Chapter 1
Foundations of Trusted Autonomy: An Introduction

Hussein A. Abbass, Jason Scholz and Darryn J. Reid

1.1 Autonomy

To aid in understanding the chapters to follow, a general conceptualisation of autonomy may be useful. Foundationally, autonomy is concerned with an agent that acts in an environment. However, this definition is insufficient for autonomy as it requires persistence (or resilience) to the hardships that the environment acts upon the agent. An agent whose first action ends in its demise would not demonstrate autonomy. The themes of autonomy then include agency, persistence and action.

Action may be understood as the utilisation of capability to achieve intent, given awareness.[1] The action trinity of intent, capability and awareness is founded on a mutual tension illustrated in the following figure.

If "capability" is defined as anything that changes the agent's awareness of the world (usually by changing the world), then the error between the agent's awareness and intent drives capability choice in order to reduce that error. Or, expressed compactly, an agent seeks achievable intent.

The embodiment of this action trinity in an entity, itself separated from the environment, but existing within it, and interacting with it, is termed an agent, or autonomy, or intelligence.

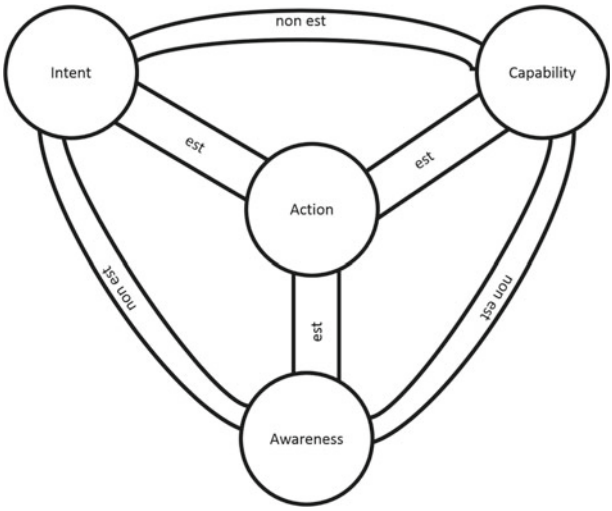
[1] D.A. Lambert, J.B. Scholz, Ubiquitous Command and Control, Intelligent Decision Technologies, Volume 1 Issue 3, July 2007, Pages 157–173, IOS Press Amsterdam, The Netherlands.

H. A. Abbass (✉)
School of Engineering and IT, University of New South Wales,
Canberra, ACT 2600, Australia
e-mail: h.abbass@adfa.edu.au

J. Scholz · D. J. Reid
Defence Science and Technology Group, Joint and Operations Analysis Division,
PO Box 1500, Edinburgh, SA, Australia
e-mail: jason.scholz@defence.gov.au

D. J. Reid
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_1



So it is fitting that Chapter 2 by Tom Everitt and Marcus Hutter opens with the topic Universal Artificial Intelligence (UAI): Practical Agents and Fundamental Challenges. Their definition of UAI involves two computational models: Turing Machines; one representing the agent, and one the environment, with actions by the agent on the environment (capability), actions from the environment on the agent (awareness), and actions from the environment to the agent including a utilisation reward (intent achievement) subject to uncertainty. The "will" that underpins the intent of this agent is "maximisation of reward". This machine intelligence is expressible - astoundingly - as a single equation. Named AIXI, it achieves a theoretically-optimal agent in terms of reward maximisation. Though uncomputable, the construct provides a principled approach to considering a practical artificial intelligence and its theoretical limitations. Everitt and Hutter guide us through the development of this theory and the approximations necessary. They then examine the critical question of whether we can trust this machine given machine self-modification, and given the potential for reward counterfeiting, and possible means to manage these. They also consider agent death and self-preservation. Death for this agent involves the cessation of action, and might represented as an absorbing zero reward state. They define both death and suicide, to assess the agent's self-preservation drive which has implications for autonomous systems safety. UAI provides a fascinating theoretical foundation for an autonomous machine and indicates other definitional paths for future research.

