Chapter 21
An Autonomy Interrogative

Darryn J. Reid

21.1 Introduction

This paper examines the highly topical imperative regarding the development and application of autonomous systems from primarily an economic perspective, meaning taking autonomous decision making to be a matter of allocating scarce resources [11, 12, 21, 29, 42]. It is no secret that despite intense international effort, the account of truly autonomous systems on hand for operational deployment in modern national defence organisations leaves something to be desired. Arguably, the outlook in autonomous systems development has been overwhelmingly dominated by technical developments, with the consequence that autonomy as a concept has become associated with increasing algorithmic and hardware sophistication. Yet given not only the technical challenges of machine intelligence, but also the difficulty of elucidating what autonomy means with respect to the kinds of operational military problems where we would like to apply it and the lack of a general concept of its utility in these circumstances, the present dearth of operational military autonomous systems ought not be so surprising. This chapter seeks to explore what is primarily an economic perspective towards autonomy, in order to inform subsequent choices of technical problems towards the realisation of operationally viable autonomous systems that solve difficult military operational problems well enough to be worthy of the title.

That many commercial enterprises possess what they hold to be autonomous systems poses the question as to why defence organisations cannot make the same claim. Yet if we regard that it takes more than to be able to act outside of direct human control to make a system autonomous, these systems may not really be autonomous at all. Specifically, they operate under highly constrained circumstances, and the predictability afforded by those operating circumstances are required to make them effective. Defence represents an especially challenging set of circumstances and difficult choices: unlike the problem contexts of extant autonomous systems, military operations, by and large, do not facilitate the luxury of tightly managing the opera-

D. J. Reid (✉)
Defence Science and Technology Group, Joint and Operations Analysis Division,
PO Box 1500, Edinburgh, SA, Australia
e-mail: darryn.reid@defence.gov.au

H. A. Abbass et al. (eds.), *Foundations of Trusted Autonomy*, Studies in Systems,
Decision and Control 117, https://doi.org/10.1007/978-3-319-64816-3_21

tional context so that the strong environmental expectations built into automations will not smash unhappily into nonconforming realities. Moreover, such failures in war and battle are potentially catastrophic. Given the fundamental uncertainty of this kind of environment [6, 53], efficiency goals that work in relatively controlled environments are largely irrelevant and may even be misleading, in the sense that the pursuit of efficiency could actually accentuate exposure to ruinous outcomes (the effect is well studied in economics, for instance this effect is understood to be a major contributing factor in the Global Financial Crisis of 2007–2008 [3, 26, 40]). Instead, operational effectiveness hinges on avoidance of irrecoverable failure, which might, incidentally, include avoiding gross inefficiency, but this is not the same thing in general as attempting to maximising efficiency.

I argue that heart of the problem of autonomy is the ability to effectively deal with fundamental uncertainty; such uncertainty is the consequence of the inherently paradoxical nature of the problems we would like our autonomous systems to be able to solve. For instance, prediction is generally a paradox of self-reference, because of the effect that making a prediction has on the predicted event; the predictions occur from within the system to which the predictions apply. Observation is similarly potentially paradoxical because the act of observing occurs from within the system and consequently may disturb the very phenomena we wish to observe. Clausewitz was well aware of these effects in war and battle, and consequently distinguished carefully between uncertainty that is unmeasurable and stochastic chance. Since the time of Clausewitz, there has been a burgeoning of interest in uncertainty, from a number of distinct perspectives. In particular, I wish herein to briefly draw on the work on uncertainty and ignorance in economics, which particularly originates in its more philosophically oriented branches, and some of the studies of incompleteness and uncomputability phenomena in mathematics and computer science [19, 35, 39, 45], and to relate the two together.

Newcomb's Dilemma [4] is a good example of how the kind of paradoxes behind mathematical incompleteness generate real-world uncertainty. This thought experiment involves a paradox of prediction that interestingly brings two distinct decision-making modes, which mostly seem to apply under different circumstances, into direct conflict in a single situation; the puzzle thus highlights the inapplicability of simple conventions of rationality as reward maximisation in settings involving fundamental uncertainty. The first decision-making mode is the impetus to act in order to produce a desired outcome, and the second is the impetus to act only when the action can alter the outcome. You are presented with two boxes: one open and one closed. The open box contains a thousand dollars, while the closed box contains either a million dollars or nothing. You can choose either the closed box or you can choose both boxes. The contents of the closed box has been prepared in advance by an oracle machine that almost perfectly predicts what you will choose: if you choose both boxes, then it has put nothing in the closed box, while if you choose just the closed box then it has place the million dollars in it.

The solution to the dilemma seems obvious to almost everyone; the catch is that people divide about evenly on which solution is the obviously correct one. In other words, the dilemma poses two incommensurate choices that are both justifiable. The

dilemma was recently resolved [54] using game theory to show that the course of action in the game depends on the observer's beliefs - which are necessary to making a choice but not rationally justifiable - about the ability of the oracle machine to predict their actions, in much the same manner as the observation of quantum states depends on the observer.

The connection from autonomy to economics rests on the proposition that an autonomous agent is exactly an economic agent, by another name. An autonomous agent—be it living, a social construct, or a machine—is an entity that exercises choice, by making decisions to act one way or another. Yet any decision to act in a particular way is just an allocation of resources under that agent's control to one of the course of action options as understood by the agent, and this matches the definition of an economic agent. To be slightly more formal: an economic agent is any actor (e.g. an individual, company, government body or other organisation) that makes decisions in aiming to solve some choice problem within some economic system, and an economic system is any system involving production, consumption and interchange of resources by such agents [11, 12, 21, 29, 42].

The difference between an autonomous agent and an economic one is a matter or emphasis that directs subsequent research problem choices. When we talk of artificial intelligence, it usually means we are mostly interested in developing algorithms and their technical implementations in machines; when we talk of economic agents, it means we are mainly concerned with individual and system-wide outcomes given different mixes of different kinds of interacting resource-allocating agents under various environmental conditions.

In particular, I am concerned with autonomy as particularly featuring decisions about the allocation of self under uncertainty, and hypothesise that self-allocation corresponds with the same boundary that delineates a notion of uncertainty that has come to be known as ontological uncertainty [13, 25, 30, 44, 50]. Interacting self-allocating agents entails logical paradox in general, which underpins formal limits on knowledge and thus the intrinsic uncertainty of such systems. This uncertainty arises without any recourse to exogenic 'shocks'. Hence familiar notions of agent utility-maximising rationality that operate within worlds of linear certainty and stochastic risk necessarily break down under conditions of fundamental uncertainty. This kind of effect has been studied extensively in the economic literature. For instance, suppose that agents have a high ability to predict a policy-maker's actions, that at least one endogenous variable is completely controlled by the policy-maker, and that the reward to the policy-maker of their actions is dependent on whether the policy is anticipated. Under these simple conditions [16], there is no unique rational course of action for the policy-maker, in general. Rather, different incommensurate yet equally viable theories may indicate distinct optimal policies, and consequently the agents cannot form rational expectations because any rational expectation must contain a theory of the policy-maker's behaviour. In a setting such as this, economically rational behaviour is basically incompatible with the formation of economically rational expectations. This is an instance of an incompleteness phenomenon occurring in a purely economic setting, and one that bears directly on autonomous machines.

21.2 Fundamental Uncertainty in Economics

21.2.1 Economic Agency and Autonomy

The tight connection between notions of economic agency and autonomy is well established in the economic literature: indeed, the history of artificial intelligence research and economics are intertwined because economics is, in essence, about human decision-making under different resource allocation mechanisms and conditions [36]. Economics has constantly looked to artificial intelligence for models, especially in response to the realisation in the 1970s [43] that economics had been weak on both the nature of human information processing capabilities and on how we achieve these capabilities, despite the central place that these problems must have in economic theory. The boundaries have become increasingly blurred, on the economics side at least, because while economics requires descriptive models of human decision-making, the available tools are primarily the prescriptive models offered by traditional artificial intelligence and applied mathematics. Furthermore, economics has itself developed prescriptive models in the course of trying to describe and predict economic system behaviours (see for instance, [15]).

While autonomous systems research has yielded many prescriptive agents that work well enough in tightly constrained environments, the economic requirement for descriptive models arguably first highlighted the yawning gap between machine reasoning algorithms and the robustness and flexibility of human decision-making. Smith recently argued [44] that in both economics and in artificial intelligence, the underlying assumptions driving research about agent learning and decision-making have typically neither sufficiently nor even explicitly emphasised the significance of fundamental uncertainty; current models remain structurally very similar to those of the past. We remain largely tied to probabilistic and statistical methods of learning and reasoning and thus to their inherent limits.

