of market manipulation.
3. Attempts to create new arbitrage channels, for example through introduction of new (possibly redundant) financial instruments, or deliberate fragmenting of markets.
4. Malicious actions to subvert markets: for example propagating misinformation, obtaining improper access to information, or direct violation of market rules.

Our description of ARB-BOT in Section 2.3 above is limited to the first level. However, algorithmic trading can and does operate at all of these levels, and the higher levels in particular may entail significant risk. In principle, the automated search for arbitrage opportunities can include the aggressive actions defining levels 2–4.

[4] Schuldenzucker [2016] proposes a related theorem-proving approach, where starting from contract descriptions expressed in a formal logic, the prover uses no-arbitrage principles to derive inequality relations on security prices. If the inequalities are violated in the market, then an arbitrage opportunity exists.

3.1 Level 1: Passive Arbitrage Search

As Fama [1970] observes, in an ideal capital market, the prices of financial assets reflect all available information, thereby allowing for better resource allocation across the economy. Persistent mispricing reflects a market inefficiency, so that financial arbitrage has the beneficial effect of aligning prices with fundamental values. Automating arbitrage in such instances is thus also beneficial, as it reduces the length of time for which such mispricing can exist. However, there may be no additional benefit from exploiting arbitrage situations that would have resolved in milliseconds anyway, without need for the ARB-BOT. In such situations, instantaneous exercise of arbitrage opportunities can even be counterproductive in terms of the overall welfare of participants. For example, Wah and Wellman [2013] find that the high-frequency trading practice of latency arbitrage between fragmented markets can reduce total surplus in the market. In addition, the practice may engender a costly latency arms race [Budish et al., 2015].

High-frequency traders often act as market makers. The practice of *market making*—providing liquidity to a market by simultaneously posting buy and sell orders—can be viewed as a form of statistical arbitrage across time. Market making is often socially beneficial [Wah and Wellman, 2015], but without sufficient competition among market makers, may detract from the surplus of patient traders.

3.2 Level 2: Market Manipulation

The SEC defines *manipulation* as "intentional conduct designed to deceive investors by controlling or artificially affecting the market for a security". Key conditions in this definition are express intent, and the effect of misleading others about market conditions. Intent is difficult to attribute, as noted above. Defining what it means to mislead is also quite tricky, for example, hiding information (e.g., about one's own values or demand) is usually considered a perfectly reasonable and ethical tactic in negotiation and trading. Clearly, to constitute manipulation a behavior must cross a higher line involving taking affirmative action to shape false signals for others.

Kyle and Viswanathan [2008] argue that the focus should be on distinguishing socially harmful kinds of manipulation from those that merely give one agent advantage over another. They propose more elaborate and stringent criteria, which require that a manipulation compromises pricing accuracy as well as market liquidity.

One special case of manipulation is *spoofing*, which is defined in the 2010 Dodd-Frank Act, §747, as "bidding or offering with the intent to cancel the bid or offer before execution". That is, a spoofing strategy places orders with the intention not to trade, but rather to falsely signal demand or supply. The spoof orders are typically priced just outside the current price quotes, and withdrawn with high probability before any market movement could trigger a trade.

A notorious case of spoofing commodity markets was perpetrated by Michael Coscia, convicted in November 2015 [Louis and Hanna, 2015]. Coscia employed a strategy called *dynamic layering*, which involved placing and withdrawing orders placed at prices away from the best buying and selling prices available in the market,

to mislead other market participants and improve conditions for real trades. In preliminary work [Wang and Wellman, 2017], we have reproduced a similar spoofing strategy in simulation, and found that it can effectively move prices in an environment where other agents make use of order book information. The presence of a spoofer degrades welfare (meets the criteria of Kyle and Viswanathan [2008] for a harmful manipulation), and when agents anticipate spoofing they will make less use of the order book information.