In this action trinity of intent, capability, and awareness, it is intent that is in some sense the foremost. Driven by an underlying will to seek utility, survival or other motivation, intent establishes future goals. Chapter 3 Benjamin Johnson, Michael Floyd, Alexandra Coman, Mark Wilson and David Aha consider Goal Reasoning and Trusted Autonomy. Goal Reasoning allows an autonomous system to respond more successfully to unexpected events or changes in the environment. In relation to UAI, the formation of goals and exploration offer the massive benefit of exponen-

tial improvements in comparison with random exploration. So goals are important computationally to achieve practical systems. They present two different models of Goal Reasoning: Goal-Driven Autonomy and the Goal Lifecycle. They also describe the Situated Decision Process (SDP), which manages and executes goals for a team of autonomous vehicles. The articulation of goals is also important to human trust, as behaviours can be complex and hard to explain, but goals may be easier because behaviour (as capability action on the environment) is driven by goals (and their difference from awareness). Machine reasoning about goals also provides a basis for the "mission command" of machines. That is, the expression of intent from one agent to another, and the expression of a capability (e.g. a plan) in return provides for a higher level of control with the "human-on-the-loop" applied to more machines than would be the case of the "human-in-the-loop". In this situation, the authors touch on "rebellion", or refusal of an autonomous system to accept a goal expressed to it. This is an important trust requirement if critical conditions are violated that the machine is aware of, such as the legality of action.

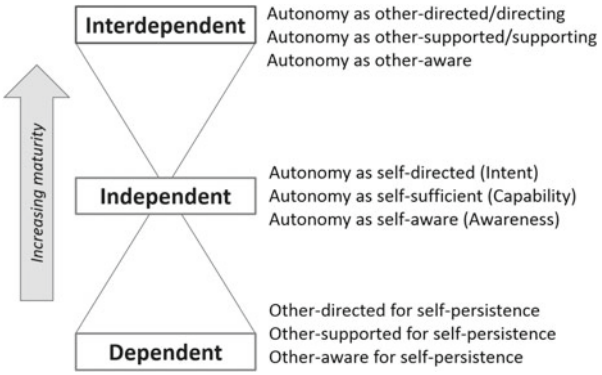
The ability to reason with and explain goals (intent) is complemented in Chapter 4 by consideration of reasoning and explanation of planning (capability). Tim Miller, Adrian R. Pearce and Liz Sonenberg examine social planning for trusted autonomy. Social planning is machine planning in which the planning agent maintains and reasons with an explicit model of the humans with which it interacts, including the human's goals (intent), intentions (in effect their plans or in general capability to act), beliefs (awareness), as well as their potential behaviours. The authors combine recent advances to allow an agent to act in a multi-agent world considering the other agents' actions, and a Theory of Mind about the other agents' beliefs together, to provide a tool for social planning. They present a formal model for multi-agent epistemic planning, and resolve the significant processing that would have been required to solve this if each agent's perspective were a mode in modal logic, by casting the problem as a non-deterministic planning task for a single agent. Essentially, treating the actions of other agents in the environment as non-deterministic outcomes (with some probability that is not resolved until after the action) of one agents own actions. This approach looks very promising to facilitate computable cooperative and competitive planning in human and machine groups.

Considering autonomy as will-driven (e.g. for reward, survival) from Chapter 2, and autonomy as goal-directed and plan-achieving (simplifying computation and explanation) from Chapters 3 and 4, what does autonomy mean in a social context? The US Defense Science board[2] signals the need for a social perspective,

> it should be made clear that all autonomous systems are supervised by human operators at some level, and autonomous systems' software embodies the designed limits on the actions and decisions delegated to the computer. Instead of viewing autonomy as an intrinsic property of an unmanned vehicle in isolation, the design and operation of autonomous systems needs to be considered in terms of human-system collaboration.