He argues that the advances in machine intelligence of recent years overwhelmingly do not even attempt to address this failing and consequently do not represent a base advance in overcoming the difficulties of modelling and reproducing human decision-making, increasing success in tightly defined domains notwithstanding. Implicated in maintaining these base assumptions unchallenged is the documented anthropomorphic tendency to label constructions in artificial intelligence with human decision-making features or human functions, in the absence of any convincing argument establishing a similarity [34]. These wishful mnemonics have arguably helped to foster the hype cycles experienced repeatedly during the course of the history of artificial intelligence research and application, while masking the limitations of the available methods to cope with non-stochastic uncertainty.

The implications of this for the development of autonomous systems capable of operating in high uncertainty environments such as those of military operations are straightforward. The failure of machine learning in comparison with human performance is poor generalisation ability; the failure of machine reasoning is brittleness, meaning a poor ability to handle environments in which future states are the

manifestations of interconnected events, including the actions of the agent itself. In both cases, the failure manifests as an inability to operate successfully—where success here means extended survival more than reward maximisation—in environments where future events are not foreseen and may not be foreseeable at all.

21.3 The Inadequacy of Bayesianism

The connection between autonomy and economics runs very deep: questions about the formation and application of beliefs in decision-making, and hence criteria delineating rational from irrational beliefs, are as prominent in economic theory as they are in artificial intelligence. For instance, the central place of equibria in economics and game theory is really a statement of a convention about the rationality of beliefs, wherein rational beliefs are held to be exactly those that coincide with equibrium solutions [18]. The Bayesian interpretation of belief is predominant in the economic literature, as it also is in artificial intelligence; the criticisms of Bayesianism in economic theory [18] apply equally in autonomous systems research and may be seen, in essence, as another recognition of the inability for a currently dominant paradigm to adequately handle fundamental uncertainty.

Bayesianism is not really a single position, but any mix of three main postulates. Firstly, an agent should have probabilistic beliefs about unknown facts, typically given as a probability measure over all possible relevant states. Note that economics often uses a stronger version than computer science and artificial intelligence by dropping the relevance qualification and thus presuming that agents must have beliefs as probability measures about absolutely everything. Secondly, Bayesian priors should be updated to yield a posterior measure in accordance with Bayes' Law. Lastly, each agent should choose the decisions that maximise expected utility (or sometimes equivalently minimise expected cost) with respect to that agent's Bayesian beliefs. While the third postulate, in conjunction with the first and second, is hardly unknown in computer science, it is particular common in economic theory.

No combination of these postulates has been immune from serious criticism in the economic literature. The Bayesian approach presumes a prior, and thereby does not deal with the manner in which the prior is obtained. This situation is sometimes known as state space ignorance or sample space ignorance. Though the second postulate seems technically safe, it has been shown to be descriptively inadequate [51]. Moreover, the economic literature arguably under-estimates the importance of complexity limitations on Bayesian updating; indeed, the computer science literature has been long focussed on this as the main difficulty. The third postulate has been attacked repeatedly in the economics literature since the Allais Paradox [2], which revealed strong inconsistencies between observed human choices and what maximising util-

ity would predict (see for instance [23] and [33]). The entire field of Behavioural Economics[1] is substantially based on the rejection of this third postulate.

For present purposes, it is the criticism of the first postulate that warrants particular notice: the lack of convincing general descriptive and prescriptive explanations regarding the origin of the prior translates into an important epistemological question. That is, this criticism amounts to a different way of stating that, in general, the agent cannot possibly have at its disposal sufficient time and resources needed to obtain the complete set of possible relevant states over which to draw belief measures. The economic literature goes even further, however, by also proposing an ontological notion of uncertainty that asserts that, beyond limits to knowing, questions about outcomes pertaining to agent decision-making are simple unanswerable at all.

21.4 Epistemic and Ontological Uncertainty

To see how uncertainty plays out in economic settings, consider the essential problem of human agents operating in an economy: they must make investment choices that will play out a future they can neither predict nor really control. Knight [27] famously formalised a distinction between risk and uncertainty on the basis that economic agents operating in a dynamic environment must do so with imperfect knowledge about the future. Knight distinguished risk, as applying to the situation when the outcome is unknown yet the chances are measurable, from uncertainty, when the information needed to measure the chances cannot be known in advance of the outcome; we do not have a sufficiently long history with the system as it currently stands to be able to establish a measure. Knight maintained that risk can be converted into an effective certainty[2]; the practice of setting hurdle rates as the rate of return on the next best option having a comparable risk profile as a mechanism for soft capital rationing is one example of this kind of conversion of a risk into a cost.

In Knight's conception, uncertainty is distinct from stochastic risk in that it is not amenable to measurement and consequently cannot be meaningfully converted to a cost in this manner. Note that this conception of uncertainty effectively raises to a more general setting the first objection that was discussed earlier in relation to Bayesian assumption of a prior, which insists that a measure can be defined over the space of possible relevant states.[3] Knight's uncertainty is epistemological: we lack knowledge of what future outcomes might occur, but at least we can be aware that we lack it.

---

[1] Behavioural Economics is basically about the study of observed behaviour of real economic agents, in contrast to the normative approach that neo-classical economics takes to behaviour.

[2] This kind of conversion is a ubiquitous standard practice. Knight made the point that this practice is indeed justified, so long as we are dealing only with measurable stochastic risks.

[3] Hence the field of Behavioural Economics, based substantially as it is on the rejection of utility maximising postulate of Bayesianism in economics, assumes as its basic position a Knightian epistemological view of uncertainty [13, 50].

The importance of the distinction has not been lost in analyses of recent behaviour of finance firms, which, in the years leading to the global financial meltdown of 2008, operated with highly regarded, highly precise risk assessments that were all based on the basic premise that the relevant conditions and outcomes are all measurable [3, 13, 26, 32, 40, 41]. When, in a mass financial panic, institutional investors suddenly realised that the assumptions about their risks being measurable were deeply flawed, financial markets collapsed in the kind of event that has been described as a 'destructive flight to quality'. Investors clambered over the top of each other to dump everything but the safest and hence least profitable assets, in a sudden cascading systemic failure having global ramifications that continue to reverberate nearly a decade later. The distinction between risk and uncertainty in economics is not an esoteric matter, but a very practical pervasive kind of problem with potentially huge ramifications. Indeed, the aftermath of the economic crisis has seen a considerable resurgence of interest among economists in fundamental uncertainty, after a long period where the primary interest was in formalising economics - and particularly finance - using precision stochastic risk models (see for instance [26]).

In contrast to Knight, Keynes [25] argued for a deeper ontological notion of fundamental uncertainty: we do not even know what we do not know. In this view, the future is fundamentally uncertain because it is simply not possible even in principle to conceive of all relevant possible future outcomes in advance.[4] Investment, Keynes maintained, is then the allocation of resources on the basis of expectations formed under conditions of this kind of uncertainty. Keynes formulated a set of conventions to describe how agents try to cope under such uncertainty, by resorting to what amount to superstitious rituals from the point of view of utility-maximising economic rationality. They tend to presume that their existing opinions are a valid guide. They tend to conform to majority views. They just ignore what is unknown: widespread reliance on risk models under the assumption that all the uncertainty is measurable as briefly described above is a prevalent contemporary example.[5] They presume that present circumstances will be stable. They rely on the opinion of experts who concoct grand predictions from the economic tea leaves. Perhaps most importantly, in a mammoth and sometimes even wilful act of self-deception, they assume far greater veracity for these measures than what any frank examination of the past would ever support.

In short, agents invest on the basis of strongly predictive models, and the consequences are far-reaching. It means that markets cannot be in stable efficiency equilibria, that investment is highly volatile, and that expectations are extremely fragile. It also means that complex behaviour such as economic bubbles and bursts and overall unpredictability of economic systems can be wholly generated endogenously [1, 11, 21, 26, 29, 31, 32, 41, 46, 52]. This turbulent picture stands in contrast to the text-

[4] I will come back to the reason for this in detail in a subsequent section. Briefly: the questions we might need to answer may contain hidden self-reference, which opens the possibility that they might be logically paradoxical and hence unanswerable from inside the system within which we have to operate.

[5] I will further discuss these instances in subsequent sections, particular in relation to the GFC, trading strategies and market bubbles and crashes generally.

book neoclassical account, which maintains that the expectations of economic agents about future returns are correct on average over time, under the so-called "rational expectations" doctrine.