This form of spoofing manipulation is applied through the order stream of an exchange. Manipulators may also employ broader channels, including social media, traditional media, or direct communication (e.g., pump-and-dump schemes). For example, about one year ago a manipulator filed a fake SEC EDGAR report of a takeover, successfully boosting Avon stock [Goldstein and Gelles, 2015] long enough to earn a small profit.[5]

Market manipulation in all these forms is illegal, though legal scholars have noted the difficulty of precisely characterizing illegal manipulation [Ledgerwood and Carpenter, 2012]. Presumably the illegality of spoofing activity is unaffected by whether one implements the intended manipulation strategy manually or with assistance of an automated trading agent. But what if the spoofing behavior is automatically synthesized by an agent with the sophistication of ARB-BOT? Attributing intent to an automatically synthesized strategy is a dicey prospect, as there is no observable indicator of the rationale behind the strategy's design. The ARB-BOT's master may have given it only the objective of earning profits; manipulation is an intermediate means to that end.

The difficulty arises because any of the individual actions comprising the proscribed behavior can be rationalized on other grounds. For example, a spoofing strategy will typically submit misleading orders, canceling them before they actually trade. However, large numbers of order cancellations are also characteristic of legitimate strategies [Aldridge, 2013], including potentially beneficial behavior such as market making. Pending significant advances in detection technology, our ability to define improper trading agent behaviors exceeds our ability to prevent them or even to identify and prosecute them after the fact.

3.3 Level 3: Artificial Arbitrage

An agent that is particularly capable of exploiting arbitrage opportunities has a large incentive to expand these opportunities. Recall from Section 2.1 that arbitrage is inevitably based on an identity relation. In principle, one can create new identities by cloning objects, creating new multiplicities of items that are the same (or the same after applying some transform). If these items are traded on separate markets, there are now correspondingly more market opportunities. Since financial securities are essentially informational objects—contracts associating financial returns with economic

[5] The Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system maintained by the SEC is a system for filing (and subsequent retrieval) of electronic forms by public companies in the US. The database is freely available to the public, including investors (`www.sec.gov/everythingedgar`).

events—they are relatively easy to clone and transform. Similarly, markets themselves are computational entities (mechanisms that map messages to trades), and can be constructed simply from computational resources and information.

Given operators for creating new financial securities or markets, ARB-BOT can in principle multiply its arbitrage opportunities. Though doing so intuitively corresponds to a high level of initiative, in general such activities are not illegal or sanctioned in any way. Historically, creating new financial instruments and exchanges has even been encouraged in the name of competition. The degree of fragmentation seen in US equity markets today is now seen as excessive, and an unintended consequence of regulation changes designed to promote such competition. How to draw the line between creative design of financial innovations on the one hand, and gratuitous cloning to obfuscate and generate arbitrage wedges is an open problem, one that may merit more serious attention if this level of activity is successfully automated.

3.4 Level 4: Malicious Action

In the science fiction novel *The Fear Index* [Harris, 2011], an AI algorithm figures out that it can induce large market price swings (and thus a bounty of arbitrage opportunities) by taking real-world violent actions that create fear and uncertainty. This is obviously an extreme instance of level-4 activity, but the logic is not really different than information attacks that could conceivably be executed by sophisticated algorithms. Information attacks by humans that move markets occur regularly, for example the April 2013 hacked AP Twitter account reporting a White House bombing. These are arguably different only in degree from the manipulations of level 2, though perhaps one could draw boundaries based on the relation of the misinformation to fundamental security concerns.

Other forms of malicious behavior would also be included in level 4, such as denial-of-service or other cyber-attacks on electronic markets, presentation of false credentials, or corruption of critical records. Such descriptions render obvious the unethical and (usually) illegal nature of the activities, though no doubt at the boundaries some clarifications will be required in order to maintain financial integrity in the face of autonomous trading agents adept at finding and exploiting loopholes.

4 Controlling ARB-BOT

Who should control ARB-BOT, and how? Responsibility for maintaining a well-functioning market rests with trading companies that may design such agents, other market participants, self-regulatory organizations such as the Financial Industry Regulatory Authority (FINRA), stock exchanges, and regulators such as the SEC. Each of these parties therefore has an incentive in controlling the behavior of an autonomous trading agent such as ARB-BOT.