[2]U.S. Defence Science Board, Task Force Report: The Role of Autonomy in DoD Systems, July 2012, pp. 3–5.

The Defense Science Board report goes on to recommend "that the DoD abandon the use of 'levels of autonomy' and replace them with an autonomous systems reference framework". Given this need for supervision and eventual human-system collaboration, perhaps a useful conceptualisation for autonomy might borrow from psychology as illustrated in the following figure.



Here, a popular definition[3] of 'autonomy as self-sufficient and self-directed' is situated in a setting of social maturity and extended to include 'awareness of self'. Covey[4] popularises a maturity progression from dependence (e.g. on parents) via independence to interdependence. The maladjusted path is progression from dependence to co-dependence. Co-dependent agents may function but lack resilience as compromise to one agent affects the other(s) thus directly affecting own survival or utility. For the interdependent agent cut off from communication there is the fall-back state of independence.

So, if this might be a preferred trajectory for machine autonomy, what are the implications a strong and independent autonomy? In Chapter 5, Bobby D. Bryant and Risto Miikkulainen consider a neuroevolutionary approach to adaptive multi-agent teams. In their formulation, a similar and significant capability for every agent is posed. They propose a collective where each agent has sufficient breadth of skills to allow for a self-organized division of labour so that it behaves as if it were a heterogeneous team. This division is dynamic in response to conditions, and composed of autonomous agents occurs without direction from a human operator. Indeed in general, humans might be members of the team. This potentially allows for massively-scalable resilient autonomous systems with graceful degradation, as losing any agent affects a loss of role(s) which might be taken up by any other agent(s) all of which have requisite skills (capability). Artificial neural networks are used to learn teams with examples given in the construct of strategy games.

Furthering the theme of social autonomy in Chapter 6, John Harvey examines both the blessing and curse of emergence in swarm intelligence systems. We might

[3] J.M. Bradshaw, The Seven Deadly Myth of Autonomous Systems, IEEE, 2013.

[4] S. R. Covey, The Seven Habits of Highly Effective People, Free Press, 1989.

consider agents composing a swarm intelligence as "similar" and ranging to identical, but not necessarily "significant" capabilities, with the implications that resilience is a property of the collective rather than the individual. Harvey notes that swarm intelligence may relate to a category within the complexity and self-organisation spectrum of emergence characterised as weakly predictable. Swarms do not require centralised control, and may be formed from simple agent interactions, offering the potential for graceful degradation. That is, the loss of some individuals may only weakly degrade the effect of the collective. These and other "blessings" of swarm intelligence presented by the author are tempered by the shortcomings of weak predictability and controllability. Indeed, if they are identical, systematic failure may also be possible as any design fault in an individual is replicated. The author suggests a future direction for research related to the specification of trust properties, might follow from the intersection of liveness properties based on formal methods and safety properties based on Lyapunov measures. Swarm intelligence also brings into question the nature of intelligence. Perhaps it may arise as an emergent property from interacting simpler cognitive elements.

If a social goal for autonomy is collaboration, then cooperation and competition (e.g. for resources) is important. Furthermore, interdependent autonomy must include machines capable of social conflict. Conflict exists where there is mutually exclusive intent. That is, if the intent of one agent can only be achieved if the intent of the other is not achieved. Machine agents need to recognise and operate under these conditions. A structured approach to framing competition and conflict is in games. Michael Barlow, in Chapter 7 examines trusted autonomous game play. Barlow explains four defining traits of games that include a goal (intent), rules (action bounds), a feedback system (awareness), and voluntary participation. Voluntary participation is an exercise of agency where an agreement to act within those conditions is accepted. Barlow examines both perspectives of autonomy for games and games for autonomy. Autonomous entities are usually termed AIs in games, and may serve a training purpose or just provide an engaging user experience. So, improving AIs may improve human capabilities. Autonomous systems can also benefit from games, as games provide a closed-world construct for machine reasoning and learning about scenarios.