Keynes further held that economic agents making decisions under uncertainty hold more liquid assets - especially money - as an asset in response to doubt about future returns; this store of wealth in liquid form is essentially a concrete and measurable manifestation of the agents' confidence regarding future outcomes (but not a measure of the uncertainty of future outcomes). Lower confidence requires higher interest rates to inveigle our agents to draw their capital from safe but unprofitable liquid deposits and invest it in volatile but potentially profitable illiquid assets. So although the agents themselves can behave pretty miserably, Keynes concludes that wise governmental moderation can, in principle at least, considerably stabilise an economic system by setting monetary policy to moderate the billow and bounce of otherwise outrageous market circumstance.[6] This proposition has been repeatedly echoed in subsequent experimental and theoretical studies concluding that well designed control policies can be very effective in moderating or eliminating asset bubble formation; some investigators report that dynamic policy control is superior in this regard to static controls [32].

These economic agents face a bimodal impetus, compelled to avoid loss on the one hand and to seek profit on the other, but they must undertake this activity under conditions of irreducible uncertainty. Interestingly, the source of macroeconomic volatility appears to lie less in the presence of fundamental uncertainty, which is unavoidable anyway, and more in the manner by which our agents attempt to avoid dealing with it. They retreat from it through the fallacious appeals that Keynes describes. Left unmanaged, the consequences of this behaviour are that numerous small failures accumulate unrecognised and unreconciled, eventually erupting in the sudden system-wide failure broadly known as a 'crash' when the edifice of false confidences in these measures can no longer be maintained under the accumulated burden of hidden errors.

Keynes' conventions seems to provide a concise summary of managerial methods, which makes sense from the point of view that management practices and investment behaviour are inextricably intertwined. A deeper connection is that one of the pillars of management theory is the inversion of the economics of externalities[7] [9]. The

[6] This does not intend to imply that Keynesian economics is without assumptions nor subject to limitations, but merely to convey the component of the theory that pertains to uncertainty and its effects in relation to autonomy as an economic question. For instance, the details of how governments intervene really matter: the Keynesian interventions in the 1930 s that directly supported broader populations deeply impacted by the Great Depression were clearly more successful than the bailouts after the GFC in which trillions were poured into the large failing financial enterprises whose activities had caused the bubble and ensuing crash, and which produced exploding deficits. The outcomes of the current downturn is leading many to question capitalist economic systems themselves, amidst a growing view that capitalism is inherently unstable, inefficient, antidemocratic, and not - as often assumed - synonymous with the presence of markets; such debates substantially question the premise as well as the limits of Keynesian intervention to stabilise capitalist systems.

[7] An externality, or transaction spill-over, is a cost or benefit that is not transmitted through the resource allocation mechanism and is instead incurred by a third party not involved in the transaction.

actions of the agents are reliant on strong assumptions about the measurable nature of future outcomes as a basis for supposedly justified action; the well documented over-reliance on risk models in finance is about precisely this kind of justification [26]. The conventions also capture the essence of the behavioural assumptions often built into automated systems, which brings us to what I regard as the fundamental question of autonomous systems development: if the uncertainty is not measurable, then how can we build systems that can effectively deal with it? The failures of machines and organisations that similarly incorrectly assume that the uncertainties of their operational environments are measurable will similarly see failures tend to accumulate into sporadic cascading systemic distress, and I warrant that it is largely in recognition of this unacceptable potential under extreme forms of uncertainty that military operational settings have proven largely unyielding to the best current technology has to offer.

How, then, should we conceive of autonomy? If we consider an autonomous agent as an economic agent self-allocating under fundamental uncertainty, then such agents - be they humans, organisations or machines - display a property I term plasticity: the ability to countenance unpredicted, and unpredictable, future states of the operating environment, in a social setting, within acceptable limits. The name is in reference to the implied need for the thing exhibiting the property to change itself in response to changing environmental conditions. Autonomy will then apply specifically to machines that satisfy this condition. This conception of autonomy has nothing to do with the inherent sophistication of algorithms, power supplies, sensors, actuators and circuits, but instead motivates technical finesse specifically to the extent and in the direction needed to fulfil the plasticity imperative adumbrated by the intended operational setting; this position reflects the primarily economic viewpoint of this chapter, in distinction to the algorithmic or robotics emphasis widely seen in the literature.

21.5 Black Swans and Universal Causality

Taleb famously coined the parable of black swans [48] to describe the occurrence of unforeseeable rare events having high consequences, and previously described the strong tendency of humans to find simple, though erroneous, explanations for their occurrence, after the fact [47]. It has since been established that this description of uncertainty draws the same basic distinction between stochastic risk and Knight's

Externalities may lead to inimical outcomes by upsetting the resource allocation mechanism, and there is typically a large magnification effect whereby a small benefit to one or both parties in the transaction generates a disproportionately large cost to the third party. Economics attempts to limit such effects, as represented most famously by Coase's Nobel-prize winning work on externality elimination cited in the text. In contrast, management theory contains a branch that aims specifically at generating externalities for the benefit of a specific party in the transaction (privatisation of profits) and the cost of other parties, which might include the second party in the transaction (socialisation of costs).

intractable epistemological uncertainty [13, 14, 50]. The black swan anecdote serves to illustrate Knight's epistemic uncertainty outside an obviously economic or financial setting. Taleb's description of the behaviour of humans in concocting reasons for the event after the fact mirrors Keynes' conventions describing economic agent behaviour, but does so in a manner that highlights the role in shaping expectations - whether economic or otherwise - of a widely discredited position usually known as universal causality in philosophy.

Taleb's observation that humans regard events as being much more attributable to determinable causes than they really are has a long history in philosophy as the notion of universal causation. Universal causation maintains that all events are the result of prior events, and this belief connects the construction of Taleb's post hoc explanations for rare events with Keynes' view that economic agents maintain undue reliance on the veracity of prediction. The post- economic meltdown criticisms of financial firms assuming that the relevant risks are all measurable is a modern manifestation of the same effect.

The intuitive appeal of this view lies in that whenever we ask simple questions after the fact about why a particular event occurred, we can obtain a plausible explanation for its occurrence in terms of some causal chain of earlier events. So it would seem on the face of it that every event is caused by something, albeit probabilistically, and hence that every event follows from prior events according to some governing logic. Yet the more deeply we dig, the more ostensibly antecedent sets of conditions look like reasons for deciding to act in a certain way, or, more pertinently, for not acting a certain way, and the less they look like the inevitable causes of an event that we first supposed. In other words, alleged causes are really only epistemic factors that influence the decisions we make from within a problem context, rather than immutable ontological features of an environment into which the agent peers from the outside.

Universal causality connects to predictability through causal determinism, which holds that every event is necessitated by some set of prior events. This claim is then the antecedent to so-called 'scientific' determinism, which concludes therefore that the world is basically predictable. To elaborate: 'scientific' determinism alleges that the structure of the world is such that future events can be predicted with precision depending on that of knowledge of the governing laws of the phenomena of interest and the accuracy of the account of past events. It is worth noting that 'scientific' determinism is poorly named, for it is not actually about determinism at all: determinism refers to the absence of arbitrary choices in the application of transition operators, with non-determinism then being the admission of arbitrary choices. Rather, 'scientific' determinism is an assertion about predictability, equivalent to assuming stability and logical completeness. Though perhaps intuitively appealing, the inference from all events having necessary causes to predictability is flatly wrong: it is well established that even fully deterministic systems in which the current state completely determines the transition to a unique subsequent state can be nonetheless savagely unpredictable [20, 22]. Conversely, non-deterministic systems can also be completely predictable. Even if we limit ourselves to completely deterministic sys-

tems and hold that this is a correct characterisation of the world, strong predictability remains the exception, not the rule.[8]

So the premise to the conclusion of predictability that we can rely on for, say, making investment decisions or for autonomously operating in a complex operational environment, is untenable in general, and we draw this conclusion without needing to reject the notion that some events may have causes, or even the stronger assertion that every possible event has causes. It collapses merely when we admit that causes may be only necessary but not sufficient for at least some events that matter to us in terms of our decision-making some of the time. Moreover, across a wide range of contemporary fields, including mathematics, computer science and physics, as well as economics, it appears increasingly clear that there are also events that just do not really have any cause at all [7, 8]. This ties in very closely with Keynes' ontological notion of uncertainty.