The levels of ARB-BOT initiative are summarized in Figure 1. The primary value of defining these levels is that they provide a natural place to set boundaries on behavior. Level 1 behavior (passive arbitrage) is usually beneficial or neutral. It may

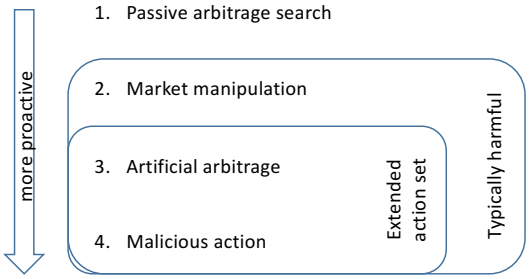

Fig. 1 Levels of trading agent initiative.

be possible to better characterize the situations where it can be harmful, and if that can be accomplished may be a place to consider refined distinctions. Until then, we advise focusing ethical and regulatory attention on levels 2–4, which are typically harmful and easier to set apart from beneficial level 1 behavior. Note that current legal proscriptions cover (or attempt to cover) levels 2 and 4, leaving qualitatively lighter regulation around levels 1 and 3.

If the action space available to ARB-BOT can be limited, an organization seeking to control ARB-BOT can reasonably prevent it from crossing into level 3 or level 4. Both generally require actions that go beyond the typical market operations of querying for information and placing orders. Therefore even an algorithm that is highly effective in finding profitable strategies based on such primitives will not produce behavior at the two highest levels defined here.

The same is not true at level 2. A designer of ARB-BOT can build an agent without an explicit strategy to construct false signals to mislead others, but cannot prevent the agent from stumbling onto such strategies. Any actions create signals, and those signals will affect others' actions. An agent, particularly one learning from real or simulated observations, may learn to generate signals that effectively mislead.

To deal with unintentional unethical behavior in agents controlled by designers or regulatory organizations, they may have to be scrutinized in the same way that agents developed by a third party would be. This means monitoring them and looking for patterns that are characteristic of unethical behavior. Some patterns may become known through experience with malicious agents (human or algorithm). To find out about others, a regulatory organization may have to devote effort to develop strategies that cross the boundaries, and experiment with them in the laboratory to understand their characteristic signatures. This approach will not produce perfect confidence in detecting unethical behavior, but it may contribute to that capacity.

When it comes to regulating the behavior of autonomous trading agents, one can also imagine a role for third-party monitors or certifiers. An example that comes to mind is something analogous to credit rating companies (which provide a rating to public debt issues) for autonomous trading agents. Alternatively, if regulators or self-regulatory organizations require trading entities to obtain insurance against the misbehavior of autonomous agents, insurance companies could perform this risk assessment role.

5 Conclusion

Because the financial world is essentially built on information, autonomous agents are proliferating rapidly in this domain. Much of this is for the good, reducing transaction costs and making markets more efficient. The full ramifications are poorly understood, however, and there is ample reason for concerns that financial markets are vulnerable to agent misbehavior, whether accidental or purposeful, legal or illegal, ethical or unethical.

We have attempted to lay out in this essay some issues presented by the invasion of autonomous trading agents in financial markets, both specific to the financial domain and as a case study for autonomous agents in general. Our contribution so far is mainly to raise questions about how to map ethical and legal concepts delineating acceptable trading behavior from the human to computational realm. In particular, the reliance on intent to distinguish legitimate from illegitimate actions may be particularly challenging to apply to automated activity.

We have also proposed a framework for thinking about autonomous trading capabilities, employing the idea of searching for arbitrage opportunities. Defining the problem at that level of generality suggests that autonomous agents may become capable of operating at high degrees of initiative, which could present increasingly serious concerns as technology develops. Finding effective solutions to the regulation of autonomous agents in this domain is important in its own right, and may also prove illuminating for addressing the broader problem of AI control.