These chapters take us on a brief journey of some unique perspectives, from autonomy as individual computational intelligence through to collective machine diversity.

1.2 Trust

Trust is a ubiquitous concept. We all have experienced it one way or another, yet it appears to hold many components that we may never converge on a single, precise, and concise definition of the concept. Yet, the massive amount of literature on the topic is evidence that the topic is an important one for scientific inquiry.

The main contribution of this part of the book is to showcase the complexity of the concept in an attempt to get a handle on its multifaceted nature. This part of the book is a brief inquiry into the meaning of trust, how it is perceived in human-human interaction and in human-machine interaction, and attempts to confine the ambiguity of the topic through novel perspectives and scientifically-grounded opinions.

It initially sounded logical to us to start this part of the book with those chapters discussing trust in its general form before the chapters discussing the trusted autonomy literature. As logical as this idea may sound, it is arguably biasing in a methodological treatment of trust in trusted autonomy.

The previous structure reflects the path that most research in the literature has been following. First, an attempt is made to understand the concept in the human social context then we use this understanding to define what aspect of the concept can be mapped to the human-machine interaction context. Why not? After all, we would like the human to trust and accept the machine as part of our social system.

The previous argument is the strength and weakness of the rationale behind that logic. It is a strong argument when we investigate human-machine interaction; when trust in this relationship is only a means to an end. The ultimate end is the human accepts the machine, accepts its decision, and accepts its role within a context.

However, this view falls short methodologically to study trust in trusted autonomy. In the ultimate form of trusted autonomous systems, the parties of a trusting relationship are both autonomous; thus, both parties need to establish trust in themselves, and then in each other. If one party is a human and the other is a machine, the machine needs to trust the human (machine-human trust) and the human needs to trust the machine (human-machine trust). Therefore, to merely assume that the machine needs to respect what trust is in a human system limits our grasp on the complexity of trust in trusted autonomy.

The nature of trust in a machine needs to be understood. How can machines evaluate trust is a question whose answers need to stem from studies that focus on the nature of the machine.

We then decided to flip the coin in the way we structure this part of the book. We start the journey of inquiry with a chapter written by Lewis, Sycarab and Walker. The chapter entitled “The Role of Trust in Human-Robot Interaction” paves the way to understand trust from a machine perspective. Lewis et al. present a thorough investigation of trust in human-robot interaction, starting with the identification of factors affecting trust as means for measuring trust. They conclude by calling for a need to establish a battery of tasks in human-robot interaction to enable researchers to study the concept of trust.

Kate Devitt in her chapter entitled “Trustworthiness of Autonomous Systems” starts a journey of inquiry to answer three fundamental questions: who or what is trustworthy? how do we know who or what is trustworthy? and what factors influence what or who is trust worthy? She proposes a model of trust with two primary dimensions: one related to competency and the second related to integrity. The author concludes the chapter by discussing the natural relationship between risk and trustworthiness; followed by questioning who and what should we trust?

Michael Smithson investigates the relationship between trust and uncertainty in more depth in his chapter entitled "Trusted Autonomy Under Uncertainty". His first inquiry into the relationship between trust and distrust, takes the view that an autonomous system is an automaton and investigates the human-robotic interaction from this perspective. The inquiry into uncertainty leads to discussing the relationship between trust and social dilemmas up to the issue of trust repair.

Andrew Dowse in his chapter "The Need for Trusted Autonomy in Military Cyber Security" presents on the need for trusted autonomy in the Cyber space. Dowse discusses the requirements for trust in the Cyber space by discussing a series of challenges that needs to be considered.

Bruza and Hoenkamp bring the field of quantum cognition to offer a lens on trust in their chapter "Reinforcing trust in autonomous systems: a quantum cognitive approach". They look into the interplay between system 1 - the fast reactive system - and system 2 - the slow rationale thinking system. They discuss an experiment with images, where they found that humans distrust fake images when they distrust the subject of the image. Bruza and Hoenkamp then presents a quantum cognition model of this phenomenon.