21.6 Ontological Uncertainty and Incompleteness

21.6.1 Uncertainty as Non-ergodicity

An ergodic system [20, 22] is one that tends towards a limiting form that is independent of its initial conditions; in a dynamical system sense, this means the phase-space average is the same as the infinite time average, for all Lebesgue-integrable functions almost everywhere (meaning except possibly in sets of measure zero). In other words, an ergodic system is one for which sampling - collecting more data - actually gives more information about the underlying system, so ergodicity characterises the precise conditions under which obtaining additional data provides additional information. The mechanisms governing the system are stationary, so they remain constant over time, and they satisfy some regularity conditions, essentially meaning that they are well behaved. Ergodicity effectively means that we can float detached from the world about which we make observations and predictions, thus avoiding the observation and prediction paradoxes that produce fundamental uncertainty.

Paul Davidson argues that the rational expectations hypothesis and efficient market hypotheses of textbook economics are worthless, and indeed positively dangerous, on the grounds that real economic systems are inherently non-ergodic: such systems are not regular or not stationary or both and consequently it is unreasonable to expect that they will converge to any equilibrium distribution, and they cannot be amenable to reliable forecast as a result [14]. Davidson holds that Keynes' ontological uncertainty pertains to the behaviour of such non-ergodic processes, in contrast to

[8] I will pick up on the question of exactly what are the conditions under which strong predictability in principle holds in the next section. In short, the conditions amount to the delineation of Keynes' ontological uncertainty, which is about absolute limits on what is knowable in principle. Knight's epistemic uncertainty and Taleb's black swans amount to further practical limits on the tractability of knowing.

Knight's epistemic uncertainty and Taleb's Black Swans, which still presume the presence of ergodic processes. In the latter case, the surprising outcome simply lies far out on the tail of a nonetheless fixed and well-behaved distribution, with apparent uniqueness given by the inordinate time between re-occurrences. In the new edition of Taleb's book, he concedes that there is a difference between non-ergodic processes and black swan events but dismisses the difference between the two as irrelevant.[9]

But there is a world of difference. The various species in the zoo comprising the ergodic hierarchy have very different properties, particularly in terms of the kind and degree of uncertainty they manifest, essentially in terms of the kinds of questions we might want to ask about the future behaviour of such systems and which of those questions can be answered in advance of simply waiting to see what happens. The most important distinction is between the class of ergodic processes and all the other classes of systems that fall somewhere in the much larger world of processes that are non-ergodic [20, 22].

To be slightly more formal: the ergodic hierarchy is a classification scheme for deterministic dynamical Hamiltonian systems, in terms of their relative unpredictability. Ergodic systems characterising certainty, stochastic risk and epistemic uncertainty are at the bottom of this hierarchy, being the most highly restrictive and correspondingly the lowest in terms of the level of uncertainty they can manifest. Weak mixings are next, then strong mixings, above which are K-systems, whose behaviour is already very strongly unpredictable, and the topmost currently recognised classification are Bernoulli systems, whose behaviour is the most deeply unpredictable in the hierarchy.[10] The criteria for strong mixings have been convincingly proposed as the demarcation of what is commonly regarded as deterministic chaos [55]. There are also interesting systems that straddle between K-systems and Bernoulli systems, known variously as C-systems or Anosov systems, but their relation to the other levels mentioned here in terms of unpredictability is more complicated and beyond the scope of the present discussion.

Note that non-ergodic systems do not undergo arbitrary change at any moment; the presence of fundamental uncertainty does not require or entail total disorder. To the contrary, systems that are non-ergodic will typically fall into transient states of apparent stability that dissipate as suddenly and unexpectedly as they start, never to repeat themselves. Strong prediction about future system behaviour is impossible, in the sense that the kinds of questions we might want to ask about the future behaviour of the system are not solvable, with higher classes in the hierarchy representing a situation of having fewer such problems for which there is a solution. Yet this does not mean that we cannot cope at with life in such a system - as individual economic agents, people and organisations certainly manage to do so - but rather that we cannot

[9]This dismissal of the significance of the difference is understandably not well received amongst researchers in ergodic theory and nonlinear dynamics, for reasons that will become clear.

[10]Interestingly, entropy is not sufficient to classify K-systems, meaning that there are uncountably many K-systems with the same entropy but that are not isomorphic; thus Ornstein's isomorphism theorem does not work for K-systems. All K-systems are also Bernoulli systems, but not vice versa; Bernoulli systems potentially manifest greater unpredictability. Yet Ornstein Theory is sufficient to classify Bernoulli systems.

expect to do so very successfully by using by relying on methods that presume that uncertainty is measurable, or that presume ergodicity, typically by using relatively strong predictions about what will happen in futures delineated by the time periods over which decisions will play out.[11]

The power of complex systems to produce long sequences of apparent predictability[12] is highly seductive. Self-reinforcing beliefs about predictability of future returns and thus future market behaviour that drive the formation of market bubbles is a highly visible example of this [1, 31, 32, 41, 46, 52]. The precision risk models heavily implicated in the mortgage-backed securities bubble preceding the Global Financial Crisis of 2007–2008 provides almost innumerable practical examples of the catastrophic failure of ergodic models in what are actually non-ergodic environments.[13] They will tend to fail suddenly and disastrously rather than smoothly, but often will do so only after mendaciously long periods of apparent positive success. Agents are more easily deceived because holding ergodic expectations about the world will naturally mean that they will also expect failures to be similarly be ergodic, and thus relatively benignly behaved and predictable. Yet there is simply no basis for this expectation. The distinctly irregular and non-stationary quality of the collapses of ergodic models in what turned out to be highly non-ergodic systems, came as something of a surprise to those invested in them, to say the least.[14]

The difference between Keynesian ontological uncertainty and Knightian epistemological uncertainty is that the former position holds that some things are simply not knowable, while the latter entails that with better information and greater ability to process it we could, in principle, calculate the probabilities for more kinds of events. The epistemic uncertainty notion ultimately sees the universe still as a collection of ergodic processes, and uncertainty as essentially a consequence of limitations that computer science calls tractability. Roughly speaking, tractability limits occur

[11] There is a growing general awareness in economics, as exemplified here by Davidson's work [13], that economic systems worth worrying about are inherently non-ergodic. The inescapable fact that real economic agents can, do and always have successfully operated under these conditions should suffice to refute the proposition that it is not possible to operate adequately in such an environment so we should not bother to do so in artificial intelligence research.

[12] Even a completely random sequence produces such sequences of lengths that are logarithmic with respect to the overall observed sequence length, which is deceptively long [35].

[13] The economic analyses typically describe this in terms of the catastrophic cascading failure of the application of precision risk models under conditions where the falsity of the strong underpinning assumptions to the effect that all relevant risks are measurable was never examined (nor were these assumptions even stated, so much were they taken for granted). Under Davidson's direct mapping from ontological uncertainty in economics to non-ergodicity, we have the stated interpretation.

[14] John Meriweather famously described the financial collapse of 2007 as a ten-sigma event, which means, according to the predominant economic models, that it should occur no more than about three times in the entire history of the universe. The models were designed from the outset to eschew the very possibility of catastrophic failure. Apparently it did not occur to the adherents of the orthodoxy even after the fact that their models might be flawed, despite the manifest empirical failure and the clear absurdity of many of their base assumptions. Even in 2010, Bernanke argued that the problem was not that the economic models failed to see the economic crash coming, but rather it was that the economic crisis was an event that was just not supposed to happen. Apparently, reality should consider itself refuted.

because although a problem may be formally solvable, the time and space requirements to solve it rapidly explode beyond the ability to meet them as the problem size increases. So the inability to sample a system for long enough in order to observe occurrences of ultra-rare events -black swan evens - are a tractability constraint of the type that characterises epistemic uncertainty.

21.7 Uncertainty and Incompleteness

The Keynesian position maintains that obtaining answers to some questions is just not ontologically possible, in exactly the same grain as the deep mathematical uncertainty expressed in results such as the incompleteness of every formal axiomatic system that contains arithmetic, the existence of recursively inseparable sets, and the unsolvable nature of most computing problems, most famously the Halting Problem. In other words, ontological uncertainty amounts to the fact that problems of determining future outcomes in non-ergodic economic systems are generally paradoxical, because such systems allow the possibility of self-reference. So the Keynesian uncertainty concept amounts to an economic manifestation of algorithmic randomness, which rests on computability theory and amounts to the modern study of incompleteness phenomena, by regarding effective procedures as compressions of potentially unbounded sequences of data generated by the system of interest and asking about what sequences have finite compressions.