References

I. Aldridge. Market microstructure and the risks of high-frequency trading. Technical report, ABLE Alpha Trading, 2013. Available at SSRN, 2294526.

F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81:637–654, 1973.

J.-F. Bonnefon, A. Shariff, and I. Rahwan. The social dilemma of autonomous vehicles. *Science*, 352:1573–1576, 2016.

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

E. Budish, P. Cramton, and J. Shim. The high-frequency trading arms race: Frequent batch auctions as a market design response. *Quarterly Journal of Economics*, 130:1547–1621, 2015.

M. Davis, A. Kumiega, and B. V. Vliet. Ethics, finance, and automation: A preliminary survey of problems in high frequency trading. *Science and Engineering Ethics*, 19:851–874, 2013.

D. Duffie. *Dynamic Asset Pricing Theory*. Princeton University Press, 1992.

D. Easley, M. M. López de Prado, and M. O'Hara. The volume clock: Insights into the high frequency paradigm. *Journal of Portfolio Management*, 39(1):19–29, 2012.

E. F. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417, 1970.

E. Gatev, W. N. Goetzmann, and K. G. Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, 19:797–827, 2006.

M. Goldstein and D. Gelles. A phantom offer sends Avon's shares surging. *New York Times*, 14 May 2015.

R. Harris. *The Fear Index*. Hutchinson, 2011.

J. C. Hull. *Options, Futures, and Other Derivatives*. Prentice Hall, fourth edition, 2000.

M. Kearns and Y. Nevmyvaka. Machine learning for market microstructure and high frequency trading. In *High Frequency Trading: New Realities for Traders, Markets and Regulators*. Risk Books, 2013.

R. Kurzweil. *The Singularity Is Near: When Humans Transcend Biology*. Penguin, 2006.

A. S. Kyle and S. Viswanathan. How to define illegal price manipulation. *American Economic Review*, 98:274–279, 2008.

S. Ledgerwood and P. R. Carpenter. A framework for the analysis of market manipulation. *Review of Law and Economics*, 8:253–295, 2012.

B. Louis and J. Hanna. Swift guilty verdict in spoofing trial may fuel new prosecutions in U.S. *Bloomberg News*, 3 November 2015.

V. C. Müller and N. Bostrom. Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller, editor, *Fundamental Issues of Artificial Intelligence*. Springer, 2016.

S. Russell, D. Dewey, and M. Tegmark. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4):105–114, 2015.

S. Schuldenzucker. An axiomatic framework for no-arbitrage relationships in financial derivatives markets. Technical report, University of Zurich, 2016.

Securities and Exchange Commission. In the matter of Knight Capital Americas LLC, order instituting administrative and cease-and-desist proceedings. Release 70694, Oct. 2013.

R. Tonkens. A challenge for machine ethics. *Minds and Machines*, 19:421–438, 2009.

H. R. Varian. The arbitrage principle in financial economics. *Journal of Economic Perspectives*, 1(2):55–72, 1987.

E. Wah and M. P. Wellman. Latency arbitrage, market fragmentation, and efficiency: A two-market model. In *Fourteenth ACM Conference on Electronic Commerce*, pages 855–872, 2013.

E. Wah and M. P. Wellman. Welfare effects of market making in continuous double auctions. In *Fourteenth International Conference on Autonomous Agents and Multi-Agent Systems*, pages 57–66, 2015.

W. Wallach and C. Allen. *Moral Machines*. Oxford University Press, 2009.

T. Walsh. The singularity may never be near. In *IJCAI-16 Workshop on Ethics for Artificial Intelligence*, 2016.

X. Wang and M. P. Wellman. Spoofing the limit order book: An agent-based model. In *AAAI-17 Workshop on Computer Poker and Imperfect Information Games*, 2017.

R. V. Yampolskiy. Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In V. C. Müller, editor, *Philosophy and Theory of Artificial*

Intelligence, pages 389–396. Springer, 2013.