Jason Scholz in his chapter "Learning to Shape Errors with a Confusion Objective" presents an investigation into class hiding in machine learning. Through class re-weighting during learning, the error of a deep neural network on a classification task can be redistributed and controlled. The chapter addresses the issue of trust from two perspectives. First, error trading allows the user to establish confidence in the machine learning algorithm by focusing on classes of interest. Second, the chapter shows that the user can exert control on the behavior of the machine learning algorithm; which is a two-edge sword. It would allow the user the flexibility to manipulate it, while at the same time it may offer an opportunity for an adversary to influence the algorithm through class redistribution.

The last chapter in this part show cases a few practical examples from work conducted at the University of British Columbia. Hart and his colleagues in their chapter on "Developing Robot Assistants with Communicative Cues for Safe, Fluent HRI" list examples of their work ranging from Car Door Assembly all the way to the understanding of social cues and how these communicative cues can be integrated in a human-robot interaction tasks.

1.3 Trusted Autonomy

Part III of the book has a distinctively philosophical flavour: the basic theme that runs through all of its chapters concerns the nature of autonomy, as distinct from automation, and the requirements that autonomous agents must meet if they are to be trustworthy, at least. Autonomy is more or less understood as a requirement for operating in complex environments that manifest uncertainty; without uncertainty relatively straightforward automation will do, and indeed the autonomy is generally seen here as being predicated on some form of environmental uncertainty. Part III

is heavily concerned with the centre point of autonomy in terms of intrinsic motivation, computational motivation, creativity, freedom of action, and theory of self. Trustworthiness is largely seen as a here as a necessary but not sufficient condition for such agents to be trusted by humans to carry out tasks in complex environments, with considerable implications for the need for controls on agent behaviour as a component of its motivational processes.

Sun argues that agents need to have intrinsic motivation, meaning internal motivational processes, if they are to deal successfully with unpredictable complex environments. Intrinsic motivation is required under such conditions because criteria defining agent control cannot be specified prior to operation. The importance of intrinsic motivation in regards to the successful operation and acceptance by humans under conditions of fundamental uncertainty represents a challenge that requires serious redress of familiar but outdated assumptions and methodologies.

Furthermore, the ability to understand the motivation of other agents is central to trust, because having this ability means that the behaviour of other agents is predictable even in the absence of predictability of future states of the overall environment. Indeed, the argument is that predictability of the behaviour of other agents through understanding their motivations is what enables trust, and this also explains why trust is such an important issue in an uncertain operating environment.

The chapter presents an overview of a cognitive architecture – the Clarion cognitive architecture – supporting cognitive capabilities as well as intrinsic and derived motivation for agents; it amounts to a structural specification for a variety of psychological processes necessary for autonomy. In particular, the focus of the chapter in this regard is on the interaction between motivation and cognition. Finally, several simulations of this cognitive architecture are given to illustrate how this approach enables autonomous agents to function correctly.

Merrick et al. discussion on computational motivation extends a very similar argument, by arguing that computational motivation is necessary to achieve open-ended goal formulation in autonomous agents operating under uncertainty. Yet it approaches this in a very different manner, by realising computational motivation in practical autonomous systems sufficient for experimental investigation of the question. Here, computational motivation includes curiosity and novel-seeking as well as adaptation, primarily as an epistemic motivation for knowledge increase.

Agents having different prior experiences may behave differently, with the implication that intrinsic motivation through prior experience impacts trustworthiness. Thus trust is a consequence of how motivational factors interact with uncertainty in the operating environment to produce an effect that is not present under closed environments containing only measurable stochastic risk, where essentially rationality and thus trustworthiness is a definable in terms of an optimality condition that means that agents operate without a much scope for exercising choice.