Formal axiomatic systems are mathematical languages allowing us to talk formally about phenomena in which we are interested, including other axiomatic systems, so it's difficult to over-state their significance to autonomy. They each consist of a set of basic terms and a set of reasoning rules defined by axioms that describe, essentially, what we might conclude from what given conditions. Such a language allows us to formally state propositions, some of which might be provably true in the sense of being logically entailed by the axioms, such as "there is no largest prime number" in Peano Arithmetic (the basic theory of numbers, see for instance [24]). Other expressible propositions, such as "adding two positive numbers together yields a number smaller than either" in Peano Arithmetic, reduce to contradiction, which is a primitive term of a system that is false in all interpretations. The most basic question about whether we have a viable axiomatic system to use is whether or not the axioms themselves entail contradiction; if so then the system is said to be inconsistent, and it does not represent a viable basis for reasoning because in such a system it is possible to conclude absolutely anything. As famously shown by Kurt Gödel [19], this fundamental question of the consistency of formal reasoning systems turns out to be anything but trivial.

Some formal axiomatic systems, such as Turing machines [39, 45], describe computation and thus set the ultimate basis for machine intelligence. In this setting, the complexity of any other system is defined as the size of the smallest procedure, with respect to some reference machine model, that reproduces the data observed from that system [35]. The remarkable fact is that this complexity is asymptotically

independent of the particular machine model, up to additive constants.[15] The conditional complexity of a sequence with respect to some information is the smallest effective procedure that takes the information as an input and produces the sequence as an output. A sequence is then said to be incompressible when the smallest size effective procedure is asymptotically comparable to the size of the sequence. For infinite sequences, the complexity is the limit as the length of an initial segment of the sequence approaches infinity of the size of the smallest effective procedure producing the segment, divided by the length of the initial segment of the sequence. If the complexity in the limit is non-zero then the sequence is incompressible, meaning that it represents a fixed individually random mathematical object, indistinguishable from the flips of a coin by any possible statistical test.

As indicated earlier, an irreducibly random sequence of this type provably contains surprisingly long sequences, of about a logarithmic function of the observed sequence length, that appear to be regular and stationary [35]. There is a deep connection between the non-linear dynamics view discussed earlier and the algorithmic information view of unpredictability: the trajectories of a non-linear dynamic system can be encoded as infinite sequences by dividing the state space of the system up into numbered cells and tracing the trajectories through these cells. A trajectory is random when there is a partitioning of the state space into cells such that the encoding of the trajectory is algorithmically random. Predictability means having a compression - in terms of some effective procedure - for anticipating the outcome in advance. Yet there are simply not enough compressions to go around. It is not even close: the shortfall is exponential, meaning vast majority of possible systems are left with no compression by which their trajectories can be predicted that is shorter than simply waiting around to see what eventually happens.

The underlying reason is that prediction in general is paradoxical: while a contradiction is both true and false, a paradox is neither true not false within the logical frame of reference in which it is stated. Though we usually think of paradoxes as obviously self-referential statements, they are usually not so obviously discernible because paradoxes actually do not need not be visibly self-referential. The famous Halting Problem for Turing Machines, the Busy Beaver Problem,[16] and their equivalents in other computational models, as well as the compression problem are all actually paradoxes that do not look like it on the face of it because their self-referential nature is hidden from immediate view.

The reason for this is that the self-referential expression is not as primitive a notion as it might at first seem: numerous systems of logic come with various kinds of implicit function theorem by which self-referential statements can be turned into equivalent statements, called "fixed-points", that lack obvious self-referentiality [5]. Paradoxes are normal, natural, and extremely common, and formal mathematical sys-

[15] This is one of those mathematical facts that seems remarkable upon first discovery, and entirely natural to the point of obviousness afterwards.

[16] Give me the largest natural number that can be generated by an effective procedure with respect to some model of computation - a program in your favourite programming language, if you prefer - limited to at most a given size.

tems are full of them[17] [7, 8]. Turing's original proof of the unsolvability of the Halting Problem for Turing Machines, by a Cantor Diagonalisation Argument,[18] further reveals the detailed nature of mathematical paradoxes, and hence of the irreducible nature of fundamental uncertainty: they are kind of folded-up infinite regresses. The proof in question is essentially just an infinite successive unfolding of the Halting Problem paradox to yield what amounts to an impenetrably unknowable number [8].

The basic lesson of Gödel's Incompleteness Theorems [19] is that any system that allows for the possibility of self-reference - any system containing basic arithmetic will do - will give rise to paradoxes, and this will manifest as uncertainty in the form of the presence of problems we might like to solve but to which there can be no solution from within the system. A bigger system might be able to provide a solution, but we don't in general have the luxury of stepping out to it and peering into the phenomena with which we are concerned from the outside. The basic example of this is that we are bound to compute things from within the limitations incumbent in the models of computation, which are all known to be equivalent and absolute, and under the Church-Turing thesis are not surmountable by any other realisable system either [39, 45].

Can the ergodic hierarchy provide a formal mathematical basis for characterising plasticity - the property of being able to survive in an unpredictable environment? I suggest so: if the observations that an agent makes of its environment meets the criteria of, say, a K-system, and yet that agent is able to survive in that system for better than a logarithmic function of time, where logically time would be taken in terms of successive observations the agent makes of its environment, then we know that the agent must be doing better than merely taking advantage of an appar-

[17] As an example of this, I recently played around with paradoxical statements about Peano Arithmetic - axioms about the behaviour of the natural numbers under the usual operators - using Provability Logic. Provability Logic system consists of familiar propositional logic with a modal operator $\Box$ meaning "it is provable that", its dual $\Diamond$ meaning "it is not disprovable that", and Löb's Theorem, which states that in any system containing Peano Arithmetic, any time we can prove that something implies its truth we may conclude that it is provable. We can use Provability Logic to explore and even to write computer programs to generate arbitrarily many generalisations of Gödel's Second Incompleteness Theorem [19] for us, by feeding it with paradoxical statements. Provability Logic's implicit function theorem guarantees that we have unique solutions to a large class of self-referential expressions. For instance, the solution $\neg(\ulcorner\Box\bot) \rightarrow (\ulcorner\Box\bot)$ to a paradoxical statement $p \leftrightarrow \ulcorner\Box p$ happens to be direct restatement in Provability Logic of The Second Incompleteness Theorem, asserting that the system cannot prove that it is consistent, or equivalently, that if it can prove that it is consistent then it must be inconsistent. Here, the symbol $\leftrightarrow$ stands for if and only if, $\rightarrow$ is implication, and $\bot$ stands for contradiction. The statement $\ulcorner\Box\bot$ says that the system is consistent.

[18] Cantor's Diagonalisation Argument was first used to prove that there are infinite sets that cannot be put into one-one correspondence with the natural numbers, and later to prove that the real numbers are uncountable, Russell's Paradox whereby attempted formulations of set theory prior to Zermelo set theory are inconsistent, and Gödel's First Incompleteness Theorem, as well as the unsolvability of Turing's Halting Problem [39, 45].

ently regular transient state, and thus must be dealing with some effectiveness with K-system uncertainty.[19]

21.8 Decision-Making Under Uncertainty

Keynes' ontological uncertainty corresponds to incompleteness phenomena in the same manner that Knight's epistemic uncertainty mirrors intractability, meaning that the root cause of ontological uncertainty is the possibility of self-reference and thus of logical paradox. Agents self-allocating in an environment where their actions affect the future states of that environment and that expectations about the future states of the environment impact on the agents' decision is logically self-referential. This is why I consider such self-allocation to be a feature delineating autonomy. The self-referential nature of self-allocation means that ultimately problems of maximising utility or efficiency in the customary sense must be formally unsolvable, so the question remains what we can do in terms of developing autonomous systems. In economics and finance, an examination of trading strategies is a good place to start for a solution, in light of the huge literature on both these kinds of strategies and their outcomes for both the agent and for the overall systems of which they are a part. To illustrate this: speculation trading specifically relies on the fact that asset prices are non-stationary, for there simply is no profit to be had for an asset speculator in a stationary price environment.

Finance economics identifies two basic types of trading behaviour: those who attempt to predict future price movements by looking for patterns in historical data are termed chartists, or sometimes technicians; those who trade on the basis of trying to determine the financial fundamentals of assets - their 'real' value - are termed fundamentalists (this terminology is common in the empirical studies of market dynamics; see for example [31, 32, 37, 46, 52]). Of course, real traders may represent a mix of trading strategies, and may alter this mix over time, so these types should be read as pertaining more to agent behaviours rather than to the agents themselves. The chartists are the speculators, and are willing to purchase assets at prices above their fundamental value, in the intention of making gains by selling those assets at still higher prices. Chartists thus operate essentially on the basis of scepticism about the rationality of other traders. They are traditionally held to be the bad guys insofar as asset bubble formation is concerned, because this speculative demand is well known to tend to build on itself in a self-reinforcing manner: speculative trading means higher demand, which pushes asset prices higher and higher above their fundamental values, resulting in bubble formation.

The fundamentalists no longer get off the hook so easily with respect to asset bubble either: the problem lies in the difficulty in assessing assets to determine their

[19] I have started to try to formalise this concept. The difficulty seems to lie in defining a general notion of what 'surviving' should mean in formulating a deterministic dynamical Hamiltonian model of agents interacting in an environment.

fundamental value and thus rational price. The crudeness of commonly used asset valuations is obvious from even a cursory examination of the kinds of models typically in use, such as the Gordon Growth Model, 2-stage model, 3-stage model and H-model [37], yet the broader issue facing the fundamentalist is that uncertainty must be an inevitable limitation with any asset value model, irrespective of its mathematically sophistication.[20] Asset valuation turns out to be effectively just a different mode of speculation (see for instance [31, 41]). The resulting systematic valuation errors may drive asset bubble formation, even in the absence of the speculation-driven demand of the chartists.

A market containing such traders allows the possibility of self-reference, because the behaviour of the market is a consequence of the behaviours of all the traders who operate within it, and whose behaviours are, in turn, deeply effected by the current state and history of the market. Consequently we should expect fundamental uncertainty here: the problems the agents are trying to solve are without complete solution, and indeed this manifests in real markets in the form of unpredictability, or what economists know as "volatility". In the final analysis, economically rational beliefs are harder to come by than it may appear at first blush. Maximising reward might make perfect sense when we consider an individual agent in isolation. Yet the collective consequences of agents effectively relying on strong assumptions about the independence of the future states with respect to their individual actions in the present means that reward maximisation becomes self-defeating, when, during the subsequent crash, almost everybody following such strategies loses.

21.9 Barbell Strategies

A barbell strategy[21] [37] is formed when a trader invests seeks to increase risk-adjusted returns by investing in a combination of safe long duration investments, and the small remaining portion in short duration securities, with nothing in intermediate duration investments. Closely related is laddering,[22] which avoids reinvesting assets in unfavourable economic environments, by investing in multiple instruments having

[20] Indeed, this is an example where increasing sophistication is actually dangerous, by creating false impressions about the reliability of the model. As explained elsewhere in the text, this factor of over-confidence in precision financial risk models is heavily implicated in the GFC of 2007–2008, and thus gives us a highly topical real-world example of the phenomenon.

[21] The term is very common in financial economics. See for instance http://www.forbes.com/forbes/2005/0509/144.html and http://www.investopedia.com/articles/investing/013114/barbell-investment-strategy.asp for brief overviews.

[22] Not to be confused with a type of insider trading known by the same name. Laddering is also used as a term to denote a process whereby insiders purchase stock at lower prices while artificially inflating the price to permit them to then sell at a higher price, by agreeing upon purchase at the lower price to also purchase additional shares at some higher price. This practice was a target of SEC investigations in the wake of the Global Financial Crisis of 2007–2008 [Fjesme, S.L, Initial Public Offering Allocations, Price Support and Secondary Investors, Journal of Financial and Quantitative Analysis (JFQA), Forthcoming].

different maturity dates; the difference is that laddering spreads investment across short, intermediate and long maturity instruments. Laddering can be seen as a kind of nesting of barbells; while real autonomous systems may in general have to ladder in this manner, I will focus here on barbells as the primitive basis of such a plan of attack.

The opposite of a barbell is a bullet strategy [37], where a trader invests in intermediate duration securities, to build a portfolio that has securities that mature consistently over time. Note that both fundamentalists and chartists typically employ a bullet strategy, just with respect to different choices of problem: the fundamentalist works on the basis of discounted average expected future dividend returns, while the chartist operates using expectations about patterns in asset price movements. The barbell strategy rests on a different base choice of problem: surviving the unexpected, rather than maximising expected returns. A barbell on only the shortest and longest bonds in a bond market is, under a simplistic forward rates assumption, known to be a maximiser of the modified excess return [10].

Taleb [49] invokes a version of the barbell strategy emphasising a large majority in extremely safe instruments that pay poor returns and the remaining in highly risky but potentially highly profitable instruments as a strategy to insulate against black swan events. Taleb presents barbells as applicable outside trading markets; this leap is not a large one having undertaken to regard decision-making in general as resource allocation. The strategy works best in periods of high inflation: put options are cheaper under high interest rates in accordance with the Black Scholes Option Pricing Formula [37], and market crashes tend to coincide with periods of high interest rates. In other words, the strategy works well under conditions of high volatility, or when viewed over the long term where such periods will manifest (usually quickly and unexpectedly), which is precisely the conditions of Taleb's claim.

Indeed, a formalisation of the uncertainty of asset distributions as entropy maximisation, without any utility assumption - most of the mathematical finance literature dealing with entropy assumes entropy minimisation as the optimisation goal - yields the barbell portfolio as the optimal solution [17]. In this sense, the barbell strategy seems to constitute a kind of fixed-point solution to an entropic formulation of survival in high-uncertainty environments, where utility assumptions cannot properly apply; recall that a fixed-point is an invariant that resolves a self-referencing question by removing the direct self-reference. Their apparent success across a wide variety of environments and conditions seems to support this interpretation, and makes them a viable starting point for defining the kinds of behaviour that self-allocating autonomous agents might use.

Keynes [25] provided a solid macroeconomic basis for barbell strategies with his explicit separation of the impetus for profit seeking from that for failure exposure, which sets up a difficult trade-off with which agents in such a system must grapple, and which provided the basis for the necessity of governmental moderating control, particularly through monetary policy. Barbells in investment amount to replaying this split at a microeconomic level: instead of retreating into prophesies, going along with majority views, and trying to optimise future returns by riding the middles of

economic waves, agents that employ a barbell (or to a lesser extent laddering) are effectively attempting to match the bimodal nature of their investment problem with a bimodal solution aimed specifically at producing a favourably asymmetric effect. To put it more in terms suited to autonomous decision-making, such agents manage their exposure to disastrous outcomes they can envisage but that they cannot brook on the one hand, while investing whatever they can thereafter reasonably afford to lose on a selection of bets that they expect will usually bomb but that will sporadically and unpredictably return disproportionately large rewards. The first component is about hedging against unacceptable outcomes, and it only requires that agents determine their sensitivity failure and be able to determine hedging actions against it, not that they predict the future states of their environment.

The second component is about taking advantage of possible opportunities but doing so only with what the agent can bear to lose, and again this requires only that agents be able to recognise potentially propitious junctures, not that be able to predict what will happen with them in the future. This reasoning about affordable bets on sporadic high return investments is not the same as the more frequently notion of high risk and high reward, which refers to a symmetric situation in which there is high chance of an unacceptable outcome and a low chance of a very high reward. Such barbell agents will not be interested in situations of high risk and high reward in the usual sense, which would at least implicitly presuppose that they are able to reason about known or at least in principle knowable distributions over known or at least knowable sample sets of outcomes. The basic mechanism is about making asymmetric opportunity bets: the rewards are potentially high, the risk of failure is unquantifiable, but losing the bet is affordable.

Under this approach, the largest part of available resources, both intellectual and physical, are typically devoted to simply avoiding exposure to decisive failure; it would be a rare circumstance where this concern did not dominate the allocation of limited resources in a military operation. Modern defence forces arguably already operate along the lines of a barbell investment, though it seems to have not previously been formulated in these terms. Military forces plan and then plan again. They reconnoitre, looking for exposure to their vulnerabilities as they best conceive of them at the time. They inefficiently keep a third in reserve, without knowing in advance how or if it will be required, and position assets to have them available to respond quickly to the unexpected, arguably given more determination of their critical vulnerabilities than positive predictions about the future will hold. They constantly review plans in light of sudden unpredicted experiences of subordinate organisations or shifts in the strategic context, and they unabashedly change our whole approach, in principle, at least, at a moment's notice, if the available evidence compels them to refute their plans. They do everything they can to cope with the reality of being constantly disrupted.

Having so hedged against unacceptable outcomes, to the extent possible under the circumstances, with whatever resources remain, military command and control will put a little something into trying to create and exploit opportunities that just might reap disproportionately large benefits. The central point in barbell strategies is that such opportunity investment is restricted to that which the agent can afford to lose -

noting that just what it might have to be prepared to lose is a highly context-specific determination. Note that nowhere in a barbell strategy are we concerned with directly considering efficiency and maximising expected utility, nor does it heavily rely on knowledge of distributions over known outcomes. Although Taleb's version of the barbell is aimed at addressing the problem of rare events on the tails of stationary distributions, the barbell strategy appears to potentially applicable to Keynes' ontological uncertainty as well: it remains feasible in principle even within a non-ergodic system, even when the agent cannot determine the sample sets of the relevant possible outcomes.[23] It seems that our would-be autonomous machines capable of handing high uncertainty environments including those of military operations could follow the same approach.