The chapter concludes that the empirical evidence presented is consistent with the thesis that intrinsic motivation in agents impacts trustworthiness, in potentially simultaneously positive and negative ways, because of the complex of overlapping and sometimes conflicting implications motivation has for privacy and security. Trustworthiness is also impacted by what combination of motivations the agents employ

and whether they operate in mixed or homogeneous agent environments. Finally, if humans are to develop trust in autonomous agents, then agent technologies have to be transparent to humans.

General computational logics are used by Bringsjord and Naveen as the basis for a model of human-level cognition as formal computing machines to formally explore the consequences for trust of autonomy. The chapter thereby sets formal limits on trust very much akin to those observed for humans in the psychology literature, by presenting a theorem stating, under various formal assumptions, that an artificial agent that is autonomous (A) and creative (C) will tend to be, from the standpoint of a fully informed rational agent, intrinsically untrustworthy (U). The chapter thus refers to the principle for humans as *PACU*, and the theorem as *TACU*. The proof of this theorem is obtained using ShadowProver, a novel automated theorem proving program.

After building an accessible introduction to the principle with reference to the psychology maintaining it for humans and empirical evidence for its veracity, the chapter establishes a formal version of the principle. This requires establishing formalisations of what it means to be an ideal observer, of what it means to be creative, and of what it means to be autonomous, and a formal notion of collaborative situations. The chapter describes the cognitive calculus DeLEL in which TACU is formalised, and the novel theorem prover ShadowProver used to prove the theorem.

More broadly, the chapter seeks not just to establish the theorem, but to establish the case for its plausibility beyond the specific assumptions of the theorem. Beyond the limitations of this particular formalisation - and the authors invite further investigation based on more powerful formalisations - the TACU theorem establishes the necessity of active engineering practices to protect humans from the unintended consequences of creative autonomous machines, by asserting legal and ethical limits on what agents can do. The preconditions of autonomy and creativity are insufficient; just as with humans, societal controls in the form of legal and ethical constraints are also required.

Derwort's concerns relate to the development of autonomous military command and control (C2). Autonomous systems in military operational environments will not act alone, but rather will do so in concert with other autonomous and manned systems, and ultimately all under broad national military control exercised by human decision-makers. This is a situation born of necessity and the opportunity afforded by rapidly developing autonomous technologies: autonomous systems and the distributed C2 across them is emerging as a response to the rapid increase in capabilities of potential military adversaries and the limited ability to respond to them with the development of traditional manned platforms.

The chapter outlines a number of past scenarios involving human error in C2, with tragic consequences, to illustrate the limitations of human decision-making, and plausible military scenarios in the not-too-distant future. There are no doubt risks involved with taking the human out of the decision-making in terms of responsibility, authority and dehumanising of human conflict, yet any rational discussion on the use of autonomy in war and battle needs to also be moderated by due recognition of the inherent risks of having humans in the decision-making processes.

Autonomous systems are merely tools, and the cost of their destruction is merely counted in dollars. Therein lies a particular strength, for autonomous systems with distributed C2 has enormous potential to create and implement minimal solutions in place of the more aggressive solutions to tactical problems to which stressed humans are prone. Autonomy offers the potential to intervene in in the face of unexpected circumstances, to de-escalate, to improve the quality as well as speed of military decision-making. Therein may lie its most serious potential for military operational use.

Young presents on the application of autonomy to training systems and raises questions about how such systems will impact the human learning environments in which they are used. Chapter 19 explores this starting from the pivotal premise of traditional teaching whereby the students must have trust in the teacher to effectively concede responsibility to the teacher. What does this mean if the teacher is a machine? The chapter seeks to explore what is possible with autonomy in the classroom, and what we might reasonably expect to be plausible.

A map is presented showing the interconnected functional components of a training system, including both those that are provided by human trainees and those that might be provided by machines. It includes the functions of the teacher and the learner, including the training topic and measurement of learning. The authors present three key drivers likely to determine the future of autonomous systems in training and education: autonomous systems development, training systems, and trust. Some of the functions required for a learning environment are already being provided by machines, albeit in relatively limited ways; the advance of autonomous systems technologies will expand the potential for delegating more of these functions to machines.