21.10 Theory of Self

There is a further complication to consider. In economics, capital rationing refers to the process by which limits are imposed on the ability of economic agents to invest of resources [11]. An economically efficient market implies access to capital markets to obtain resources, whereby an agents could, in principle at least, access virtually any amount of capital at market rates in order to pursue any and all investment opportunities that promise a positive return, allowing for the cost of the capital and other expenses and some margin dependent on the perceived risk of the project. In contrast, an agent operating under capital rationing faces potentially high decision-making complexity because of the investment options can no longer be considered in isolation.

Soft capital rationing occurs when the agent itself exercises an internal policy restricting the size of investments, which can be understood as an attempt to manage exposure to uncertainty. Examples of soft capital rationing include internal budget allocations, setting aside capital for unforeseen contingencies, and setting a hurdle rate, which is a minimum rate of return as a required compensation for the perceived risk of the option. A hurdle rate can also be viewed as an opportunity cost, which might be evaluated as the rate of return from the next best investment opportunity having a similar perceived risk profile. Hard capital rationing is externally imposed, where the agent cannot raise capital through equity or debt. Regulatory capital requirements on banks, an inability to raise capital because of previous low performance, and legal prohibition on national defence organisations from accessing capital markets are all examples of hard capital rationing.

In addition to forcing the agent to simultaneously consider multiple options against one another, the presence of capital rationing violates the conditions of market effi-

[23]The dismissal mentioned earlier that Taleb makes towards the distinction between rare events in an ergodic system and events in a non-ergodic system being inconsequential might be generously re-interpreted as a recognition of the potential applicability of barbell strategies to classes of situations of non-ergodic uncertainty, and hence to the weaker ergodic rare events of Knightian uncertainty with which Taleb is concerned as well.

ciency, so we should not expect to have an economically efficient allocation of resources. This is a central point in this chapter: conventional notions of utility maximisation break down rapidly under complexity. Moreover, the quality of decision-making of agents under capital rationing is especially sensitive to unforeseen changes in the future cost of capital, which means for autonomous agents potentially exquisite sensitivity to failure under uncertainty.

Given hard and soft constraints under ontological uncertainty, decisions regarding how much resource to put into each side of the strategy amounts to judgements about own tolerance to loss, more than it is about the potential for the environment to dish up favourable or unfavourable outcomes. So at the core of the strategy is the requirement for autonomous machines to be equipped with a theory of self, specifically for the purpose of evaluating exposure to unacceptable failure and deciding hedging plans, and for recognising feasible opportunity for reward and determining investment plans given a range of such opportunities. Theory of mind in the usual sense is then really an extension of theory of own mind as the more fundamental concept for autonomous systems research. The strategy amounts, therefore, to a mechanism for substituting the unapproachable problems of prediction and knowledge acquisition in uncontrolled unstructured environments in general with the much more manageable problem of self-knowledge.

Note that soft capital rationing is self-imposed, and amounts to judgements and this condition appears to very directly imply a requirement for an agent theory of self. It also seems that a theory of self would imply that, in a sense, such agents would effectively talk to themselves, much as humans do [28, 38], constructing a kind of narrative of self as they debate with themselves about different investment options, and moderate and alter their own beliefs and expectations.

There is a deep issue lurking here that motivates and underpins this proposition about agency. Incompleteness means that self-knowledge, and thus knowledge of one's own sensitivity to failure, is inherently limited; after all, we all observe ourselves from inside ourselves. A theory of mind supporting the development and application barbell-type strategies accommodates this in two ways simultaneously. Firstly, the judgement caveat on plasticity and thus on autonomy concerning limits we consider to be operationally acceptable is about making visible to the agent itself the consequences of the limits of its own self-knowledge in terms of managing the effects of the limited ability of anything - or anyone - to determine its own failure modes. Secondly, with respect to determining potential exposure to unacceptable failure, I am advocating a defensive kind of posture: exposure to decisive failure is a matter of choosing boundaries beyond which unacceptable failure is a potential rather than a certainty. We cannot determine failure sensitivity completely or with exactitude, but we can choose those boundaries conservatively or optimistically and in priority order depending on our faith in the broader social enterprise to absorb the possible consequent failures. It seems that this problem of self-knowledge is much more manageable, however, both by virtue the fact that the system we then have to deal with is much smaller than that of the entire environment, which, after all, includes the agent itself, and by virtue of the fact that we humans are a testament to how successful in uncertain worlds agents armed with self-knowledge can be.

21.11 Conclusion

There is a deep connection between economics and autonomous systems research, for the simple reason that the two fields have at their core questions about the nature of agency as autonomous decision-making. The difference is roughly that economics is mainly concerned with descriptive models of agency, while artificial intelligence is squarely focussed on engineering prescriptive models. The interface has been, and must be, permeable. Both fields face the same basic issues about the nature of agency and, in particular, suffer from the inadequacy of current approaches with respect to decision-making under conditions of fundamental uncertainty. Previous work [44] in the economic literature has sought to exert influence primarily on economics audiences about poor representations of human agency, has noted the role in artificial intelligence of wishful mnemonics in masking the severe limitations incumbent in standard assumptions, and has concluded that artificial intelligence has not even begun to replicate the abilities of real humans to cope with fundamental uncertainty as a result.

This chapter is firstly about raising awareness among machine learning, automated reasoning and robotics communities about the relevance of the economic literature on decision-making under uncertainty. In particular, economic theory distinguishes between stochastic risk and unmeasurable epistemological uncertainty, and between epistemological uncertainty and a deeper notion of ontological uncertainty. Secondly, it is about relaying and extending the mathematical underpinnings of this economics literature. The connection from the basic economic notions of uncertainty to non-linear dynamics and ergodic theory that has been relatively recently established in economics [13, 50], with epistemological uncertainty and ontological uncertainty formally distinguishable on this basis. I have also sought here to extend this viewpoint with reference to the well established mathematical practices of formulating questions about future behaviours of non-linear systems as computational problems, whereby ontological uncertainty then manifests as formally unsolvability and incompressibility [35, 39].

Yet the formal unsolvability of computational problems is a particularly deep extension to the slightly earlier results establishing the incompleteness of non-trivial systems of formal reasoning, which means that within such systems there are always questions that cannot be answered, even with infinite resources. Ontological uncertainty can perhaps be best understood on this basis: the possibility of self-reference in any system means that some problems we might like to solve, such as predicting what will happen in the future, are paradoxical in the sense of being unresolvable within the system as either true or false.

Far from constituting an abstruse irrelevance, the practical consequences in economics of failures to handle this kind of uncertainty are difficult to overstate. The sophisticated risk models heavily implicated in the 2007–2008 GFC [3] are now widely acknowledged as having failed so spectacularly for the precise reason that they fail to address epistemological - let alone ontological - uncertainty. There is abounding circumstantial evidence that the same kinds of failures have been felt

with respect to applications of artificial intelligence for a very long time, particularly in the form of the manifest dearth of genuinely autonomous military operational systems. The failure of ergodic models in non-ergodic environments will be non-ergodic, meaning that questions about the future occurrences of failure will be unanswerable, and thus failure will be observed as unpredictable and sporadic after seductively long periods of apparent success.

In addition to the prime example of this type of failure, financial economics provides a window into the kinds of strategies that have already been applied successfully in securities markets to cope with uncertainty. Such environments display ontological uncertainty because they allow the possibility of self-reference, because future outcomes are dependent on agent behaviour, which depends on expectations about future outcomes. The consequence is that supposedly rational reward-maximising behaviour for an individual agent may easily be ultimately self-defeating; what makes apparent sense for the individual does not necessarily make sense for the system as a whole.

Barbell trading strategies and their relatives aim to change the problem choice from expected reward maximisation to survival in face of intrinsically unknowable futures. The idea is to divide resource allocation into two logical steps, with the first step being allocation of resources to avoid exposure to unacceptable outcomes. In the second step, remaining resources can be utilised to pursue opportunities, which will usually amount to affordable failures yet will sporadically reap large returns. While not additive, both concerns have to be considered together under conditions that amount economically to hard capital rationing, which make decision making much more difficult; moreover, the resource limits concerned may not be fully determinable in advance. At the centre of this picture is a requirement for self-knowledge: a theory of self seems necessary to recognising failure sensitivity and opportunity in high uncertainty environments under partially observable hard resource limits. The limits of self-knowledge together with the complexity of decision-making under hard capital rationing with the possibility of unexpected budget changes appears to imply that autonomous problem problem solving must be intrinsically social.