Trust is presented as a function of familiarity, which is consistent with the view of trust in some preceding chapters as requiring predictability of other agents' behaviours even within a complex environment that is inherently unpredictable. Trust is held to be central to learning, and trust through familiarity over time is the basis for exploring a number of future scenarios. The first revolves around the frustration that might be the result of the perceived artificiality of autonomous teachers, compounded by inconsistencies between different autonomous teachers over subsequent time periods. The second concerns the social dislocation and potential incompetence resulting from machines taking over simpler tasks from humans and thereby denying the humans knowledge of those tasks and thereby effecting the quality of higher-level human decision-making. The third is a scenario in which the machine responsible for teaching the human grows up with the human in a complex relationship marked by mutual trust, suggesting that the human's trust in the machine is symbiotic with the development of the machine's trust in the human.

Boyce and Griffin begin with an elucidation of the harshness and remoteness of space, marked by extreme conditions that can degrade or destroy spacecraft. Manoeuvres in orbits near earth or other large objects are complex and counter-intuitive. Gravitational fields are not uniform, interactions between multiple objects can produce significant errors, and space is becoming increasingly crowded, requiring the ability to conduct evasive actions in advance of potential collisions. Close human

operation is inefficient and dangerous, mandating the use of autonomy for a wide range of spacecraft functions.

With increasing miniaturisation of spacecraft, traffic management and collision avoidance are becoming pressing problems driving greater degrees of spacecraft autonomy. Yet the lack of trust ascribed to the limitations of automated code generation, runtime analysis and model checking for verification and validation for software that has to make complex decisions is a large barrier to adoption of higher-level autonomy for spacecraft. Linked to this is the need for human domain experts to be involved in the design and development of software in order to build trust in the product.

The chapter concludes with some possible space scenarios for autonomy, the first of which might be achieved in the near future, involving greater autonomous analysis of information from different sources. The second concerns autonomy in space traffic management, linked to all spacecraft that have the ability to manoeuvre, that includes the decision-making and action currently undertaken by humans. The final scenario concerns distributed space systems that can self-configure with minimal human input, both to achieve capabilities not achievable using single large spacecraft and to respond to unexpected events such as partial system failure.

The final chapter presents a picture of autonomous systems development primarily from an economic point of view, on the basis that an economic agent is an autonomous agent; the difference being that economics is primarily concerned with analysing overall outcomes from societies of decision-makers while AI is squarely focussed on decision-making algorithm development. The connection between economics and AI is probably more widely understood in economics - which has long utilised and contributed, in turn, to the development of machine learning and automated reasoning methods - than it is in autonomy research. Thus the chapter treats autonomy as the allocation of scarce resources under conditions of fundamental uncertainty.

The main thrust of the chapter is an economic view of uncertainty, which distinguishes between epistemic uncertainty and ontological uncertainty, and its consequences for autonomy. Ontological uncertainty is the deeper of the two: epistemic uncertainty amounts to ignorance of possible outcomes due to sampling limits, while ontological uncertainty relates to the presence of unsolvable paradoxical problems; the chapter thus draws out the connection between the economic notion of ontological uncertainty and the famed incompleteness theorems of Gödel, the unsolvability of the Halting Problem of Turing, and incompressibility theorems of Algorithmic Information Theory.

Drawing on both financial economics and macroeconomic theory, commonplace investment strategies are presented in the context of this notion of uncertainty, noting that, under conditions of ontological uncertainty, what might be seemingly rational for an individual agent in the short-term need not be rational in the long-term nor from the perspective of the entire social enterprise. Certain well-known bond investment strategies, however, appear to have the potential to strike a healthy balance and yield desirable long-term properties for both the agent and the broader system of which it

is a component, and thus may offer a basis for autonomous systems. Interestingly, implementing such a strategy in an agent seems to require a theory of self, to provide the kinds of motivational processes discussed in other chapters as well.