Autonomy boils down to developing decision processes for machines for solving problems in complex environments; ontological uncertainty, which I have emphasised in importance over epistemic uncertainty, boils down to the common occurrence in complex environments of seemingly straightforward decision-making problems for which there can be no solution. The future of autonomous research - as with economic theory - will be about changing the problem choices of the past, and, in doing so, effectively altering what it means for the problem to be acceptably solved, than it will be about advancing the technical development of most currently established methods. The technical developments matter, but are subordinate to wiser choices about how they are applied. Fundamental uncertainty has to feature as the central concern of robotics, machine learning and automated reasoning, for otherwise the account of genuinely autonomous operationally usable systems will surely remain at zero.

References

1. K. Adam, A. Marcet et al., *Booms and busts in asset prices* (Technical report, Institute for Monetary and Economic Studies, Bank of Japan, 2010)
2. M. Allais, Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de líècole amèricaine. Econometrica **21**(4), 503–546 (1953)
3. M. Auerback. Risk vs uncertainty: the cause of the current financial crisis. Technical Report Occassional Paper No 37, Japan Policy Research Institute, 2007
4. M. Bar-Hillel, A. Margalit, Newcomb's paradox revisited. Br. J. Philos. Sci. **23**(4), 295–304 (1972)
5. G. Boolos, *The Logic of Provability* (Cambridge University Press, Cambridge, 1995)
6. C. Carr, *The Book of War: Sun-Tzu's "The Art of War" & Karl Von Clausewitz's "On War"* (Random House Inc., New York, NY, USA, 2000)
7. G.J. Chaitin, *The Unknowable* (Springer Science & Business Media, New York, 1999)
8. G.J. Chaitin, *The Limits of Mathematics: A Course on Information Theory and the Limits of Formal Reasoning (Discrete Mathematics and Theoretical Computer Science)* (Springer, New York, 2003)
9. R.H. Coase, *The problem of social cost, Classic Papers in Natural Resource Economics* (Springer, New York, 1960), pp. 87–137
10. P. Cotton. When a Barbell Bond Portfolio Optimises Modified Excess Return (2012)
11. T. Cowen, *Modern Principles: Microeconomics* (Worth Publishers, Basingstoke, 2011)
12. P. Davidson, *Post Keynesian Macroeconomic Theory* (Edward Elgar Publishing, Aldershot, UK, 1994)
13. P. Davidson, Black swans and knight's epistemological uncertainty: are these concepts also underlying behavioral and post-walrasian theory? J. Post Keynesian Econ. **32**(4), 567–570 (2010)
14. P. Davidson, Is economics a science? should economics be rigorous? Real-World Econ. Rev. **59**, 58–66 (2012)
15. I. Erev, A.E. Roth, Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. Am. Econ. Rev. **88**(4), 848–881 (1998)
16. R. Frydman, G.P. O'Driscoll, A. Schotter, Rational expectations of government policy. South. Econ. J. **49**, 311–319 (1982)
17. D. Geman, H. Geman, N.N. Taleb, Tail risk constraints and maximum entropy. Entropy **17**(6), 3724–3737 (2015)
18. I. Gilboa, A. Postlewaite, D. Schmeidler, Rationality of belief or why bayesianism is neither necessary nor sufficient for rationality. Technical Report Cowles Foundation Discussion Papers 1484, Cowles Foundation for Research in Economics, Yale University (2004)
19. K. Gödel, *On formally undecidable propositions of principia mathematica and related systems* (Dover Publications, New York, 1992)
20. R.M. Gray, R.M. Gray, *Probability, Random Processes, and Ergodic Properties* (Springer, New York, 1988)
21. F. Hahn, R.M. Solow, *A Critical Essay on Modern Macroeconomic Theory* (MIT Press, Cambridge, MA, 1997)
22. P.R. Halmos, *Lectures on Ergodic Theory* (Martino Fine Books, Eastford, 2013)
23. D. Kahneman, A. Tversky, Prospect theory: an analysis of decision under risk. Econometrica. J. Econ. Soc. **47**, 263–291 (1979)

24. R. Kaye, *Models of Peano Arithmetic* (Cndon Press, Oxford, 1991)
25. J.M. Keynes, *The General Theory of Employment, Interest and Money* (Macmillan, London, 1973)
26. M. King, *The End of Alchemy: Money, Banking, and the Future of the Global Economy* (WW Norton & Company, New York, 2016)
27. F.H. Knight, *Risk, Uncertainty and Profit* (Beard Books, Washington DC, 2002)
28. E. Kross, E. Bruehlman-Senecal, J. Park, A. Burson, A. Dougherty, H. Shablack, R. Bremner, J. Moser, O. Ayduk, Self-talk as a regulatory mechanism: how you do it matters. J. Pers. Soc. Psychol. **106**(2), 304 (2014)
29. P. Krugman, R. Wells, Microeconomics (2012)
30. A.D. Lane, R.R. Maxfield, Ontological Uncertainty and Innovation. J. Evol. Econ. **15**(1), 3–50 (2005)
31. V. Lugovskyy, D. Puzzello, S. Tucker, An experimental study of bubble formation in asset markets using the tâtonnement trading institution. Technical Report Working Papers in Economics 11/07, University of Waikato, Department of Economics (2011)
32. V. Lugovskyy, D. Puzzello, S. Tucker, A. Williams, et al., Can concentration control policies eliminate bubbles? Technical Report Working Papers in Economics 12/13, University of Waikato, Department of Economics (2012)
33. M.J. Machina, Choice under uncertainty: problems solved and unsolved. J. Econ. Perspect. **1**(1), 121–154 (1987)
34. D. McDermott, Artificial intelligence meets natural stupidity. ACM SIGART Bull. **57**, 4–9 (1976)
35. L. Ming, P. Vitányi, *An introduction to Kolmogorov complexity and its applications* (Springer, New York, 2008)
36. P. Mirowski, *Machine Dreams: Economics Becomes a Cyborg Science* (Cambridge University Press, Cambridge, 2002)
37. P. Moles, N. Terry, *The Handbook of International Financial Terms* (OUP, Oxford, 1997)
38. A. Morin, Self-talk and self-awareness: on the nature of the relation. J. Mind Behav. **14**(3), 223–234 (1993)
39. A. Nies, *Computability and Randomness*, vol. 51 (Oxford University Press, Oxford, 2009)
40. P. Roberts, *The Failure of Laissez Faire Capitalism* (Clarity Press, Atlanta, 2013)
41. J.R. Shiller, Speculative asset prices. Am. Econ. Rev. **104**(6), 1486 (2014)
42. O. Shy, *The Economics of Network Industries* (Cambridge University Press, Cambridge, 2001)
43. A.H. Simon, Rationality as process and as product of thought. Am. Econ. Rev. **68**(2), 1–16 (1978)
44. R.E. Smith, Idealizations of uncertainty, and lessons from artificial intelligence. Econ.: Open-Access Open-Assess. E-J. **10**, 1 (2016)
45. R.I. Soare, *Turing Computability: Theory and Applications* (Springer, New York, 2016)
46. K. Steiglitz, D. Shapiro, Simulating the madness of crowds: price bubbles in an auction-mediated robot market. Comput. Econ. **12**(1), 35–59 (1998)
47. N. Taleb, *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets* (Random House, New York, 2005)
48. N.N. Taleb, *The Black Swan: The Impact of the Highly Improbable*, vol. 2 (Random House, New York, 2010)
49. N.N. Taleb, *Antifragile: Things that Gain from Disorder* (Random House, New York, 2012)
50. A. Terzi, Keynes's uncertainty is not about white or black swans. J. Post Keynesian Econ. **32**(4), 559–566 (2010)
51. A. Teversky, D. Hahneman, Judgement under uncertainty: Heuristic biases. Science **185**(4157), 1124–1131 (1974)
52. G.A. Timmermann, How learning in financial markets generates excess volatility and predictability in stock prices. Q. J. Econ. **108**(4), 1135–1145 (1993)
53. T. von Ghyczy, B. von Oetinger, C. Bassford, *Clausewitz on strategy: Inspiration and insight from a master strategist* (Wiley, New York, 2002)

54. T.A. Weber, A robust resolution of newcomb's paradox. Theory Decision **81**(3), 339–356 (2016)
55. C. Werndl, What are the new implications of chaos for unpredictability? Br. J. Philos. Sci. **60**(1), 195–220 (2009)